Journal of Nuclear Medicine, published on November 5, 2021 as doi:10.2967/jnumed.121.262567

Nuclear Medicine and Artificial Intelligence: Best Practices for Algorithm Development

Tyler J. Bradshaw[1], Ronald Boellaard[2], Joyita Dutta[3], Abhinav K. Jha[4], Paul Jacobs[5], Quanzheng Li[6], Chi Liu[7], Arkadiusz Sitek[8], Babak Saboury[9], Peter J.H. Scott[10], Piotr J. Slomka[11], John J. Sunderland[12], Richard L. Wahl[13], Fereshteh Yousefirizi[14], Sven Zuehlsdorff[15], Arman Rahmim[16], Irène Buvat[17]

[1]Department of Radiology, University of Wisconsin - Madison
[2]Department of Radiology & Nuclear Medicine, Cancer Centre Amsterdam, Amsterdam University Medical Centres
[3]Department of Electrical and Computer Engineering, University of Massachusetts Lowell
[4]Department of Biomedical Engineering and Mallinckrodt Institute of Radiology, Washington University in St. Louis
[5]MIM Software Inc., Cleveland, United States
[6]Department of Radiology, Massachusetts General Hospital and Harvard Medical School
[7]Department of Radiology and Biomedical Imaging, Yale University
[8]Sano Centre for Computational Medicine
[9]Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health
[10]Department of Radiology, University of Michigan Medical School
[11]Department of Imaging, Medicine, and Cardiology, Cedars-Sinai Medical Center
[12]Departments of Radiology and Physics, University of Iowa
[13]Mallinckrodt Institute of Radiology, Washington University in St. Louis
[14]Department of Integrative Oncology, BC Cancer Research Institute
[15]Siemens Medical Solutions USA, Inc., Hoffman Estates, United States
[16]Departments of Radiology and Physics, University of British Columbia
[17]Institut Curie, Université PSL, Inserm, Université Paris-Saclay, Orsay, France

Corresponding Author:
Tyler J. Bradshaw
Department of Radiology
University of Wisconsin – Madison
Email: tbradshaw@wisc.edu
ORCID: 0000-0001-9549-7002
Phone: 608-262-9249
Fax: 608-262-2413

Running title: Best Practices for Algorithm Development

Word Count: 8,342

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

NOTEWORTHY

- AI studies are being published with increasing frequency in nearly all subspecialties of nuclear medicine (Page 1).
- Common pitfalls to AI studies include poor reproducibility, overly optimistic performance statement, lack of generalizability, and insufficient transparency (Page 1).
- Technical best practices to AI algorithm development can help ensure reproducible scientific gains and accelerated clinical translation (Page 1).
- Some general recommendations include: work closely with domain experts, collect representative datasets, develop models using cross validation, follow published reporting guidelines, make models and codes available, and be fully transparent about dataset characteristics and algorithm failure modes (Pages 3-7).
- Some specific recommendations for nuclear medicine subspecialties include: evaluate image enhancement algorithms through reader studies, use multiple annotators for training and evaluating segmentation and diagnostic algorithms, algorithms that perform clinical tasks should be interpretable, and remove redundant features from radiomics analysis (Pages 7-11).

ABSTRACT

The nuclear medicine field has seen a rapid expansion of academic and commercial interests 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 for technical best practices for developing AI algorithms in nuclear medicine are provided, beginning with general recommendations followed by descriptions on 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.

INTRODUCTION

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 (Figure 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 (see Figure 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 could 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 committed by 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 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 subspecialities of nuclear medicine. The report primarily focuses on machine learning (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 it 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 (Figure 3) and provides recommendations that are common to most machine learning applications in nuclear medicine. The *Supplemental Data* presents 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/dissemination, illustrating the recommendations listed below (Supplemental Figure 1).

Study Design

The first step in AI algorithm development is to carefully define the task to be performed by the algorithm (Figure 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 (see 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). The large majority of 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 due to limited datasets and insufficient clinical evaluation techniques. Once an algorithm has shown technical promise in a method development study, it would then move on to a clinical evaluation phase in which a trained algorithm's biases and limitations are evaluated on a clinical task to provide evidence to substantiate a clinical claim. Evaluation studies must be performed using datasets that are external to the development dataset, and should use "frozen" algorithms that are beyond the training stage (e.g., commercial software). Evaluation studies might include reader studies, phantom studies, and potentially multicenter blinded 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 studies 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 employ preregistered statistical analysis plans (*9*).

AI algorithms will require post-deployment 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 non-conformance and to 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. A 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 due to the lower volumes of exams compared to other modalities and applications.

A data collection strategy should be designed with a goal to avoid the biases that might result from an insufficiently representative training dataset. Biases can be clinical (how well the training data reflects the clinical condition or pathological features), technical (scanner models, acquisition protocols, reconstruction settings), demographic (racial and socioeconomic demographics, age, gender, habitus, etc.), and selection-based (e.g., tertiary versus community hospital). For each of these biases, 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, more data is better, as long as the data is high quality (i.e., capturing the data distribution of targeted population). No formal guidelines exist for training set size estimation, although some practical approaches have been described (*10*), and therefore trial and error are 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 its target label, dataset sizes can be artificially increased (*13*). Also, using a different dataset to pre-train a model can enhance the model's capability to learn certain features and associations when labeled data is 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, etc. 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 intra-observer variability.

Due to 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*).

Due to 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 semi-supervised learning algorithms (*17*) active learning, contrastive learning, pre-training 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, use of neural networks or decision trees, etc. 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/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 hold out 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 and/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 prevent 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/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 one-time partitions (i.e., holdout testing) (*23*). The latter should only be used in development studies when cross validation is technically infeasible or for large datasets.

Federated learning can be considered for multi-institution studies in which pooling data across institutions is challenging or prohibited due to 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

Following 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 (Figure 5). The test set should have similar data and class distributions as the target population. The target population must be explicitly defined (e.g., "Hodgkin's 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. Evaluation metrics should be selected 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 will often include visual inspection of the model output. It is recommended to include challenging cases in the test set to probe the model's limitations. Investigators should also directly compare the AI model's performance to another acceptable standard, such as standard of care. It is recommended to conduct subgroup analysis to identify if there are cohorts the algorithm is biased against.

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/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 by iteratively perturbing or occluding parts of the input data (e.g., Shapley additive explanations (SHAP)) (*26*).

Reporting and Dissemination

The quality of the reporting of AI studies is a key determinant of its subsequent impact in 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*).

For development studies, journals should make publication contingent on the models and either the source code (preferred) or executables being made accessible. Publications on development studies should contribute to the technical advancement of the field, which is often only accomplished 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, so 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 could consist of reproducibility checklists (*35*) and/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 post-deployment monitoring of a trained algorithm. Following 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 and/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 (Figure 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 (Figure 3).

Image Reconstruction

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

In general, two 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 merits of end-to-end approaches versus penaltybased approaches are needed. Furthermore, for end-to-end algorithms, innovative solutions are needed to handle the large size of 3D time-of-flight sinograms, as graphics processing unit (GPU) memory constraints have limited methods to either single-slice and non-time-of-flight applications or have required sinogram rebinning (*38*). Solutions might include multi-GPU parallelization or dimensionality reduction strategies.

The large impact that AI-based reconstruction methods could 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, and/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 could prove more economical 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., readers studies).

Post-Reconstruction 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 consists of pairs of images reconstructed from fully-sampled and subsampled listmode data. Subsampling should span the entire length of the exam time so that motion and tracer distribution are consistent between the image pairs. Investigators should compare the performance of denoising networks to other denoising approaches, such as Gaussian smoothing and more advanced methods such as non-local 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 and/or model observer studies.

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

Image Analysis

AI is anticipated to automate a number of image analysis tasks in nuclear medicine, such as in oncological 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 inter-observer 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 different types of datasets and different labeling techniques than 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 could 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 (Figure 4), clear annotation instructions should be distributed to qualified labelers to guide them on viewing settings, how to handle functional-anatomic misregistration, and other conditions that might affect segmentation. Expert contours will inevitably have inter-observer 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 (*48*)). 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 auto-segmentation methods (*49*). Due to the sparsity of large, high quality labeled datasets in nuclear medicine, phantom and/or realistic simulation data can also be used for model pre-training (*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 biological 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 (*51*). AI is therefore a useful adjunct to radiomics.

First, ML can be used to identify deep radiomic features whose definitions completely depend on the data and on the task, unlike handcrafted radiomic features that are mathematically pre-defined 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-dimensional spaces, or design prediction models. Unsupervised ML might be used to combine correlated input features into a smaller, more tractable set of factors (*52*), or to select feature relevant to a task. Redundancy in features can arise from technical causes (e.g., mathematical equivalence of radiomics features), or if they measure the same underlying biological factor, or as the result of a biological causal relationship (some biological factor influences multiple feature values). By distinguishing between these three situations, investigators can better to approach dimensionality reduction (*53*). For example, mathematical 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-dimensional datasets might necessitate a multi-step 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.

Following 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 to the discovery rate found in the real dataset. Investigators should also test different models and architecture to see if the discovered relationships hold, as it is unlikely that a real association will only be identified by one model.

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

Detection and Diagnosis

Computer-aided diagnosis (CADx) and detection (CADe) have long histories of successes and failures in radiology, but the recent advancements in AI have made widespread use of CADx and CADe 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, amputations, etc). 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 CADx or CADe application (Figure 4). Labels from specialists are superior to those from trainees or generalists, and labels resulting from multiple readers (adjudication/consensus) are superior to those from single readers. Labels extracted from clinical reports are considered inferior to those obtained from dedicated research readings (*57*). Intra-observer and inter-observer 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., SHAP) 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. This can be accomplished 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 CADe and CADx 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 for different exam types. Heterogeneity in language can be more challenging for automation than 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 semi-supervised approaches due to the large volume of unlabeled data available in clinical picture archiving and communication system

(PACS) systems. Various model types will likely be applied in this domain, but language models may need to be adapted to consider the unique vocabulary in nuclear medicine which might not be represented in typical medical text corpora (e.g,. “SUV”). Due to challenges in de-identification of radiology reports (*59*), federated learning should be considered to enable privacy-protected multi-institutional 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 exam 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 healthcare.

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 machine learning-based solutions (*61*). For example, machine learning has been used to estimate 2D and 3D position-of-interaction for detectors (*62*). Other promising applications include timing pickoff for detector waveforms, inter-crystal 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 in administration (*65*) of radiopharmaceuticals is only beginning to be explored. Potential applications include predicting drug-target interactions (*66*), predicting and optimizing radiochemical reactions, and de novo drug design (*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 above, including those 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. Yet, these recommendations should serve as a guide to developers and investigators at a time where 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 as 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, as it will in many fields. But the potential pitfalls of AI warrant a careful and methodical 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.

DISCLOSURES

Tyler Bradshaw receives research support from GE Healthcare. Ronald Boellaard is (unpaid) scientific advisor 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 research grant from Siemens Medical Solutions, is a consultant for IBA, and receives royalties from Cedars-Sinai for nuclear cardiology software. No other potential conflicts of interest relevant to this article exist.

ACKNOWLEDGMENTS

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

REFERENCES

1. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3:199-217.

2. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586:E14-E16.

3. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med*. 2018;15:e1002683.

4. Buvat I, Orlhac F. The T.R.U.E. checklist for identifying impactful AI-based findings in nuclear medicine: is it True? Is it Reproducible? Is it Useful? Is it Explainable? *J Nucl Med*. Epub ahead of print .

5. Xue Y, Xu T, Zhang H, Long LR, Huang X. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*. 2018;16:383-392.

6. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep*. 2017;7:10117.

7. Andrearczyk V, Oreiller V, Jreige M, et al. Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. *Lecture Notes in Computer Science*. 2020;12603:1-21.

8. Center for Devices and Radiological Health. Software as a medical device (SAMD): clinical evaluation - guidance for industry and Food and Drug Administration staff. Food and Drug Administration website. Accessed Apr 01, 2021. https://www.fda.gov/media/100714/download.

9. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. *Proc Natl Acad Sci USA*. 2018;115:2600-2606.

10. Dirand A-S, Frouin F, Buvat I. A downsampling strategy to assess the predictive value of radiomic features. *Sci Rep*. 2019;9:17869.

11. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.

12. Smith SM, Nichols TE. Statistical challenges in “big data” human neuroimaging. *Neuron*. 2018;97:263-268.

13. Hwang D, Kim KY, Kang SK, et al. Improving the accuracy of simultaneously reconstructed activity and attenuation maps using deep learning. *J Nucl Med*. 2018;59:1624-1629.

14. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. arXiv.org website [Cornell University]. https://arxiv.org/abs/1902.07208. Submitted February 14, 2019. Accessed May 5, 2021.

15. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402-2410.

16. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264-1272.

17. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal*. 2019;54:280-296.

18. Zhu J, Li Y, Hu Y, Ma K, Zhou SK, Zheng Y. Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis. *Med Image Anal*. 2020;64:101746.

19. Papers with Code website. Accessed May 10, 20201. https://paperswithcode.com/.

20. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22.

21. Zhao K, Zhou L, Gao S, et al. Study of low-dose PET image recovery using supervised learning with CycleGAN. *PLoS One*. 2020;15:e0238455.

22. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv.org website [Cornell University]. https://arxiv.org/abs/1811.12808. Submitted November 13, 2018. Accessed May 2, 2021.

23. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statist Surv*. 2010;4:40-79.

24. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag*. 2020;37:50-60.

25. Yang J, Sohn JH, Behr SC, Gullberg GT, Seo Y. CT-less direct correction of attenuation and scatter in the image space using deep learning for whole-body FDG PET: potential benefits and pitfalls. *Radiol Artif Intell*. 2021;3:e200137.

26. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. *Phys Med Biol*. 2021;66:04TR01.

27. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health*. 2020;2:e549-e560.

28. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26:1364-1374.

29. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26:1320-1324.

30. Sengupta PP, Shrestha S, Berthon B, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist. *JACC Cardiovasc Imaging*. 2020;13:2017-2035.

31. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27:2011-2015.

32. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat Med*. 2020;26:807-808.

33. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393:1577-

1579.

34. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. 2021;27:186-187.

35. Pineau J, Vincent-Lamarre P, Sinha K, et al. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 Reproducibility Program). arXiv.org website [Cornell University]. https://arxiv.org/abs/2003.12206. Submitted March 27, 2020. Accessed May 5, 2021.

36. Reuzé S, Orlhac F, Chargari C, et al. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. *Oncotarget*. 2017;8:43169-43179.

37. Reader AJ, Corda G, Mehranian A, d. Costa-Luis C, Ellis S, Schnabel JA. Deep learning for PET image reconstruction. *IEEE Trans Radiat Plasma Med Sci*. 2021;5:1-25.

38. Whiteley W, Luk WK, Gregor J. DirectPET: full-size neural network PET reconstruction from sinogram data. *J Med Imaging (Bellingham)*. 2020;7:032503.

39. Yu Z, Rahman MA, Schindler T, Laforest R, Jha AK. A physics and learning-based transmission-less attenuation compensation method for SPECT. *Proc SPIE, Medical Imaging 2021: Physics of Medical Imaging*. 2021:1159512.

40. Katsari K, Penna D, Arena V, et al. Artificial intelligence for reduced dose 18F-FDG PET examinations: a real-world deployment through a standardized framework and business case assessment. *EJNMMI Phys*. 2021;8:25.

41. Lu W, Onofrey JA, Lu Y, et al. An investigation of quantitative accuracy for deep learning based denoising in oncological PET. *Phys Med Biol*. 2019;64:165019.

42. Yu Z, Rahman MA, Schindler T, et al. AI-based methods for nuclear-medicine imaging: Need for objective task-specific evaluation. *J Nucl Med*. 2020;61:575-575.

43. Weisman AJ, Kim J, Lee I, et al. Automated quantification of baseline imaging PET metrics on FDG PET/CT images of pediatric Hodgkin lymphoma patients. *EJNMMI Phys*. 2020;7:76.

44. Capobianco N, Meignan M, Cottereau A-S, et al. Deep-learning 18F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-cell lymphoma. *J Nucl Med*. 2021;62:30-36.

45. Weisman AJ, Kieler MW, Perlman SB, et al. Convolutional neural networks for automated PET/CT detection of diseased lymph node burden in patients with lymphoma. *Radiol Artif Intell*. 2020;2:e200016.

46. Weisman AJ, Kieler MW, Perlman S, et al. Comparison of 11 automated PET segmentation methods in lymphoma. *Phys Med Biol*. 2020;65:235019.

47. Leung KH, Marashdeh W, Wray R, et al. A physics-guided modular deep-learning based automated framework for tumor segmentation in PET. *Phys Med Biol*. 2020;65:245032.

48. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004;23:903-921.

49. Hatt M, Lee JA, Schmidtlein CR, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM task group No. 211. *Med Phys*. 2017;44:e1-e42.

50. Liu Z, Laforest R, Mhlanga J, et al. Observer study-based evaluation of a stochastic and physics-based method to generate oncological PET images. *Proc SPIE, Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*. 2021:1159905.

51. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med*. 2021;27:136-140.

52. Peeters CFW, Übelhör C, Mes SW, et al. Stable prediction with radiomics data. arXiv.org website [Cornell University]. https://arxiv.org/abs/1903.11696. Submitted March 27, 2019. Accessed May 2, 2021.

53. Pfaehler E, Mesotten L, Zhovannik I, et al. Plausibility and redundancy analysis to select FDG-PET textural features in non-small cell lung cancer. *Med Phys*. 2021;48:1226-1238.

54. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother Oncol*. 2019;130:2-9.

55. Li Z, Kitajima K, Hirata K, et al. Preliminary study of AI-assisted diagnosis using FDG-PET/CT for axillary lymph node metastasis in patients with breast cancer. *EJNMMI Res*. 2021;11:10.

56. Betancur J, Hu L-H, Commandeur F, et al. Deep learning analysis of upright-supine high-efficiency SPECT myocardial perfusion imaging for prediction of obstructive coronary artery disease: a multicenter study. *J Nucl Med*. 2019;60:664-670.

57. Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the Radiology editorial board. *Radiology*. 2020;294:487-489.

58. Panayides AS, Amini A, Filipovic ND, et al. AI in medical imaging informatics: current challenges and future directions. *IEEE J Biomed Health Inform*. 2020;24:1837-1857.

59. Steinkamp JM, Pomeranz T, Adleberg J, Kahn CE, Cook TS. Evaluation of automated public de-identification tools on a corpus of radiology reports. *Radiol Artif Intell*. 2020;2:e190137.

60. Kwiecinski J, Tzolos E, Meah M, et al. Machine-learning with 18F-sodium fluoride PET and quantitative plaque analysis on CT angiography for the future risk of myocardial infarction. *J Nucl Med*. 2021;Epub ahead of print.

61. Arabi H, Zaidi H. Applications of artificial intelligence and deep learning in molecular imaging and radiotherapy. *Eur J Hybrid Imaging*. 2020;4:1-23.

62. Gong K, Berg E, Cherry SR, Qi J. Machine learning in PET: from photon detection to quantitative image reconstruction. *Proc IEEE*. 2020;108:51-68.

63. Müller F, Schug D, Hallen P, Grahe J, Schulz V. A novel DOI positioning algorithm for monolithic scintillator crystals in PET based on gradient tree boosting. *IEEE Trans Radiat Plasma Med Sci*. 2019;3:465-474.

64. de Almeida AF, Moreira R, Rodrigues T. Synthetic organic chemistry driven by artificial intelligence. *Nat Rev Chem*. 2019;3:589-604.

65. Nelson SD, Walsh CG, Olsen CA, et al. Demystifying artificial intelligence in pharmacy. *Am J Health Syst Pharm*. 2020;77:1556-1570.

66. Wen M, Zhang Z, Niu S, et al. Deep-learning-based drug-target interaction prediction. *J Proteome Res*. 2017;16:1401-1409.

67. Ståhl N, Falkman G, Karlsson A, Mathiason G, Boström J. Deep reinforcement learning for multiparameter optimization in de novo drug design. *J Chem Inf Model*. 2019;59:3166-3176.

68. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). U.S. Food and Drug Administration website. Accessed Apr 14, 2021. https://beta.regulations.gov/document/FDA-2019-N-1185-0001.

69. IEEE artificial intelligence medical device working group. IEEE Standards Association website. Accessed April 14, 2021. https://sagroups.ieee.org/aimdwg/.

TABLES

Table 1. Proposed standards for development studies versus evaluation studies

	Development Studies	**Evaluation Studies**
Make code/models/executable accessible	Necessary for publication	Encouraged
Use of external datasets	Encouraged	Required
Subgroup analysis for biases	Encouraged (if applicable)	Required (if applicable)
Clinical claims	None	Required
Annotation quality	Fair to high	High
Ablation studies	Encouraged (if applicable)	Not necessary
Comparison of architectures	Encouraged (if applicable)	Not necessary
Novelty in technology or application	High (for publication)	Not necessary (for publication)
Data splitting	Cross validation	Holdout/external

.

Table 2. Summary of recommendations.

Category	Topic	Recommendation
Study Design	Task definition	Collaborate with domain experts, stakeholders
	Study types	Publications should identify as development studies or evaluation studies
	Risk assessment	A study's degree of rigor should depend on the risk the algorithm poses to patients
	Statistical plan	Prospective studies should preregister statistical analysis plans
Data Collection	Bias anticipation	Collect data belonging to classes/groups that are vulnerable to bias
	Training set size estimation	Based on trial and error, or prior similar studies
	Evaluation set size estimation*	Guided by statistical power analysis
	Data decisions	Inclusion/exclusion criteria should be justified, objective, and documented
Data Labeling	Reference standard	Labels should be regarded as sufficient standards of reference by the field
	Label quality	Label quality should be justified by the application, study type, and clinical claim (Figure 4)
	Labeling guide*	Reader studies should produce a detailed guide for labelers
	Quantity/quality tradeoff	Consider multiple labelers (quality) over greater numbers (quantity)
Model Design	Model comparison*	Development studies should explore and compare different models
	Baseline comparison	Complex models should be compared with simpler models and/or standard-of-care
	Model selection	The model selection and hyperparameter tuning techniques should be reported
	Model stability	Repeated training with random initialization is recommended
	Ablation study*	Development studies focusing on novel architectures should perform ablation studies
Model Training	Cross validation*	Cross validation should be used for development studies; preserve data distribution across splits
	Data leakage	Information leaks from the test/evaluation set to the model during training must be avoided
Model Testing and Interpretability	Test set	Should have same data/class distribution as the target population; high quality labels
	Target population	The target population should be explicitly defined
	External sets	External sets are recommended for evaluating model sensitivity to dataset shift
	Evaluation metric	May consist of multiple metrics; often requires visual inspection of model output
	Model interpretability*	Interpretability is needed for clinical tasks
Reporting and Dissemination	Reporting	Follow published reporting guidelines/checklists
	Sharing*	Development studies must make code and models accessible
	Transparency	Be forthcoming about failure modes and population characteristics in training/evaluation sets
	Reproducibility checks	Journals should ensure that submitted materials are sufficient for replication
Evaluation	Addressed in a separate report from the AI Task Force	

*not all recommendations are applicable to all types of studies

Table 3. Resources for hosting and sharing code, models, and data.

Data type	Resources
Code	Git repository hosts (Github, Gitlab, Bitbucket), Matlab File Exchange, SourceForge
Models, containers, executables	Docker Hub, modelhub, Model Zoo, ModelDepot, TensorFlow Hub, PyTorch Hub, Hugging Face
Data	The Cancer Imaging Archive, Kaggle, paperswithcodes.com, LONI Image and Data Archive, Figshare

FIGURES

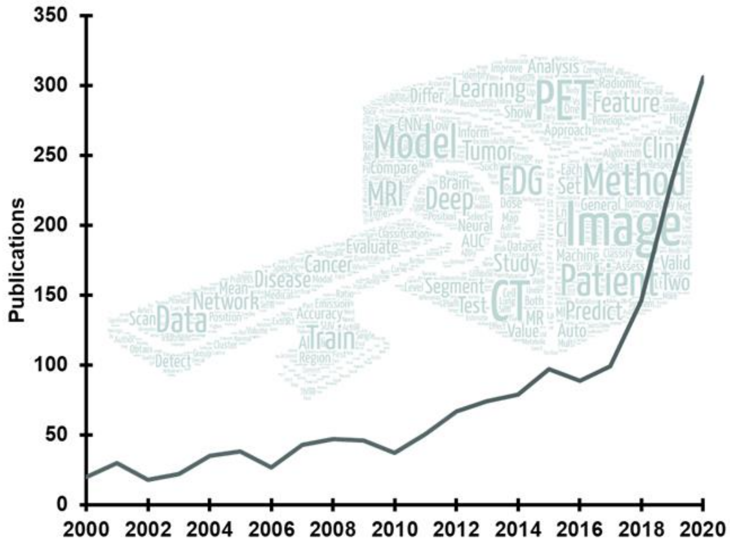


Figure 1. The trend in publications on artificial intelligence within nuclear medicine according to Scopus. The word cloud contains the most commonly-used terms in recent abstracts.

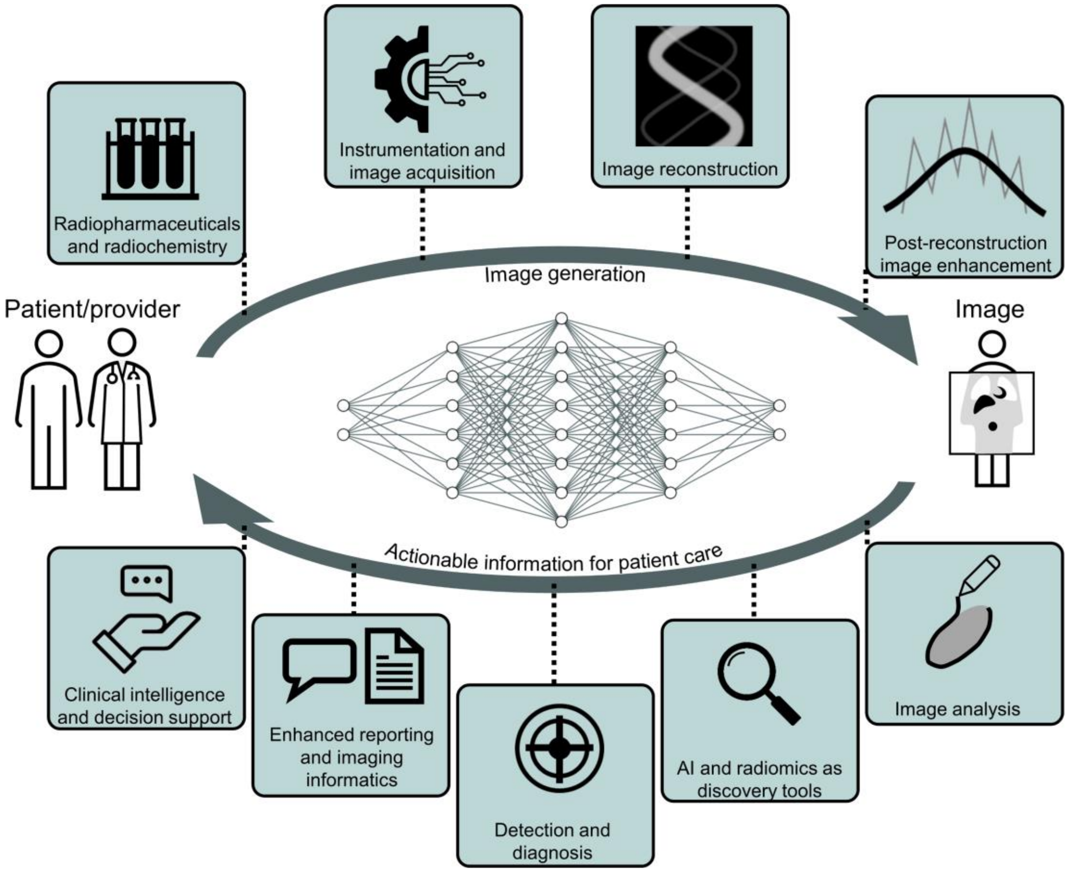

Radiopharmaceuticals and radiochemistry
Instrumentation and image acquisition
Image reconstruction
Post-reconstruction image enhancement
Image generation
Patient/provider
Image
Actionable information for patient care
Clinical intelligence and decision support
Enhanced reporting and imaging informatics
Detection and diagnosis
AI and radiomics as discovery tools
Image analysis

Figure 2. Applications of AI span the gamut of nuclear medicine subspecialties.

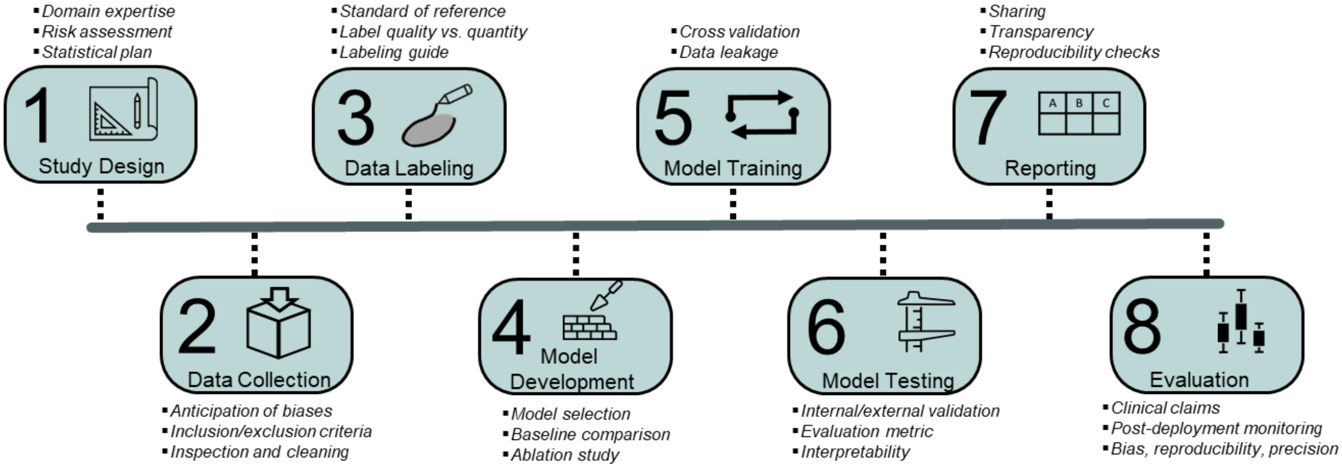


Figure 3. The pipeline for AI algorithm development together with key considerations of each stage of development.

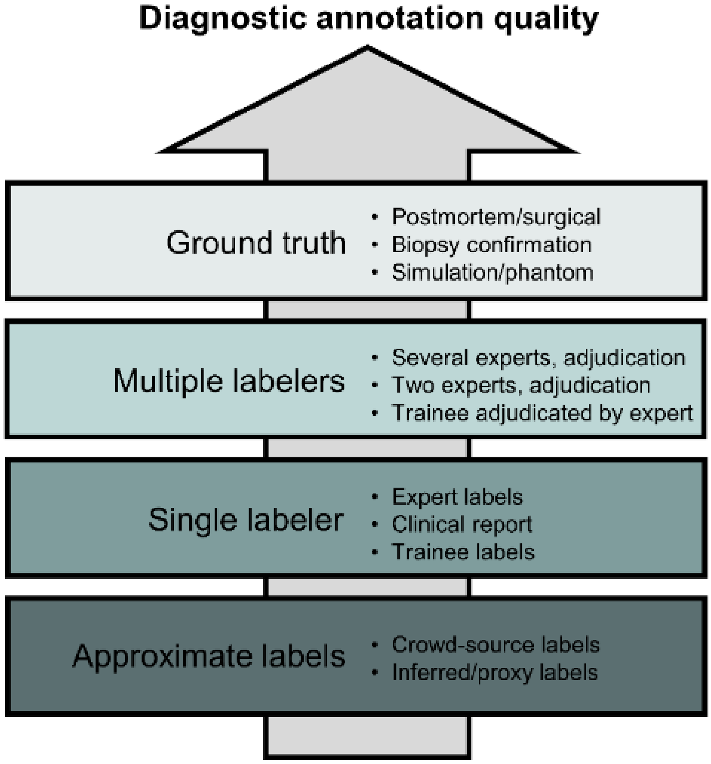


Figure 4. Annotation quality as a function of different labeling techniques for diagnostic applications. This hierarchy does not imply how useful an annotation method is (e.g., expert labels are often more useful than simulations due to the limited realism of simulated data).

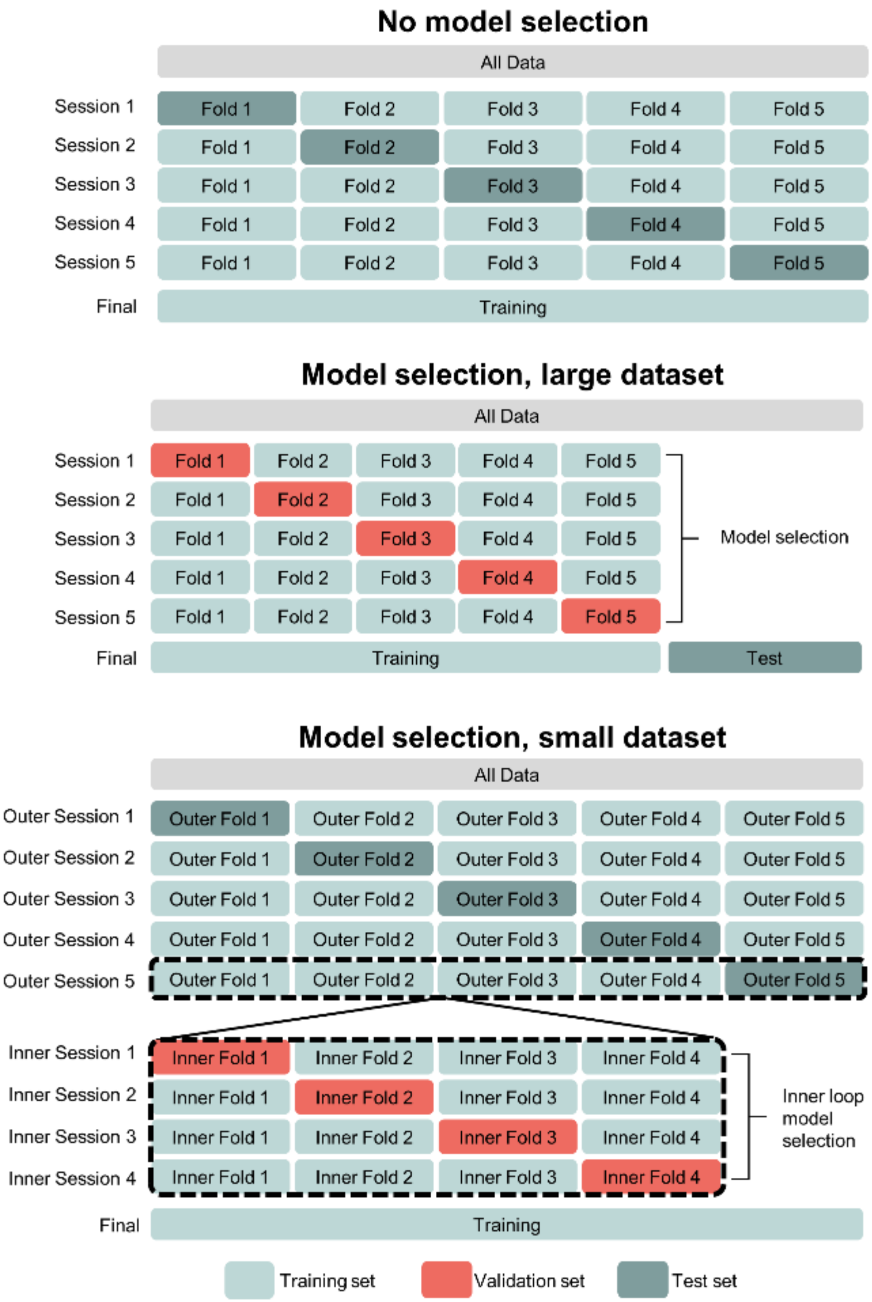

No model selection
All Data
Session 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Final Training
Model selection, large dataset
All Data
Session 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Session 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Model selection
Final Training Test
Model selection, small dataset
All Data
Outer Session 1 Outer Fold 1 Outer Fold 2 Outer Fold 3 Outer Fold 4 Outer Fold 5
Outer Session 2 Outer Fold 1 Outer Fold 2 Outer Fold 3 Outer Fold 4 Outer Fold 5
Outer Session 3 Outer Fold 1 Outer Fold 2 Outer Fold 3 Outer Fold 4 Outer Fold 5
Outer Session 4 Outer Fold 1 Outer Fold 2 Outer Fold 3 Outer Fold 4 Outer Fold 5
Outer Session 5 Outer Fold 1 Outer Fold 2 Outer Fold 3 Outer Fold 4 Outer Fold 5
Inner Session 1 Inner Fold 1 Inner Fold 2 Inner Fold 3 Inner Fold 4
Inner Session 2 Inner Fold 1 Inner Fold 2 Inner Fold 3 Inner Fold 4
Inner Session 3 Inner Fold 1 Inner Fold 2 Inner Fold 3 Inner Fold 4
Inner Session 4 Inner Fold 1 Inner Fold 2 Inner Fold 3 Inner Fold 4
Inner loop model selection
Final Training
Training set
Validation set
Test set

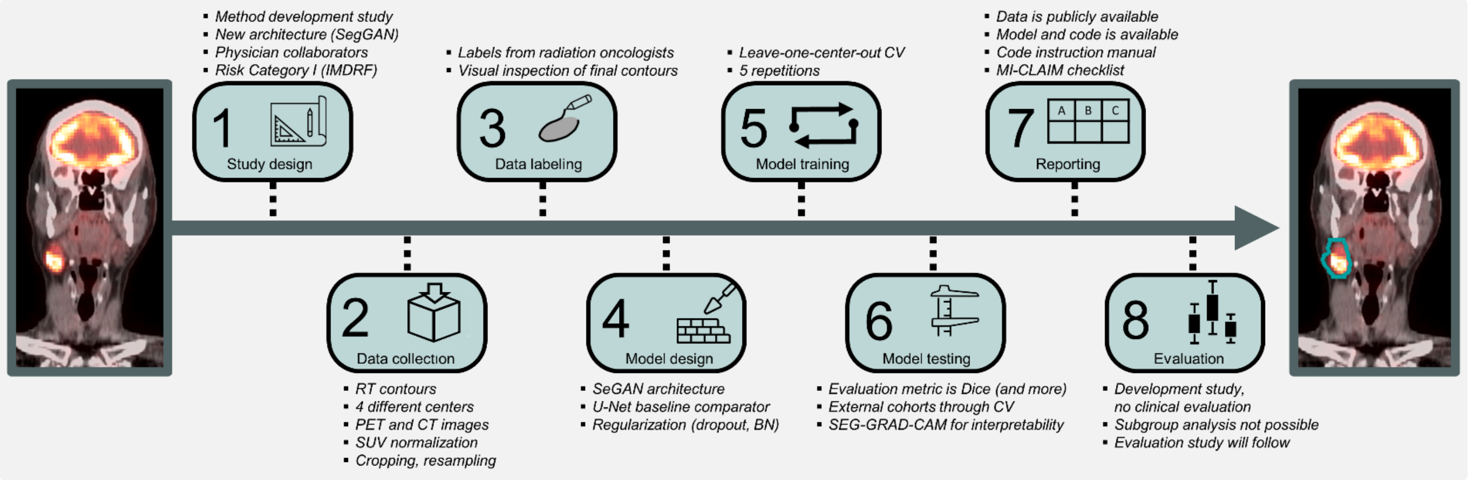
Figure 5. Different approaches to cross validation (CV), depending on the dataset size and if model selection is needed. The figure illustrates 5-fold CV without model selection/hyperparameter tuning (top), 5-fold CV with a holdout test set (middle), and nested CV (5-fold outer loop, 4-fold inner loop).

Supplemental Data

Nuclear Medicine and Artificial Intelligence: Best Practices for Algorithm Development

Example AI Application: Deep learning-based segmentation of PET-CT head and neck cancer images

This section walks readers through an example of machine learning being applied to a specific use case: automatic segmentation of oropharyngeal tumors in PET/CT images. This is a hypothetical illustration of how the different steps of algorithm development would be followed using a real-world dataset, while adhering to the recommendations listed in Table 2.



1 Study design
▪ Method development study
▪ New architecture (SegGAN)
▪ Physician collaborators
▪ Risk Category I (IMDRF)
2 Data collection
▪ RT contours
▪ 4 different centers
▪ PET and CT images
▪ SUV normalization
▪ Cropping, resampling
3 Data labeling
▪ Labels from radiation oncologists
▪ Visual inspection of final contours
4 Model design
▪ SeGAN architecture
▪ U-Net baseline comparator
▪ Regularization (dropout, BN)
5 Model training
▪ Leave-one-center-out CV
▪ 5 repetitions
6 Model testing
▪ Evaluation metric is Dice (and more)
▪ External cohorts through CV
▪ SEG-GRAD-CAM for interpretability
7 Reporting
A B C
▪ Data is publicly available
▪ Model and code is available
▪ Code instruction manual
▪ MI-CLAIM checklist
8 Evaluation
▪ Development study, no clinical evaluation
▪ Subgroup analysis not possible
▪ Evaluation study will follow

Supplemental Figure 1. The segmentation algorithm development pipeline.

1. Study Design

Task definition: The algorithm's task was to efficiently and reproducibly create segmentation masks from FDG PET/CT images of oropharyngeal tumors. The contours are to be used in image quantification for prognostication (e.g., SUV, radiomics). The study was designed with input from medical physicists with experience in image analysis, a nuclear medicine physician, and a radiation oncologist, all of whom contributed to the study design and provided extensive feedback throughout the development process. The motivation for the study is to improve outcomes for patients with head and neck cancers. Segmentation of PET imaging data can produce quantitative radiomics features that might predict tumor recurrence. Segmentation is an important step for radiomics extraction. However, manual segmentation is both time-consuming and subject to high inter-observer variability.

Study type: The study was a method development study. The study aimed to demonstrate the promise of a novel AI architecture and report on the method's segmentation performance. As a method development study, it makes no claims to the performance of the method on a clinical task (ie, predicting outcomes), but instead focuses on its technical performance. The hypothesis of the study was that automatic segmentation using a novel AI model would be non-inferior to manual physician segmentations in terms of accuracy, with the expected advantages of greater reproducibility and faster analysis. The novelty of this development study is in the application of a previously-developed AI architecture to a new task: segmentation of PET/CT images. In the current segmentation study, SegAN, a 3D generative adversarial network (GAN) based network developed by Xue *et al.* (*4*), was used for PET/CT segmentation and a combination of mean absolute error and Mumford-Shah loss function was used during training. The resulting mask is then used as the input to a localized active contour model to compute the energy function based on the region-based hybrid information of both PET and CT channels, which refines the SegAN-produced mask. This method has not been previously used for PET/CT segmentation.

Risk assessment: The use case for this algorithm would be as an input to a prognostic or predictive model to help guide treating physicians. According to the IMDRF risk categories, the application would likely belong to Category I: it would serve as input to a system that "informs patient management" with patients in "serious" condition, and would therefore be considered low risk.

Statistical plan: As an observational study, pre-registration of a statistical analysis plan is not required, but it can prevent "p-hacking" and help ensure that the study results are credible. For this study, a statistical analysis plan was designed, which described the expected metrics and statistical tests to be performed. As an optional step, prior to the analysis of the study, the document was uploaded to the academic research data management and dissemination tool figshare.com with a timestamp and assigned a digital object identifier (DOI) number. Any unplanned analyses were stated as being unplanned in the publication. While these steps are not required they do reflect best practices in science and can lend credibility and transparency to a study.

2. Data Collection

The dataset used in this study was originally described by Vallières et al. (*5*) and is publicly available on The Cancer Imaging Archive (TCIA) and provided by HEad and neCK TumOR (HECKTOR) (*6*) challenge organizers with the updated delineations in the 23rd MICCAI conference. It consists of pre-treatment FDG-PET and CT images of 201 cases with head-and-neck cancer acquired for initial staging. The images were gathered within a median of 18 days (range: 6–66) before treatment from four institution in Quebec: Centre Hôpital général juif (CHGJ) de Montréal (55 cases), Centre Hôpital Maisonneuve-Rosemont (CHMR) de Montréal (18 cases), Centre hospitalier de l'Université de Montréal (CHUM) (56 cases) and Centre hospitalier universitaire de Sherbrooke (CHUS) (72 cases). Scanner and imaging protocols are described in (*6*). According to (*5*), this study was limited to patients with head and neck squamous cell carcinoma (HNSCC) who planned to receive radiation or chemo-radiation with curative intent and excluded patients with recurrent head and neck cancer, those with metastases, and those receiving palliative treatment.

Bias anticipation: Biases in segmentation algorithms are most likely to arise from clinical factors (size, location, stage, shape, and extent of disease), technical factors (acquisition protocol, tracer uptake time, scanner model, reconstruction settings) and demographic factors. The dataset came from a variety of institutions with different scanners and data acquisition protocols and reconstruction methods, although biased towards equipment by GE Healthcare. The lack of data from other established manufacturers and the lack of data from latest scanner generations may limit the generalizability of the results. According to the Supplemental Data provided by (*5*), datasets were biased towards males (~75%), indicating that subgroup analysis according to gender should be performed. The patient data also all originated from Quebec, suggesting that results may not be applicable to other locales. The dataset did represent different TNM stages, with a bias towards later stages (III and IV). There is also a potential technical bias introduced by the inconsistent techniques used to label images: some tumors (~40%) were labeled on PET/CT images, while others were labeled on contrast enhanced CT images and then registered to PET/CT images (described below).

Training set size estimation: Training set size was selected based on trial-and-error and previous published studies of HN tumor segmentation. In our study, we use a sample size of 201.

Evaluation set size estimation: Not applicable, as cross validation was used.

Data decisions: The inclusion/exclusion criteria were determined by the investigators that collected and made the data available (*5,6*) and have been described at the beginning of this section. No additional criteria were enforced for this study.

Curation: The pre-processing steps of the segmentation study are as follows:

- DICOM images were converted to NIFTI format.
- PET/CT images were deformably registered to the planning (contrast enhanced) CT using commercial software (MIM Software Inc., Cleveland, OH).
- Resampling the CT and PET images to an isotropic 1×1×1mm voxel spacing using trilinear interpolation.
- The CT volumes are clipped in the specific range of Hounsfield Units (HU) [−150, 150].
- Cropping the head and neck data to a volume of size 144×144×144 voxels containing all tumors (primary oropharynx tumors and metastatic lymph nodes). Cropping is semi-random to prevent the tumor from always appearing in the middle, and was accomplished by randomly shifting the bounding box relative to the center point of the tumor, but still keeping all tumor edges within the bounding box. Shifts ranged from 1 to 10 voxels.
- Visual confirmation of the image quality and cropping.
- Conversion to standardized uptake values (SUV) normalized by body mass.

3. Data Labeling

Reference standard: Contours drawn on PET/CT images by board certified nuclear medicine physicians or radiologists are generally considered standards of reference for PET quantification. However, for applications involving outcome prediction following radiotherapy, planning contours drawn by board certified radiation oncologists are also generally considered as acceptable reference standards, especially for a development/proof-of-concept study.

Label quality: The labels used in this development study were gross tumor volumes (GTVs) drawn by radiation oncologists for radiotherapy planning. The original publications of the dataset (*5*) do not state the number, credentials, or experience of the radiation oncologists performing the labeling, which should normally be reported in AI studies. The lack of this information should be considered as a limitation of this study.

- Contours (GTVs) were drawn on CT images (either the CT of the PET/CT exam or the planning CT) using various treatment planning software platforms and then registered to the PET image using MIM software.
- Labels were manually adapted for quantification when appropriate. For example, air and other surrounding non-tumor tissue that often gets included within the GTV were removed, and the quality of the contour propagation following registration was visually inspected and corrected, if needed.
- All final registered contours and images were visually inspected and approved by an experienced nuclear medicine physician.

The quality of the labels could be improved by having multiple experts independently contour the lesions and perform a majority vote, or by having multiple experts come to a consensus. Multiple labelers would also allow for quantification of inter-observer variability, which is a reflection of overall label quality. However, as a method development study, the quality of the labels is of sufficient quality to demonstrate the promise of the method, especially when used for quantification alone (as opposed to being used for radiation treatment planning).

Labeling guide: No labeling guide was developed due to the retrospective nature of the dataset.

Quantity/quality tradeoff: The retrospective nature of this dataset prevented an *a priori* evaluation of the tradeoff between quality and quantity.

4. Model Development

Model comparison: An end-to-end deep learning framework, SegAN (*4*) was previously designed for medical image segmentation based on adversarial learning. We used the output of the SegAN as input to an active contour model. Our method was compared to a standard 3D U-Net model.

Baseline comparison: All deep learning models were compared to a 40% threshold of maximum SUV.

Model selection: For the CNN architectures in the SegAN framework, batch normalization layers were used to increase network convergence and dropout layers were used to avoid overfitting. The hyperparameters of the SegAN network were set to match the original SegAN publication (*4*). The hyperparameters of the U-Net architecture, including the number of layers, number of convolutional filters at each layer, and learning rate were determined using a grid search approach with a preliminary one-time random split (60:40) of the entire dataset. This preliminary split did not correspond to the data splits used during model training with cross validation, and will therefore not lead to overfitting. Ten U-Net models were trained and compared using the preliminary data split. The hyperparameters from the best performing U-Net model were then used throughout each fold of cross validation.

Model stability: Repetitions of 5 cross-validation runs were performed to assess model stability. The variation between runs is expected due to networks' random weight initializations and shuffling of the training samples.

Ablation study: The SegAN + active contour model was compared to the SegAN model alone to determine the added value of the active contour component.

5. Model Training

Cross validation: The model was trained using leave-one-center-out cross-validation. One center was used as the test set, one center as the validation set (which determined the number of epochs with which to train the model) and the remaining centers as the training set. A repetition of 5 cross-center-validation runs was performed to assess model stability. The U-Net model was also trained with the same cross validation scheme. A binary cross entropy loss function was used.

Data leakage: The use of cross validation prevents data leakage.

6. Model Testing and Interpretability

Test set: Cross-center-validation was used to evaluate the generalization of the segmentation algorithms to unknown centers. Each center served as an external test set for each fold of CV.

Targeted population: The targeted population in this application is patients in Quebec with histologically proven head-and-neck cancer prior to radiation therapy, excluding patients with recurrent disease or receiving palliative treatment.

External sets: No test sets that were external to the developmental dataset were used, although the leave-one-center-out cross validation approach does treat each center as an external to the training data set.

Evaluation metric: The DSC (Dice similarity coefficient) was used as the primary evaluation metric. Values were averaged across all test patients for each fold and repetition of CV. The structural similarity index measure (SSIM), Jaccard coefficients (JSC), Hausdorff distance (HD) measures and tumor volume error were also reported.

Model interpretability: We used gradient-weighted class activation mapping to determine which parts of the image contributed to the pixel-classification decisions of the network. Heatmaps were inspected and reported with example results.

7. Reporting and Dissemination

Reporting: When publishing results of the study, we follow the MI-CLAIM checklist (*27*) for reporting AI studies.

Sharing: All codes and trained models, with instructions on how to implement them, are posted to a public git repository prior to submission for publication. Examples, with preprocessing steps, are provided to ensure correct implementation and reproducibility.

Transparency: We visually inspected the 25 cases with the lowest model performance to identify any patterns or failure modes. Overall, we found that small, low uptake lesions less than 5 ml in volume tended to be missed by the model (i.e, the model returned no segmentation mask). When publishing, we display some of these cases.

Reproducibility checks: We followed the checklist used by the NeurIPS conference (*33*).

8. Evaluation

Clinical evaluation of the final trained model was not performed as this study was a method development study that did not aim to make any claims about the clinical performance of the algorithm. Given the promising results, evaluation studies will follow. Evaluation studies will require clinical task-specific evaluation (i.e., outcome prediction) using external datasets and involving a team of domain experts (i.e., physicians) to help properly select the appropriate evaluation dataset and evaluation strategy.