The nuclear medicine field has seen a rapid expansion of academic and commercial interest in developing artificial intelligence (AI) algorithms. Users and developers can avoid some of the pitfalls of AI by recognizing and following best practices in AI algorithm development.

In this article, recommendations on technical best practices for developing AI algorithms in nuclear medicine are provided, beginning with general recommendations and then continuing with descriptions of how one might practice these principles for specific topics within nuclear medicine. This report was produced by the AI Task Force of the Society of Nuclear Medicine and Molecular Imaging.

Key Words: computer/PACS; research methods; statistics, algorithm; artificial intelligence; best practices

DOI: 10.2967/jnumed.121.262567

Recent advances in artificial intelligence (AI) algorithms, together with the emergence of highly accessible AI software libraries, have led to an explosion of interest in AI within the nuclear medicine field (Fig. 1). AI, which is the development of computer systems able to perform tasks normally requiring human intelligence, is being explored in nearly every subspecialty in the chain of molecular imaging, from radiochemistry to physician report generation (Fig. 2).

The hype that propels the development of AI algorithms in nuclear medicine is counterbalanced by concerns about certain pitfalls of AI (1). The enthusiasm for AI is justified given its numerous potential benefits: AI might relieve physicians and staff from repetitive tasks, accelerate time-intensive processes, enhance image quantification, improve diagnostic reproducibility, and deliver clinically actionable information. AI promises to carry nuclear medicine beyond certain human limitations and biases. On the other hand, AI is susceptible to unique biases that are unlike the biases typically associated with human experts. There are also valid concerns about the reproducibility of claims made in many published AI studies (2) and the generalizability of trained algorithms (3). These serious issues must be addressed to ensure that algorithms earn the trust of care providers and care recipients (4).

This report was developed by the AI Task Force of the Society of Nuclear Medicine and Molecular Imaging and lays out good machine learning (ML) practices for algorithm development in nuclear medicine. Standards and recommendations for algorithm development, study design, and scientific reporting can help ensure safe technologies and reproducible gains. The report provides general recommendations for AI algorithm development, followed by recommendations that are specific to the individual subspecialties of nuclear medicine. The report focuses primarily on ML methods, as those are currently the predominant class of AI algorithms being explored in nuclear medicine, although many principles are applicable beyond ML. The target audience of the report is developers, including physicists and clinical scientists, who wish to develop AI algorithms in nuclear medicine, but the report can also benefit users (e.g., physicians) who wish to understand algorithm development. A forthcoming report from the AI Task Force focuses on appropriate methods of evaluating and validating AI algorithms in a clinical setting.

GENERAL RECOMMENDATIONS

The first part of this report describes the general pipeline of algorithm development (Fig. 3) and provides recommendations that are common to most ML applications in nuclear medicine.
The supplemental data (available at http://jnm.snmjournals.org) present a hypothetical tumor segmentation algorithm using a novel architecture (5) trained on a publicly available dataset (6,7) and follows it through all stages of development, from conception through reporting and dissemination, illustrating the recommendations provided here (Supplemental Fig. 1).

Study Design

The first step in AI algorithm development is to carefully define the task to be performed by the algorithm (Fig. 3). Investigators should collaborate with relevant stakeholders to understand whether and how the algorithm will be used in practice and then tailor the algorithm to the need. Early and regular feedback from users (e.g., clinicians) throughout the development process is necessary to properly align the algorithm’s functionality with the clinical need. Once the algorithm’s task is defined, studies should then be designed to train and evaluate the algorithm.

It is recommended that nuclear medicine AI studies be classified as either method development studies or evaluation studies, so that each class can be held to unique technical standards (Table 1). Method development studies are defined as studies that introduce a novel method or demonstrate the feasibility of a new application (i.e., proof of concept). Most recently published studies are method development studies. The evidence produced by these studies is insufficient to support a claim about how the trained algorithm is expected to perform clinically, often because of limited datasets and insufficient clinical evaluation techniques. Once an algorithm has shown technical promise in a method development study, the algorithm would then move on to a clinical evaluation phase in which a trained algorithm’s biases and limitations in a clinical task are evaluated to provide evidence substantiating a clinical claim. Evaluation studies must be performed using datasets that are external to the development dataset and should use algorithms that are frozen—that is, are beyond the training stage (e.g., commercial software). Evaluation studies might include reader studies; phantom studies; and, potentially, mult-center masked randomized controlled trials. Both classes of studies play important roles in advancing the field, and well-conducted studies of both classes should have a pathway to publication (potentially even in the same publication, if appropriate). Yet both classes of studies require unique design considerations. By holding both types of study to higher technical standards, it is hoped that the field can better avoid common weaknesses found in AI publications, including poor reproducibility, overly optimistic performance estimation, lack of generalizability, and insufficient transparency. The technical standards for both study types are discussed throughout this report and are summarized in Tables 1 and 2. Requirements for clinical evaluation studies will be further described in a forthcoming companion report from the AI Task Force.

The pathway that a technology will take to reach clinical adoption should depend on the degree of risk it poses to patients. Risk categories for software have been proposed by the International Medical Device Regulators Forum and adopted by the U.S. Food and Drug Administration (8). Software in the highest risk category will require prospective studies to validate clinical claims. Prospective studies should use preregistered statistical analysis plans (9).

AI algorithms will require postdeployment monitoring to ensure safety and quality. A decline in performance might occur for a variety of reasons, such as new scanners or shifting patient demographics. Developers should plan to seek extensive user feedback and gather performance data after clinical deployment to detect and mitigate algorithm nonconformance and identify opportunities for improvement.

Data Collection

Collecting and labeling data are typically the most time-consuming aspects of algorithm development but also have the greatest dividends. An ML algorithm is ultimately a reflection of its training data, and its performance can be affected by the amount and quality of its training data. In nuclear medicine, collecting large datasets can be challenging because of the lower volumes of examinations compared with other modalities and applications.

A data collection strategy should be designed with a goal of avoiding the biases that might result from an insufficiently representative training dataset. Biases can be clinical (how well the

NOTEWORTHY

- AI studies are being published with increasing frequency in nearly all subspecialties of nuclear medicine.
- Common pitfalls to AI studies include poor reproducibility, overly optimistic performance statements, lack of generalizability, and insufficient transparency.
- Technical best practices in AI algorithm development can help ensure reproducible scientific gains and accelerated clinical translation.
- Some general recommendations include working closely with domain experts, collecting representative datasets, developing models using cross validation, following published reporting guidelines, making models and codes available, and being fully transparent about dataset characteristics and algorithm failure modes.
- Some specific recommendations for nuclear medicine subspecialties include evaluating image enhancement algorithms through reader studies, using multiple annotators to train and evaluate segmentation and diagnostic algorithms, making sure that algorithms performing clinical tasks are interpretable, and removing redundant features from radiomics analysis.
training data reflect the clinical condition or pathologic features), technical (scanner models, acquisition protocols, reconstruction settings), demographic (racial and socioeconomic demographics, age, sex, habitus), and selection-based (e.g., tertiary vs. community hospital). For each of these biases, a structural or distribution mismatch between the training and deployment domains can result in unintended model outputs. Datasets should ideally be curated to contain the features and abnormalities that the algorithm is expected to face once deployed. Domain experts (e.g., clinicians) should guide the collection of representative cases.

It is challenging to determine the number of cases needed for algorithm development. For algorithm training, having more data is better, as long as the data are of high quality (i.e., capturing the data distribution of the targeted population). No formal guidelines exist for estimating the size of the training set, although some practical approaches have been described (10), and trial and error are therefore often necessary (11). For evaluation studies, however, sample sizes can be guided by statistical power calculations (12).

Data augmentation can be particularly useful for deep learning applications in nuclear medicine. By synthetically modifying the input data, being careful not to break the association between the input data and their target labels, dataset sizes can be artificially increased (13). Also, using a different dataset to pretrain a model can enhance the model’s capability to learn certain features and associations when labeled data are limited, although there is a risk of model overparameterization (14).

Data Labeling

For supervised ML, labels should reflect the desired output of the algorithm in both form and quality. Labels might be generated by expert opinion, computer simulation, or other methods. The labels should be regarded by experts in the field to be sufficient standards of reference. Different labeling techniques are typically possible for a given task, often yielding different degrees of quality as illustrated in Figure 4 for diagnostic applications. When labels are based on expert opinion, it is recommended that a detailed and thorough guide to labeling be developed and discussed among labelers to reduce inter- and intraobserver variability.

Because of the high cost of expert labeling, tradeoffs are nearly always made between the number of cases that can be labeled and the quality of those labels. For some tasks, having more labelers per sample can produce greater performance gains than using a larger dataset but with fewer labelers (15,16).

Because of the scarcity of labeled nuclear medicine datasets, methods that minimize labeling efforts and maximize the use of unlabeled data should be considered. Labeling is often a bottleneck in algorithm development, yet troves of unlabeled data sit dormant in clinical databases. Developers should consider data-efficient approaches to algorithm development, including semisupervised learning algorithms (17), active learning, contrastive learning, pretraining with proxy tasks, and self-supervised learning (18).

Model Design

Investigators are often faced with numerous options when selecting or designing a model for a particular task. Options can include supervised or unsupervised learning and use of neural networks or decision trees, among others. Benchmark datasets and data science competitions are useful resources for exploring different options (19).

For development studies, investigators should compare different model types. To avoid unnecessary complexity, investigators using large models are encouraged to also evaluate simpler models as a baseline comparison (e.g., logistic regression (20)). For a fair comparison of models, hyperparameters for all models should be sufficiently tuned. The approach used for hyperparameter optimization, including how many models were trained and compared, should be reported in the publication. For method development studies that introduce a novel architecture, ablation analysis is recommended (21).
When comparing AI models, small performance differences between candidate models have to be carefully interpreted. Random initialization of model weights can result in sizeable performance differences between training sessions even when identical architectures are trained with identical data. If feasible, repeated training with random initialization or with repeated holdout should be performed to provide confidence intervals of a model’s performance, which can be used to more rigorously compare different models.

**Model Training**

A critical part of model training is the partitioning of labeled datasets into disjoint sets. Each set serves a different purpose: the training set for updating the model’s weights, the validation set for hyperparameter tuning or model selection (if needed), and the testing set for estimating the model’s performance on unseen data. Partitioning a dataset reduces the risk of obtaining overly optimistic performance estimates due to overfitting to its own dataset. For this same reason, careful attention should be paid to preventing information from being leaked from the test set to the model during training. This can happen when, for example, a model is repeatedly retrained after evaluating it on the test set (i.e., tuning to the test set). Investigators should use the validation set to monitor model convergence (i.e., loss curves) to prevent underfitting and overfitting.

Cross validation is recommended for method development studies, whereas holdout or external test sets should be used for evaluation studies. In cross validation, the training, validation, and test datasets are repeatedly sampled from the overall dataset and a different model is trained and evaluated with each sampling. There are several approaches to cross validation (22), some of which are illustrated in Figure 5. Generally, data partitioning should aim to preserve data and class distributions in each of the data splits. A drawback of cross validation is that it creates multiple models and may not be computationally feasible for large models. However, for limited datasets, cross validation produces a less biased estimate of a method’s generalization performance than using 1-time partitions (i.e., holdout testing) (23). The latter should be used in development studies only when cross validation is technically infeasible or for large datasets.

Federated learning can be considered for multiinstitution studies in which pooling of data across institutions is challenging or prohibited because of privacy concerns. In federated learning, data cohorts reside within their respective institutional boundaries but models and weights are shared across institutions (24).

**Model Testing and Interpretability**

After model training and selection, the model’s technical performance is determined. Model testing, especially when using the developmental dataset, does not typically result in evidence to substantiate broad clinical claims.

Models are tested using a test dataset, which should be an unseen holdout dataset or—for development studies—may consist of all the data through cross validation (Fig. 5). The test set should have data and class distributions similar to those of the target population. The target population must be explicitly defined (e.g., “Hodgkin lymphoma patients scanned in our department in 2020”). Additional test cohorts that are external to the developmental data are highly desirable, as they provide an estimate of the algorithm’s sensitivity to covariate or dataset shift.

Model performance is quantified using evaluation metrics. Selection of evaluation metrics should be based on how well they reflect the failures and successes of the algorithm for the specific application. However, evaluation metrics are often unable to detect all the ways in which an algorithm fails, and summary statistics can hide meaningful errors (25). Investigators should seek to detect cases of failure and work to understand their causes. This work will often include visual inspection of the model output. It is recommended that challenging cases be included in the test set to probe the model’s limitations. Investigators should also directly compare the AI model’s performance with another acceptable standard, such as the standard of care. It is recommended that subgroup analysis be conducted to identify whether the algorithm is biased against any cohorts.

Investigators should attempt to make their algorithms interpretable to users, especially algorithms that perform clinical tasks (4). Interpretable algorithms attempt to explain their outputs by highlighting the properties of the input data that most impacted the model’s prediction. Interpretability may help identify confounding factors that are unrelated to the task or pathology yet unintentionally guide the model’s predictions (3). Popular approaches include tracking gradients through the network (e.g., gradient-weighted class activation mapping) or iteratively perturbing or occluding parts of the input data (e.g., Shapley additive explanations) (26).

**Reporting and Dissemination**

The quality of the reporting of AI studies is a key determinant of its subsequent impact on the field. Formal guidelines for reporting of AI studies are emerging (27,28), including some that have been proposed (29–31) and others that are forthcoming (32–34).

---

**TABLE 1**

Proposed Standards for Development Studies Vs. Evaluation Studies

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Development studies</th>
<th>Evaluation studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility of code, models, and executables</td>
<td>Necessary for publication</td>
<td>Encouraged</td>
</tr>
<tr>
<td>Use of external datasets</td>
<td>Encouraged</td>
<td>Required</td>
</tr>
<tr>
<td>Subgroup analysis for biases</td>
<td>Encouraged (if applicable)</td>
<td>Required (if applicable)</td>
</tr>
<tr>
<td>Clinical claims</td>
<td>None</td>
<td>Required</td>
</tr>
<tr>
<td>Annotation quality</td>
<td>Fair to high</td>
<td>High</td>
</tr>
<tr>
<td>Ablation studies</td>
<td>Encouraged (if applicable)</td>
<td>Not necessary</td>
</tr>
<tr>
<td>Comparison of architectures</td>
<td>Encouraged (if applicable)</td>
<td>Not necessary</td>
</tr>
<tr>
<td>Novelty in technology or application</td>
<td>High (for publication)</td>
<td>Not necessary (for publication)</td>
</tr>
<tr>
<td>Data splitting</td>
<td>Cross validation</td>
<td>Holdout or external</td>
</tr>
</tbody>
</table>

**Bradshaw et al.**
<table>
<thead>
<tr>
<th>Category</th>
<th>Topic</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>Task definition</td>
<td>Collaborate with domain experts, stakeholders</td>
</tr>
<tr>
<td>Study types</td>
<td></td>
<td>Identify publications as development studies or evaluation studies</td>
</tr>
<tr>
<td>Risk assessment</td>
<td></td>
<td>Assess the degree of risk that algorithm poses to patients and conduct study accordingly</td>
</tr>
<tr>
<td>Statistical plan</td>
<td></td>
<td>Preregister statistical analysis plans for prospective studies</td>
</tr>
<tr>
<td>Data collection</td>
<td>Bias anticipation</td>
<td>Collect data belonging to classes or groups that are vulnerable to bias</td>
</tr>
<tr>
<td></td>
<td>Training set size estimation</td>
<td>Estimate size on the basis of trial and error, or prior similar studies</td>
</tr>
<tr>
<td></td>
<td>Evaluation of set size estimation*</td>
<td>Use statistical power analysis for guidance</td>
</tr>
<tr>
<td></td>
<td>Data decisions</td>
<td>Use justified, objective, and documented inclusion and exclusion criteria</td>
</tr>
<tr>
<td>Data labeling</td>
<td>Reference standard</td>
<td>Use labels that are regarded as sufficient standards of reference by the field</td>
</tr>
<tr>
<td></td>
<td>Label quality</td>
<td>Justify label quality by application, study type, and clinical claim (Fig. 4)</td>
</tr>
<tr>
<td></td>
<td>Labeling guide*</td>
<td>Produce detailed guide for labelers in reader studies</td>
</tr>
<tr>
<td></td>
<td>Quantity/quality tradeoff</td>
<td>Consider multiple labelers (quality) over greater numbers (quantity)</td>
</tr>
<tr>
<td>Model design</td>
<td>Model comparison*</td>
<td>Explore and compare different models for development studies</td>
</tr>
<tr>
<td></td>
<td>Baseline comparison</td>
<td>Compare complex models with simpler models or standard of care</td>
</tr>
<tr>
<td></td>
<td>Model selection</td>
<td>Report model selection and hyperparameter tuning techniques</td>
</tr>
<tr>
<td></td>
<td>Model stability</td>
<td>Use repeated training with random initialization when feasible</td>
</tr>
<tr>
<td></td>
<td>Ablation study*</td>
<td>Perform ablation studies for development studies focusing on novel architectures</td>
</tr>
<tr>
<td>Model training</td>
<td>Cross validation*</td>
<td>Use cross validation for development studies; preserve data distribution across splits</td>
</tr>
<tr>
<td></td>
<td>Data leakage</td>
<td>Avoid information leaks from test set during model training</td>
</tr>
<tr>
<td>Model testing and</td>
<td>Test set</td>
<td>Use same data and class distribution as for target population; use high-quality labels</td>
</tr>
<tr>
<td>interpretability</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target population</td>
<td>Explicitly define target population</td>
</tr>
<tr>
<td></td>
<td>External sets</td>
<td>Use external sets for evaluating model sensitivity to dataset shift</td>
</tr>
<tr>
<td></td>
<td>Evaluation metric</td>
<td>Use multiple metrics when appropriate; visually inspect model outputs</td>
</tr>
<tr>
<td></td>
<td>Model interpretability*</td>
<td>Use interpretability methods for clinical tasks</td>
</tr>
<tr>
<td>Reporting and</td>
<td>Reporting</td>
<td>Follow published reporting guidelines and checklists</td>
</tr>
<tr>
<td>dissemination</td>
<td>Sharing*</td>
<td>Make code and models from development studies accessible</td>
</tr>
<tr>
<td></td>
<td>Transparency</td>
<td>Be forthcoming about failure modes and population characteristics in training and evaluation sets</td>
</tr>
<tr>
<td></td>
<td>Reproducibility checks</td>
<td>Ensure that submitted materials to journals are sufficient for replication</td>
</tr>
</tbody>
</table>

*Not all recommendations are applicable to all types of studies.
†Addressed in separate report from AI Task Force.
BEST PRACTICES FOR ALGORITHM DEVELOPMENT  •  Bradshaw et al.  505

For development studies, journals should make publication contingent on the models and either the source codes (preferred) or executables being made accessible. Publications on development studies should contribute to the technical advancement of the field, which is often accomplished only through sharing. Many hosting resources are available for sharing, as listed in Table 3. Investigators should work with institutional review boards to ensure that datasets can be properly anonymized and openly shared. The paucity of large, high-quality multicenter datasets is a major hindrance to the clinical translation of AI tools in nuclear medicine, and open sharing of data would greatly benefit the nuclear medicine community. When data cannot be fully shared for privacy reasons, at least sample data should be made available so that the correct implementation of the model can be tested. Code should come with a modus operandi that does not leave any room for subjective settings, including a data dictionary defining variables and any preprocessing or parameter-tuning instructions.

In publishing evaluation studies, the scientific contribution is the reporting on the efficacy of a previously reported or commercial algorithm; therefore, referring to the description of the algorithm is deemed sufficient for publication.

Journal editors and reviewers are encouraged to systematically check that all provided materials are sufficient for replicating studies. This step could consist of reproducibility checklists (35) or dedicated data-expert reviewers, similar to statistics reviewers that are solicited for articles involving sophisticated statistical analyses. These demanding but desirable actions have been adopted in other fields and will serve to accelerate development and validation of AI algorithms.

Investigators should be forthcoming about limitations and failures of their algorithm (36). Failure modes should be carefully described, along with positive results. Developers should provide detailed descriptions of the characteristics and limitations of the training and evaluation datasets, such as any missing demographic groups.

Evaluation

Algorithm evaluation refers to the quantification of technical efficacy, clinical utility, biases, and postdeployment monitoring of a trained algorithm. After a successful development study, a trained algorithm should be subjected to a thorough evaluation study. Evaluation studies should involve clinical users of the algorithm and produce evidence to support specific claims about the algorithm. Clinical evaluation of a diagnostic algorithm requires reader studies, in which expert nuclear medicine physicians or radiologists assess how AI algorithms impact image interpretation or clinical decision making, often in comparison to a reference method. There are numerous additional considerations to algorithm evaluation, and a separate forthcoming report from the Society of Nuclear Medicine and Molecular Imaging AI Task Force focuses specifically on these evaluation studies and the claims that result from them.

SPECIFIC APPLICATIONS

The following subsections deal with the application of AI in the various subspecialties of nuclear medicine (Fig. 2). Each section describes how AI might be used in the different domains of nuclear medicine, together with best practices in algorithm development for each type of application and considering the different components of the development pipeline (Fig. 3).

Image Reconstruction

There is great anticipation about the benefits that AI might provide to image reconstruction, including faster reconstruction, improved signal-to-noise ratio, and fewer artifacts. AI might also contribute to different components of image reconstruction, such as direct parametric map estimation, accelerated scatter correction, and attenuation correction for PET/MRI, PET-only, and SPECT-only systems.

In general, 2 classes of approaches are being explored in nuclear medicine reconstruction: those that incorporate neural networks into current physics-based iterative reconstruction methods, and those that directly reconstruct images from projection data (37). Studies on the merits of end-to-end approaches versus penalty-based approaches are needed. Furthermore, for end-to-end algorithms, innovative solutions are needed to handle the large size of 3-dimensional time-of-flight sinograms, as the memory constraints of graphics processing units have limited methods to either single-slice and non–time-of-flight applications or have required sinogram rebinning (38). Solutions might include multi–graphics processing unit parallelization or dimensionality reduction strategies.

The large impact that AI-based reconstruction methods might have on patient care demands that algorithms be sufficiently validated. Investigators should use figures of merit to evaluate image quality, such as mean-squared error, structural similarity index, or peak signal-to-noise ratio, but should also recognize that these metrics might be misleading, as small, diagnostically important features could potentially be added or removed from images without significantly impacting summary statistics (25). Therefore, evaluation studies will require reader studies with clinically focused tasks (e.g., lesion detection). Models that use anatomic priors (e.g., CT) should be tested for robustness to functional–anatomic
misregistration. For development studies, computational model-observer–based studies might prove more economic in identifying promising methods (39).

Overall, comparative studies of different AI-based reconstruction approaches are needed, and evaluation studies should use task-oriented figures of merit and validation methods (i.e., reader studies).

Postreconstruction Image Enhancement

AI methods can enhance reconstructed nuclear medicine images with more favorable qualitative or quantitative properties, with many of the same benefits as AI-based reconstruction, including lower noise, artifact removal, and improved spatial resolution.

Denoising of low-count PET images has been the subject of numerous publications and even commercial software (40). Training data often consist of pairs of images reconstructed from fully sampled and subsampled list-mode data. Subsampling should span the entire length of the examination time so that motion and tracer distribution are consistent between the image pairs. Investigators should compare the performance of denoising networks with other denoising approaches, such as gaussian smoothing and more advanced methods such as nonlocal means. Contrast, feature quantification, and noise levels should be systematically evaluated.

Algorithms might be sensitive to outliers (e.g., implants) or artifacts (e.g., motion) and should always be evaluated on challenging, out-of-distribution cases. For applications that use coregistered CT or MR images as inputs, networks should be evaluated for robustness to misregistration (41).

Traditional figures of merit to evaluate denoising methods may be misleading (42). Metrics such as signal-to-noise ratio, mean squared error, and quantitative bias should be used to evaluate gains in image quality while also ensuring quantitative fidelity. However, these metrics may not reflect the presence or absence of clinically meaningful features. Also, AI can create synthetic-looking or overly smooth images. Thus, evaluation should consist of human observer or model observer studies.

In short, image enhancement algorithms should undergo sensitivity studies and reader evaluation studies, and performance should be compared with existing enhancement methods.

Image Analysis

AI is anticipated to automate several image analysis tasks in nuclear medicine, such as in oncoligic imaging (e.g., lesion detection, segmentation, and quantification (43,44)), cardiac imaging (e.g., blood flow analyses), brain imaging (e.g., quantification of neurodegenerative diseases), and dosimetry, among others (44,45). Automation of these tasks has significant potential to save time, reduce interobserver variability, improve accuracy, and fully exploit the quantitative nature of molecular imaging (46,47).

AI-based segmentation algorithms should be task-specific. For instance, segmentation for radiotherapy target volume delineation requires datasets and labeling techniques different from those for segmentation for prediction of overall survival (though they are related). An algorithm might be sufficient for one metric but not another (43). Images from other modalities, such as CT and MRI, that provide complementary high-resolution information can also be considered as inputs to an algorithm if expected to be available clinically.

Segmentation algorithms are typically trained using expert-generated contours. To ensure appropriate and consistent labeling (Fig. 4), clear annotation instructions should be distributed to qualified labelers to guide them on viewing settings, on handling functional-anatomic misregistration, and other conditions that might affect segmentation. Expert contours will inevitably have interobserver variability, which should be measured and used as a point of comparison for automated methods. Various methods exist for creating consensus contours from multiple observers (e.g., simultaneous truth and performance-level estimation algorithm (49)). Investigators should also be aware of the various objective functions and evaluation metrics for segmentation and of the existing guidelines for validation and reporting of autosegmentation methods (49). Because of the sparsity of large, high-quality labeled datasets in nuclear medicine, phantom or realistic simulation data can also be used for model pretraining (47,50).

Overall, the development of AI segmentation algorithms should include meticulous, task-specific labeling practices, and published guidelines for validating and reporting of algorithms should be followed.

AI and Radiomics as a Discovery Tool

AI is expected to play a critical role in assisting physicians and scientists in discovering patterns within large biologic and imaging datasets that are associated with patient outcome. Modern ML methods have shown promise as useful tools to uncover hidden
but meaningful relationships within datasets (57). AI is therefore a useful adjunct to radiomics.

First, ML can be used to identify deep radiomic features whose definitions depend completely on the data and on the task, unlike handcrafted radiomic features that are mathematically predefined whatever the data. Second, ML is an effective way to mine large numbers of radiomic features, possibly augmented by other omics or clinical data, to identify associations, reduce redundancy, produce tractable representations in low-dimension spaces, or design prediction models. Unsupervised ML might be used to combine correlated input features into a smaller, more tractable set of factors (52) to or to select features relevant to a task. Redundancy in features can arise from technical causes (e.g., mathematical equivalence of radiomics features), from measurement of the same underlying biologic factor, or from a biologic causal relationship (some biologic factor influences multiple feature values). By distinguishing among these 3 situations, investigators can better approach dimensionality reduction (53). For example, mathematic equivalence of radiomics features can be detected by randomly perturbing the image and assessing which correlations persist through the perturbations (54).

The challenge of discovering predictive signatures in high-dimension datasets might necessitate a multistep approach. Investigators might first start with a selection of cases that represent both ends of the label’s range of values, such as short and long survival, to maximize the chances of detecting features associated with outcome but at the cost of low specificity.

After initial discovery, whatever features or relationships have been identified must be rigorously evaluated and scrutinized. Investigators must explore the relationships across the entire dataset using cross validation, aim to understand the underlying cause, and then externally validate these findings, ruling out false-positives or spurious correlations. For example, they can repeat the whole AI-analysis pipeline on sham data (e.g., randomized labels) to determine the baseline false-positive rate for their set of methods and then compare it with the discovery rate found in the real dataset. Investigators should also test different models and architecture to see whether the discovered relationships hold, as it is unlikely that a real association will be identified by only 1 model.

In short, radiomics analysis should include the removal of redundant features, and a multistep approach of discovery (high sensitivity, low specificity) followed by rigorous validation might be considered.

Detection and Diagnosis

Computer-aided diagnosis and detection have long histories of successes and failures in radiology, but the recent advancements in AI have made widespread use of computer-aided diagnosis and computer-aided detection an approaching reality for nuclear medicine. Automation of diagnostic tasks in nuclear medicine can be challenging, as diagnostic tasks are subjective, have high stakes, and must be incredibly robust to rare cases (e.g., implants or amputations). However, the incentive to develop such tools is strong, with applications including assisted reads (55), tumor detection suggestions, neuro or cardiac diagnosis tools (56), training programs for residents, and many others.

Investigators should select an appropriate labeling technique according to the accuracy that is needed for their computer-aided diagnosis or computer-aided detection application (Fig. 4). Labels from specialists are superior to those from trainees or generalists, and labels resulting from multiple readers (adjudication or consensus) are superior to those from single readers. Labels extracted from clinical reports are considered inferior to those obtained from dedicated research readings (57). Intraobserver and interobserver variability in labels is often an indicator of label quality and should be quantified and reported.

Investigators are encouraged to integrate model interpretability (e.g., Shapley additive explanations) and uncertainty signaling (e.g., Bayesian approximation) into their algorithm. Because diagnostic algorithms will be used under the supervision of a physician, algorithm decisions should ideally be explainable so that clinicians have sufficient information to contest or provide feedback when algorithms fail. Developers also need to be transparent about their algorithm development and evaluation processes, including data sources and training set population characteristics, such as by using reporting checklists such as MI-CLAIM (29).

The high visibility and public attention that AI-based diagnostic algorithms receive demands that developers make every effort to be fully transparent.

In short, for computer-aided detection and computer-aided diagnosis algorithms, label quality should be justified by the application (high quality for high-risk applications) and algorithms should be interpretable and fully transparent.

Enhanced Reporting and Imaging Informatics

ML has the potential to transform how the information within diagnostic images is translated into reports and clinical databases. AI can be used to prepopulate radiology reports, assist in real-time report generation, help standardize reporting, and perform structured synoptic reporting (58).

Algorithm development in medical imaging informatics has several unique considerations. A critical challenge is the large heterogeneity in diagnostic reporting standards and practices across institutions, individual physicians, and examination types. Heterogeneity in language can be more challenging for automation than is heterogeneity in medical images. Therefore, training data should be collected from diverse sources and annotators, and studies are
expected to require much larger sample sizes than for other applications. Tasks in this domain might be uniquely suitable to unsupervised or semisupervised approaches because of the large volume of unlabeled data available in clinical PACS systems. Various model types will likely be applied in this domain, but language models may need to be adapted to consider the unique nuclear medicine vocabulary that might not be represented in typical medical text corpora (e.g., the term SUV). Because of challenges in deidentification of radiology reports (59), federated learning should be considered to enable privacy-protected multiinstitutional studies. Reporting of model performance should be disaggregated according to data source, originating institution, and annotator.

**Clinical Intelligence and Decision Support**

Clinical intelligence and decision support are concerned with delivering actionable advice to clinicians after extracting, distilling, and consolidating clinical information across multiple data sources. These systems are expected to pull the most pertinent information generated by a nuclear medicine examination and combine it with other clinical data to best guide patient care. For example, ML can predict future myocardial infarction using PET features combined with other clinical variables (60). The development and validation of clinical decision support systems should be guided by physician needs and clinical experts, involving teams from nearly all sectors of health care.

An algorithm’s ability to explain its decisions is key to safe, ethical, fair, and trustworthy use of AI for decision support, calling for the same recommendations as discussed in the section on detection and diagnosis. An AI model should ideally be able to provide an estimate of uncertainty together with its output, possibly by using Bayesian methods, and be willing to provide a no-decision answer when the model uncertainties are too large to make the output meaningful.

**Instrumentation and Image Acquisition**

Challenging problems in data acquisition and instrumentation could be well suited to ML-based solutions (61). For example, ML has been used to estimate 2- and 3-dimensional position of interaction for detectors (62). Other promising applications include timing pickoff for detector waveforms, intercrystal scatter estimation, patient motion detection, and the prediction of scanner failure from quality control tracking.

Precise data collection is critical to the success of AI applications within instrumentation. Simulations should be performed using appropriate models that incorporate geometric, physical, and statistical factors underlying image generation. Investigators should consider possible discrepancies between in silico and physical domains and are encouraged to conduct cross-validation studies when possible (61). Physical measurements, such as point source measurements, may require high-precision motion stages and lengthy acquisition studies to collect the full range of training data. Scanner quality control applications will likely require enterprise-level tracking to obtain sufficient data on failure patterns.

Algorithms that process events in real time and need to be implemented on front-end electronics will likely be memory- and operation-limited (63), favoring simpler model architectures. Ablation analysis can help identify more parsimonious models.

**Radiopharmaceuticals and Radiochemistry**

The potential for AI to challenge the current paradigms in synthesis (64) and administration (65) of radiopharmaceuticals is only beginning to be explored. Potential applications include predicting drug–target interactions (66), predicting and optimizing radiochemical reactions, and performing de novo design of drugs (67), as well as helping optimize radiopharmacy workflows. Proper integration of AI within the radiochemistry and radiopharmacy communities will require collaborations between key stakeholders, including industry, end users, and quality control personnel, as well as experts in information technology, cybersecurity, and regulatory aspects. It is strongly recommended that groups share manufacturing data freely, as this will accelerate innovation by providing large test sets for ML that cannot be sufficiently generated at individual labs (e.g., synthesis module and cyclotron log files).

**DISCUSSION AND CONCLUSION**

The recommendations listed in this article and summarized in Table 2 are intended to assist developers and users in understanding the requirements and challenges associated with the design and use of AI-based algorithms. They focus on specificities associated with nuclear medicine applications, whereas best practices for software development, data management, security, privacy, ethics, and regulatory considerations are largely covered elsewhere. It is also acknowledged that some standards of today are likely to be superseded by new standards as technologies continue to evolve. These recommendations should serve as a guide to developers and investigators at a time when AI is booming but should not be assumed to be comprehensive or unchanging.

These recommendations were drawn from various sources, including the authors’ collective experiences in academia and industry, as well as other published position papers, and put into the context of nuclear medicine applications. They should be considered an add-on to other guidelines, including forthcoming guidelines from regulatory bodies (68) and relevant working groups (69).

AI is expected to influence and shape the future of nuclear medicine and many other fields. But the potential pitfalls of AI warrant a careful and methodic approach to AI algorithm development and adoption. Standards and guidelines can help nuclear medicine avoid the mismatch between the role that AI is expected to play and what it will actually deliver.

**DISCLOSURE**

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

**ACKNOWLEDGMENTS**

The members of the AI Task Force acknowledge Bonnie Clarke for her excellent support of this project.