An Investigation of Lesion Detection Accuracy for Artificial Intelligence–Based Denoising of Low-Dose ^{64}Cu-DOTATATE PET Imaging in Patients with Neuroendocrine Neoplasms

Mathias Loft*[1,2], Claes N. Ladefoged*[1], Camilla B. Johnbeck[1,2], Esben A. Carlsen[1,2], Peter Oturai[1,2], Seppo W. Langer[2–4], Ulrich Knigge[2,5], Flemming L. Andersen[1], and Andreas Kjaer[1,2]

[1]*Department of Clinical Physiology and Nuclear Medicine & Cluster for Molecular Imaging, Copenhagen University Hospital–Rigshospitalet & Department of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark;* [2]*ENETS Neuroendocrine Tumor Center of Excellence, Copenhagen University Hospital–Rigshospitalet, Copenhagen, Denmark;* [3]*Department of Oncology, Copenhagen University Hospital–Rigshospitalet, Copenhagen, Denmark;* [4]*Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; and* [5]*Departments of Clinical Endocrinology and Surgical Gastroenterology, Copenhagen University Hospital–Rigshospitalet, Copenhagen, Denmark*

Frequent somatostatin receptor PET, for example, ^{64}Cu-DOTATATE PET, is part of the diagnostic work-up of patients with neuroendocrine neoplasms (NENs), resulting in high accumulated radiation doses. Scan-related radiation exposure should be minimized in accordance with the as-low-as-reasonably achievable principle, for example, by reducing injected radiotracer activity. Previous investigations found that reducing ^{64}Cu-DOTATATE activity to below 50 MBq results in inadequate image quality and lesion detection. We therefore investigated whether image quality and lesion detection of less than 50 MBq of ^{64}Cu-DOTATATE PET could be restored using artificial intelligence (AI). **Methods:** We implemented a parameter-transferred Wasserstein generative adversarial network for patients with NENs on simulated low-dose ^{64}Cu-DOTATATE PET images corresponding to 25% ($PET_{25\%}$), or about 48 MBq, of the injected activity of the reference full dose ($PET_{100\%}$), or about 191 MBq, to generate denoised PET images (PET_{AI}). We included 38 patients in the training sets for network optimization. We analyzed PET intensity correlation, peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and mean-square error (MSE) of $PET_{AI}/PET_{100\%}$ versus $PET_{25\%}/PET_{100\%}$. Two readers assessed Likert scale–defined image quality (1, very poor; 2, poor; 3, moderate; 4, good; 5, excellent) and identified lesion-suspicious foci on PET_{AI} and $PET_{100\%}$ in a subset of the patients with no more than 20 lesions per organ ($n = 33$) to allow comparison of all foci on a 1:1 basis. Detected foci were scored (C_1, definite lesion; C_0, lesion-suspicious focus) and matched with $PET_{100\%}$ as the reference. True-positive (TP), false-positive (FP), and false-negative (FN) lesions were assessed. **Results:** For $PET_{AI}/PET_{100\%}$ versus $PET_{25\%}/PET_{100\%}$, PET intensity correlation had a goodness-of-fit value of 0.94 versus 0.81, PSNR was 58.1 versus 53.0, SSIM was 0.908 versus 0.899, and MSE was 2.6 versus 4.7. Likert scale–defined image quality was rated good or excellent in 33 of 33 and 32 of 33 patients on $PET_{100\%}$ and PET_{AI}, respectively. Total number of detected lesions was 118 on $PET_{100\%}$ and 115 on PET_{AI}. Only 78 PET_{AI} lesions were TP, 40 were FN, and 37 were FP, yielding detection sensitivity (TP/(TP+FN)) and a false discovery rate (FP/(TP+FP)) of 66% (78/118) and 32% (37/115), respectively. In 62% (23/37) of cases, the FP lesion was scored C_1, suggesting a definite lesion. **Conclusion:** PET_{AI} improved visual similarity with $PET_{100\%}$ compared with $PET_{25\%}$, and PET_{AI} and $PET_{100\%}$ had similar Likert scale–defined image quality. However, lesion detection analysis performed by physicians showed high proportions of FP and FN lesions on PET_{AI}, highlighting the need for clinical validation of AI algorithms.

Key Words: ^{64}Cu-DOTATATE; somatostatin receptor imaging; PET/CT; neuroendocrine neoplasms; artificial intelligence

J Nucl Med 2023; 64:951–959
DOI: 10.2967/jnumed.122.264826

Neuroendocrine neoplasms (NENs) are rare diseases that originate from the diffuse neuroendocrine system. PET based on radiotracers targeting the somatostatin receptor (SSR), overexpressed in most NENs, plays a fundamental role in the clinical management of diagnosis, staging, treatment guidance, and follow-up of patients with NENs (*1–4*). Patients may undergo lifelong annual or biannual follow-up with inclusion of SSR-based PET/CT imaging (*3*), resulting in relatively high accumulated radiation exposure that underscores the importance of adhering to the as-low-as-reasonably achievable principle (*5*).

The U.S. Food and Drug Administration–approved activity dose of the SSR PET radiotracer ^{64}Cu-DOTATATE is 148 MBq, with an effective radiation dose of 4.7 mSv (*6*). One way to reduce the PET-related radiation burden is by reducing the radiotracer activity dose. By analyzing simulated dose-reduced PET images, we previously demonstrated that the injected ^{64}Cu-DOTATATE activity could be reduced to approximately 100 MBq without loss of clinically relevant information (*7*). With activity dose reduction to less than 50 MBq, image quality was suboptimal and lesion detection sensitivity was low.

Deep learning (DL), a subtype of artificial intelligence (AI), has recently been proposed as a tool for low-count PET image noise reduction (*8*), because it has been shown to outperform conventional denoising methods while retaining lesion detectability and quantitative accuracy in oncologic PET (*9,10*). However, limited contrast recovery has been observed for smaller lesions (<1 cm^3), which challenges the use of DL methods when lesion detectability is important for clinical diagnosis.

When evaluating the performance of AI methods in medical imaging, discrepancies may arise between conventional fidelity-based metrics, for example, structural similarity index (SSIM) and mean-square error (MSE), and objective clinical task–based metrics. For example, the application of a denoising DL algorithm on simulated low-dose SPECT

Received Aug. 24, 2022; revision accepted Jan. 31, 2023.
For correspondence or reprints, contact Andreas Kjaer (akjaer@sund.ku.dk).
*Contributed equally to the work.
Published online May 11, 2023.

images in a phantom study by Yu et al. (*11*) did not improve the signal detection task despite showing improvements in fidelity-based metrics. Similarly, using a denoising DL algorithm to augment low-dose SPECT myocardial perfusion scintigraphy images, Yu et al. (*12*) found poor performance in the detection of myocardial defects, whereas the fidelity-based metrics were improved. Discrepancies are not limited to denoising algorithms. Yang et al. (*13*) found that implementation of a DL algorithm for CT-less attenuation correction of ^{18}F-FDG PET/CT images from oncologic patients resulted in false-negative (FN) lesions and the appearance of false-positive (FP) lesions when the DL PET images were reviewed by radiologists, even though convincing fidelity-based metrics were found. As highlighted in the recently published Recommendations for Evaluation of Artificial Intelligence for Nuclear Medicine (RELAINCE) guidelines (*14*), it is therefore essential to include evaluation of relevant clinical tasks early in the development of the algorithms and to not rely solely on fidelity-based metrics.

In the current study, we evaluated to what extent application of a DL-based model could assist in reducing the image noise of suboptimal, low-dose ^{64}Cu-DOTATATE PET images while retaining finer image structures such as tumor lesions. The clinical goal of SSR PET imaging is to ensure correct lesion detection, disease classification, and staging of patients with NENs. In accordance with the RELAINCE guidelines (*14*), we therefore evaluated the clinical task of detecting tumor lesions on denoised, low-dose PET images from patients with NENs, in addition to evaluation of the Likert scale–defined image quality and conventional fidelity-based metrics.

MATERIALS AND METHODS

Patient Population

The study is a continuation of our previously reported activity dose reduction PET investigation performed in patients with NENs (*7*). We retrospectively included the same 38 patients with NENs referred to a

TABLE 1
Patient Characteristics

Characteristic	Data (n = 38)	Subset for clinical image analysis (n = 33)*
Sex		
Female	21 (55)	19 (58)
Male	17 (45)	14 (42)
Age (y)		
Median	64	64
Range	37–84	37–84
Site of primary tumor		
Small intestine	21 (55)	16 (49)
Pancreas	11 (29)	11 (33)
Lung	3 (8)	3 (9)
Other	3 (8)	3 (9)
Previous treatment†		
Surgery	29 (76)	27 (82)
Somatostatin analogs	23 (61)	18 (55)
Peptide receptor radionuclide therapy	12 (32)	8 (24)
Chemotherapy	10 (26)	7 (21)
Radiofrequency ablation (liver metastases)	2 (5)	2 (6)
Ki-67 proliferation index		
<3%	9 (24)	8 (24)
3%–20%	26 (68)	22 (67)
>20%	3 (8)	3 (9)
Dose (MBq)‡		
$PET_{100\%}$	191 (169–209)	191 (172–209)
$PET_{25\%}$/PET_{AI}	48 (42–52)	48 (43–52)

*Patients with >20 lesions per organ (n = 5) were excluded for clinical image analysis. Patients used for clinical image analysis (n = 33) thus represent subset of all 38 patients included in training sets.

†Some patients received multiple treatments. Therefore, total number of treatments exceeds number of patients.

‡Dose at $PET_{100\%}$ is ^{64}Cu-DOTATATE activity dose given to patient for PET/CT. $PET_{25\%}$ and PET_{AI} dose is derived from simulated equivalent dose at 25% of $PET_{100\%}$ dose.

Data are number followed by percentage in parentheses, except for age and dose (median and range).

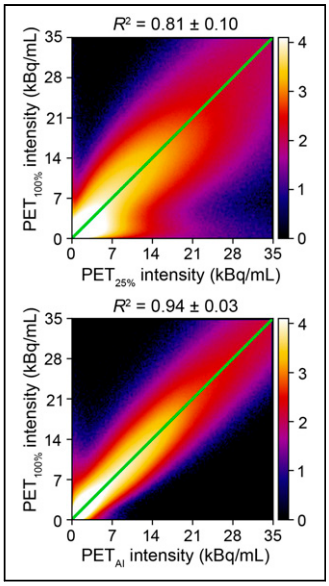

FIGURE 1. Joint histogram of PET intensity values for $PET_{25\%}$ (top) and PET_{AI} (bottom) versus reference $PET_{100\%}$. Green line is identity line, and R^2 is shown above each image. Analysis was performed on training sets (n = 38).

routine ^{64}Cu-DOTATATE PET/CT at the Department of Clinical Physiology and Nuclear Medicine, Copenhagen University Hospital–Rigshospitalet, between April and September 2019 with PET list-mode data available. The study was approved by the Danish Patient Safety Authority (reference 31–1521–453) according to Danish regulations, and the requirement to obtain written informed consent was waived.

PET/CT Acquisition and Image Reconstruction

PET/CT acquisition, PET reconstruction, and generation of reduced-dose PET equivalents were performed as previously described (*7*). Patients were injected with approximately 200 MBq of ^{64}Cu-DOTATATE based on our clinical studies (*15–18*). PET acquisition was performed approximately 1 h later with a Siemens Biograph 128 mCT PET/CT scanner with an axial field of view of 221 mm and a 4 min/bed position acquisition time. A standard routine whole-body diagnostic CT imaging series was performed. Simulated low-dose ^{64}Cu-DOTATATE PET images corresponding to 25% ($PET_{25\%}$) of the injected activity of the reference full dose ($PET_{100\%}$) were generated by randomly deleting events in the PET list-mode file using LMChopper (e7-tools; Siemens Healthineers). We created 5 realizations of the $PET_{25\%}$ images. This was done to increase the number of training samples and perform data augmentation because of noise variation among the realizations. Reconstruction of both $PET_{100\%}$ and $PET_{25\%}$ was performed using 3-dimensional (3D) ordinary Poisson ordered-subset expectation maximization with 2 iterations and 21 subsets, including time of flight at 540 ps and modeling of the point spread function, followed by smoothing by a gaussian postprocessing filter at 2 mm full width at half-maximum. The reconstructed image size was 400 × 400 × 426 voxels with a voxel size of 2.04 × 2.04 × 2.00 mm^3.

PET Image Preprocessing

$PET_{25\%}$ images were first cropped to 256 × 256 × 426 voxels to minimize the effect of voxels outside of the body. We extracted patches of 64 × 64 × 9 voxels with a stride of 9 voxels in each direction, excluding patches with maximum PET or CT values that were less than empirically selected thresholds (<10 Bq/mL or <−200 HU, respectively) to limit empty patches. A total of 762,338 patches were extracted for each of the 5 noise realizations across the 38 patients.

Network Setup and Training

To generate the denoised PET images (PET_{AI}), we implemented a parameter-transferred Wasserstein generative adversarial network (PT-WGAN) for low-dose PET noise reduction inspired by Gong et al. (*19*). The network type was chosen because the authors demonstrated it had better performance than a pure 2-dimensional (2D) or 3D convolutional neural network on the same dataset. Supplemental Appendix A gives a more detailed overview (supplemental materials are available at http://jnm.snmjournals.org). In short, the PT-WGAN consists of 2 parts, a generator and a discriminator, where the generator is a hybrid 2D and 3D U-netlike network pretrained without the discriminator to improve stability and convergence during training. The hybrid combination was introduced by Gong et al. (*19*) to limit computational resources. The model training and evaluation were done using 5-fold cross-validation. In each fold, we first reserved a test set consisting of one fifth of the 38 patients for evaluation that was not part of model training for that fold. Next, we reserved 10% of the remaining four fifths of the data for validation during training (used to detect overfitting) and trained the model on the remaining patients. After the 5 repetitions, all 38 patients had at one point been in a test set, and a PET_{AI} image was therefore created. We did not vary any hyperparameters among the folds.

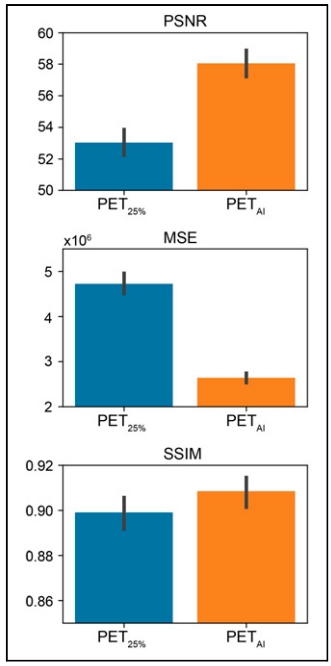

FIGURE 2. Image similarity metrics: PSNR (top), MSE (middle), and SSIM (bottom). Error bars mark 95% CI. Analysis was performed on training sets (n = 38).

Objective Visual Similarity Analysis

We evaluated the quantitative accuracy of $PET_{25\%}$ and PET_{AI} by computing a joint histogram of the PET activity relative to $PET_{100\%}$, and we compared the image fidelity using the following standard similarity comparison metrics: peak signal-to-noise ratio (PSNR), SSIM, and MSE. We restricted the comparison to voxels inside the patient volume defined using the CT image (more than −900 HU).

Clinical Image Analysis

Two readers placed side by side collectively analyzed all PET/CT scans: a board-certified nuclear medicine physician with 10 y and a nuclear medicine physician in training with 4 y of experience in reading SSR-based PET/CT scans from patients with NENs. To analyze all patients' lesions on a 1:1 basis, only PET images from a subset of patients with no more than 20 lesions in each organ system (n = 33), of the initially included 38 patients used for training, were used for the clinical image analysis. The readers were blinded to the PET image ($PET_{100\%}$ or PET_{AI}) and analyzed the images in 2 clusters, each containing either $PET_{100\%}$/CT or PET_{AI}/CT from 1 of the 33 patients, presented to the readers in random order. After 12 wk of quarantine, the second cluster was analyzed by the same readers. Mirada DBx 1.2.0 was used for the clinical analysis.

Likert Scale–Defined Image Quality

The image quality of the PET images was rated on a 5-point Likert scale: 1 (very poor), 2 (poor), 3 (moderate), 4 (good), and 5 (excellent). Scores 4 and 5 were accepted as diagnostic image quality.

Number and Certainty of Detected Lesions

On each PET image, any focus considered lesion-suspicious was annotated. The CT was used mainly to confirm the anatomic location of the PET focus. Each focus was given a certainty score for a definite lesion (C_1) and for a focus indicative of a lesion or a suspicious area (C_0), in which the presence of a lesion could not be ruled out. The images were then unblinded and the identified foci were matched on $PET_{100\%}$ and PET_{AI}. $PET_{100\%}$ was considered the standard of truth. Concordant, true-positive (TP) lesions identified on both $PET_{100\%}$ and PET_{AI} and

discordant lesions—FN lesions visible on $PET_{100\%}$ but not PET_{AI} and FP lesions visible on PET_{AI} but not $PET_{100\%}$—were grouped according to organs and regions. Organ- or region-specific and overall sensitivities and false discovery rates (FDRs) for detected lesions on PET_{AI} were calculated as TP/(TP+FN) and FP/(TP+FP), respectively, on a per-lesion basis. We evaluated the distribution of TP, FP, and FN lesions according to the number of lesions detected on $PET_{100\%}$ in the following groups: no lesions, 1 lesion, 2–5 lesions, 6–10 lesions, and more than 10 lesions. We also analyzed the per-patient sensitivities and specificities for the detection of organ- or region-specific and overall disease based on matched lesions on a per-patient basis, with $PET_{100\%}$ as the reference.

Patient Characteristics Based on Lesion Types

To analyze whether patient-specific characteristics contributed to the occurrence of FN and FP lesions, we compared patients with FN or FP lesions and patients with either TP-only or no lesions with the following variables: injected activity dose, weight, activity dose per weight, and liver background (SUV_{mean} measured in a 3-cm-diameter sphere in the right lobe of the liver in an area free of blood vessels and lesions).

Statistics

$PET_{100\%}$ was considered the standard of truth. For the clinical analysis, the proportion of PET images with Likert scale–defined image quality scores of good or excellent (considered diagnostic image quality) were analyzed with the McNemar test for paired proportions for PET_{AI} versus $PET_{100\%}$. The McNemar test was also used for analysis of the distribution of C_1 and C_0 lesion scores among TP lesions on PET_{AI} versus $PET_{100\%}$. For sensitivities, specificities, and FDR, 95% CI was calculated with the Clopper-Pearson exact method. For comparison of the patient-specific characteristics, we used Mann–Whitney *U* tests. Reference groups were patients with only TP lesions or no lesions for the patient-specific comparisons. R version 3.6.1 was used for the clinical statistical analysis. For comparison of the PET intensity correlations of $PET_{25\%}$ and $PET_{100\%}$ versus PET_{AI} and $PET_{100\%}$, we computed a goodness-of-fit value (R^2) to the identity line for each of the patients. Image fidelity metrics of $PET_{25\%}$ and $PET_{100\%}$ versus PET_{AI} and $PET_{100\%}$ were calculated with NumPy version 1.22.4 and scikit-image version 0.18.2 (*20*) in Python version 3.8.

RESULTS

Patient Characteristics

Characteristics of the patients are shown in Table 1.

Objective Visual Similarity Analysis

The AI algorithm was able to reduce the noise while improving the quantitative accuracy in the images (Fig. 1), resulting in better correlation with $PET_{100\%}$ for PET_{AI} (R^2 = 0.94) compared with $PET_{25\%}$ (R^2 = 0.81). The model increased PSNR and SSIM while decreasing MSE compared with $PET_{25\%}$ (Fig. 2).

Likert Scale–Defined Image Quality

Likert scale–defined image quality scores are shown in Figure 3. All $PET_{100\%}$ (33/33) and all but 1 PET_{AI} (32/33) had a Likert scale–defined image quality score of 4 (good) or 5 (excellent) and were thus considered diagnostic image quality. No statistically significant difference in the proportions of patients with diagnostic image quality PET was observed (P = 1.0). Figure 4 shows a representative example of the AI algorithm's ability to reduce noise and apparently restore the Likert scale–defined image quality of low-dose $PET_{25\%}$.

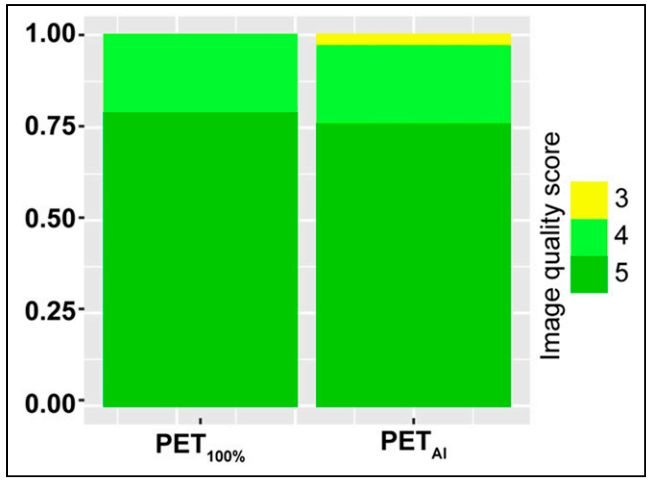

FIGURE 3. Distribution of Likert scale–defined image quality scores—3, moderate; 4, good; and 5, excellent (Likert scale–defined image quality scores 4 and 5 are considered diagnostic image quality)—on $PET_{100\%}$ and PET_{AI}. No patient had Likert scale–defined image quality score below 3. Analysis was performed on patient subset for clinical image analysis consisting of patients with ≤20 lesions per organ (n = 33).

Number of Detected Lesions

Table 2 shows the number of lesions detected on $PET_{100\%}$ and PET_{AI} grouped by organs and regions. The total number of lesions was similar on $PET_{100\%}$ and PET_{AI}, with 118 and 115 lesions detected, respectively. However, only 78 lesions were TP on PET_{AI}, yielding lesion detection sensitivity of 66% (78/118). In addition, 37 FP lesions were detected on PET_{AI}, corresponding to FDR of 32% (37/115). The same trend, with high rates of FP and

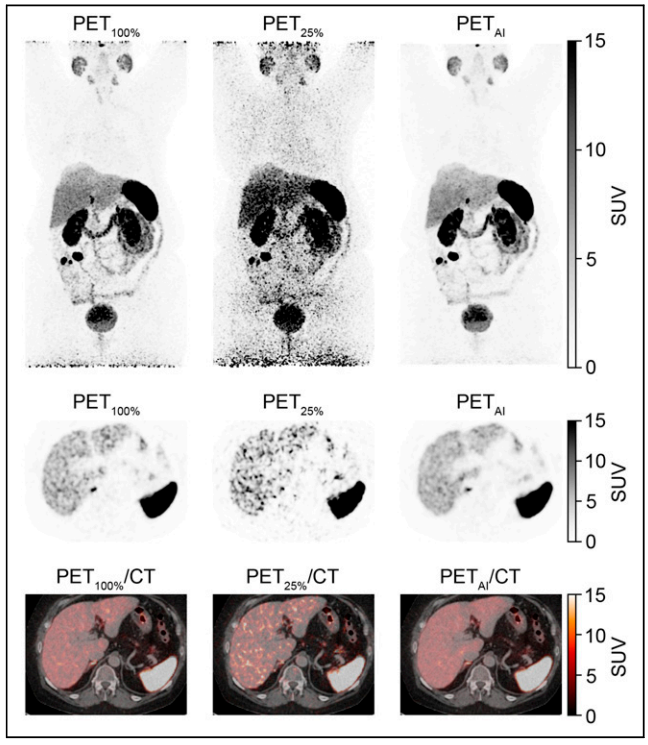

FIGURE 4. Examples of full-dose $PET_{100\%}$, low-dose $PET_{25\%}$, and denoised PET_{AI}.

TABLE 2
Number of Lesions Grouped by Organs and Regions in 33 Patients with NENs

Organ or region	No. of lesions $PET_{100\%}$	No. lesions PET_{AI}	TP	FP	FN	Sensitivity*	FDR*
Liver	36	38	17	21	19	47 (30–65)	55 (38–71)
Pancreas	6	7	6	1	0	100 (54–100)	14 (0–58)
Abdominal	49	47	36	11	13	73 (59–85)	23 (12–38)
Extraabdominal LNs	5	6	5	1	0	100 (48–100)	17 (0–64)
Bone	17	12	10	2	7	59 (33–82)	17 (2–48)
Other	5	5	4	1	1	80 (28–99)	20 (1–72)
Overall	118	115	78	37	40	66 (57–75)	32 (24–42)

*Data for sensitivity and FDR are percentages followed by 95% CI in parentheses.

Abdominal = intestines, intraabdominal carcinosis, and intraabdominal lymph nodes (LNs); other = brain (1), ovary (1), thyroid or parathyroid (1), and skin (2). Analysis is performed on patient subset for clinical image analysis consisting of patients with ≤20 lesions per organ (n = 33).

FN lesions yielding low lesion detection sensitivity and high FDR, was observed for the abdomen and liver. A representative example of a patient with a FN liver lesion is shown in Figure 5. This patient had additional TP liver lesions. Figure 6 shows a representative example of a patient with a FP lesion detected only on PET_{AI}. This was the only lesion detected on either of the scans, that is, no TP lesions. Figure 7 shows the distribution of TP, FP, and FN lesions according to the number of detected lesions on $PET_{100\%}$. Per-patient sensitivity and specificity for the detection of NEN disease across organs and regions are shown in Supplemental Table 1.

Certainty in Detected Lesions

The distributions of lesion certainty scores (C_1 and C_0) across organs and regions are shown in Table 3. Most TP lesions were given C_1 scores, suggesting that the readers were certain of the presence of a lesion on both $PET_{100\%}$ and PET_{AI}. For the FN lesions, larger fractions of C_0 lesions were observed on $PET_{100\%}$, suggesting that the readers were uncertain whether a suspicious focus indeed was a lesion in these cases. Of the 37 FP lesions detected only on PET_{AI}, 23 (62%) were given a score of C_1, suggesting that the readers were certain of the presence of a lesion.

Patient Characteristics Based on Lesion Types

Patient-specific characteristics are shown in Table 4. There was a trend of a lower weight-adjusted activity dose in the groups of patients with FP compared with the groups of patients with no lesions or only TP lesions, although this was not statistically significant.

DISCUSSION

Using randomly undersampled list-mode ^{64}Cu-DOTATATE PET data, we simulated low-dose PET images and implemented a state-of-the-art denoising PT-WGAN-based AI algorithm to test whether the image quality and lesion detection rate could be restored. Our main finding was that only 78 of 118 lesions could be detected on PET_{AI} (TP), and of 115 lesions detected on PET_{AI}, 37 were FP, corresponding to lesion detection sensitivity and FDR of 66% (78/118) and 32% (37/115), respectively. Despite the improvements of the fidelity-based metrics and the Likert scale–defined image quality performed by the AI algorithm, the discrepancies between $PET_{100\%}$ and PET_{AI} for the detection of correct lesions highlight the need for clinical validation when assessing the performance of AI algorithms.

According to the fidelity-based metrics, perceived Likert scale–defined image quality, and total number of detected lesions, the algorithm appeared successful in denoising low-dose $PET_{25\%}$. However, low lesion detection sensitivity on PET_{AI} shows that a large fraction of the lesions was not captured by the AI algorithm. Even more alarming was the high proportion of FP lesions observed only on PET_{AI}, yielding high FDR. For most of the 37 FP lesions, the readers assigned a C_1 certainty score, suggesting high certainty that the focus was indeed a lesion. Because the readers generally considered PET_{AI} to be of diagnostic image quality (Likert scale–defined image quality score of good or excellent), they may have been prone to accepting an apparent lesion-suspicious focus as a lesion without raising concern that

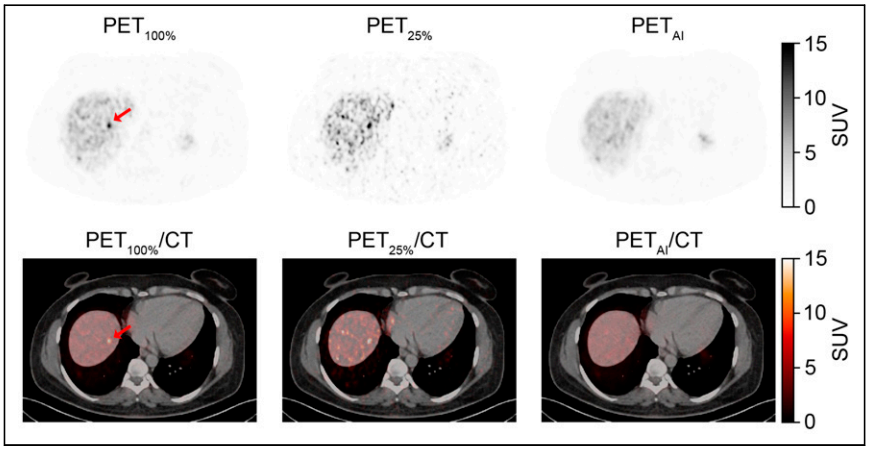

FIGURE 5. Patient with FN liver lesion. Patient had additional concordant TP liver lesions. Arrows mark lesion location on $PET_{100\%}$ and $PET_{100\%}$/CT. $PET_{25\%}$ shown for reference.

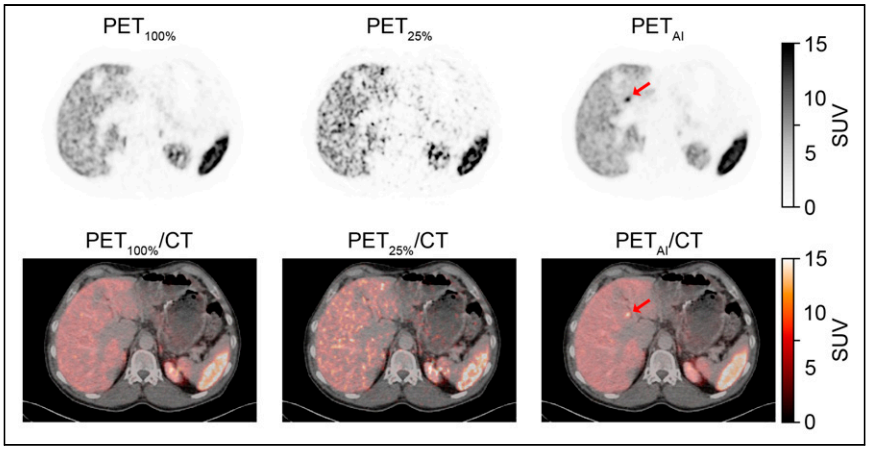

FIGURE 6. Patient with FP liver lesion on PET_{AI}. Patient had no lesions detected on $PET_{100\%}$. Arrows mark lesion location on PET_{AI} and PET_{AI}/CT. $PET_{25\%}$ shown for reference.

its appearance may result from the algorithm. Importantly, FP and FN lesions were not restricted to patients with multiple lesions on $PET_{100\%}$, in which case a single or a few FN or FP findings would have limited clinical consequences. FP and FN lesions were also found in patients with none or only 1 lesion detected on $PET_{100\%}$, in which case a single misclassified lesion could alter the patient's status from healthy to diseased, or vice versa. This was supported by low per-patient sensitivity and specificity for the presence or absence of disease across organs and regions based on matched lesions. These findings highlight the importance of focusing on the correct clinically relevant task when assessing AI algorithms, as recommended in the RELAINCE guidelines (*14*).

Compared with other advanced DL-based denoising studies on low-dose or fast-acquisition ^{18}F-FDG PET in oncologic patients who showed detection sensitivity of up to 97% (*21,22*), the detection sensitivity of our study was low. Without a comparative study, it is difficult to assess potential causes of this difference. However, the larger training cohorts, 313 patients in a study by Xing et al. (*21*) and 60 patients in a study by Sanaat et al. (*22*), could have an impact. Differences in the patients' tumor phenotypes, the physical properties of ^{18}F versus ^{64}Cu, or the biodistribution of the radiotracer may also contribute. Patients with NENs frequently have metastatic disease with multiple lesions scattered throughout the body. Metastases are often small (≤1 cm), which may impact the performance of the denoising algorithms. In line with this, Yu et al. (*11*) showed poor performance of a DL-based denoising algorithm for signal detection of small signal sizes of denoised low-dose SPECT images. In addition, the liver and the intestines are particularly lesion-prone in patients with NENs, and these organs have relatively high background uptakes of SSR-based radiotracers, making it difficult to distinguish potential lesions from surrounding tissue on low-dose PET. However, we did not find any difference in uptake of ^{64}Cu-DOTATATE in the normal liver of patients with FN or FP lesions compared with patients with no or TP-only lesions. In contrast, a trend of a lower weight-adjusted dose in the group with FP lesions was observed, which could contribute to the FP lesions because of increased image noise.

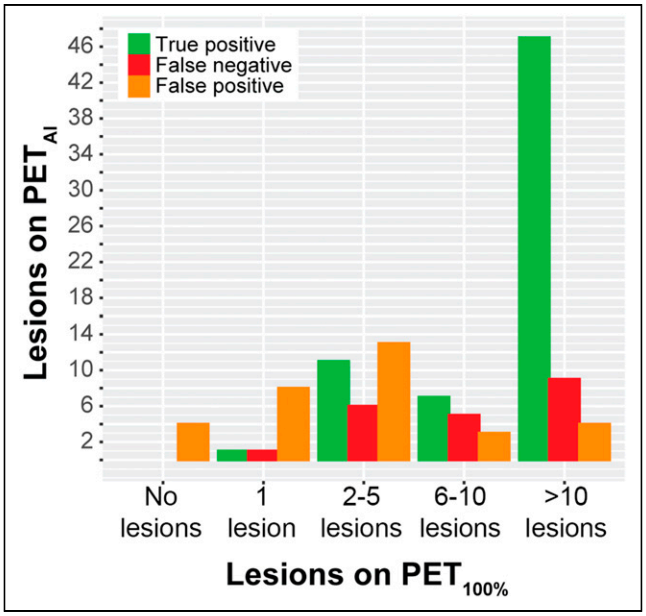

FIGURE 7. Distribution of TP, FP, and FN on PET_{AI} corresponding to number of lesions detected on $PET_{100\%}$. Analysis was performed on patient subset for clinical image analysis consisting of patients with ≤20 lesions per organ (n = 33).

If denoising AI algorithms of low-dose, whole-body SSR PET images are to be implemented clinically, the challenges of FP and FN lesions must be solved. Variations in DL training strategies, including choice of network architecture, loss function, and hyperparameters, might improve performance. We compared the effect of the network architecture by comparing PT-WGAN against traditional 3D U-netlike, residual 3D U-netlike, and adversarial network architectures and found the best performance with PT-WGAN (Supplemental Appendix B). Furthermore, we evaluated the influence of sampling patches over areas with high activity, and although this improved PSNR and MSE, there was no improvement on image structure measured with SSIM (Supplemental Appendix A). We speculate that further improvement might be achieved by incorporating a lesion-based loss term; however, this would require total tumor segmentation of the training patients and was not pursued in this work.

A limitation of the method is the low number of included patients, despite being comparable to other recently published studies of 9–31 patients (*10,19,23*). We chose to use a k-fold cross-validation training strategy to achieve a sufficient number of patients for evaluation, which is a frequently used technique to overcome a low number of patients. This is a significant limitation because clinical AI methods must be evaluated using an independent test set to show robustness and avoid potential data leakage. However, we would not expect lesion detection sensitivity and FDR to improve if tested on an independent test set. Rather, we speculate that the FP or FN findings may be even more pronounced. Although the 38 patients each contribute many data points during training because of the large, whole-body PET data files, these in turn are highly correlated with those extracted from neighboring areas. Inclusion of additional patients in the training sets may assist the AI algorithm in detecting the lesion patterns and may improve the performance. In addition, optimization

TABLE 3
Certainty of Detected Lesions in 33 Patients with NENs

Organ or region	All lesions			TP				FN			FP		
	Total	C_1	C_0	Total	C_1	C_0	*P**	Total	C_1	C_0	Total	C_1	C_0
Liver													
$PET_{100\%}$	36	31	5	17	17	0	1.0	19	14	5	N/A	N/A	N/A
PET_{AI}	38	29	9	17	17	0	N/A	N/A	N/A	N/A	21	12	9
Pancreas													
$PET_{100\%}$	6	6	0	6	6	0	1.0	0	0	0	N/A	N/A	N/A
PET_{AI}	7	7	0	6	6	0	N/A	N/A	N/A	N/A	1	1	0
Abdominal													
$PET_{100\%}$	49	45	4	36	35	1	1.0	13	10	3	N/A	N/A	N/A
PET_{AI}	47	43	4	36	36	0	N/A	N/A	N/A	N/A	11	7	4
Extraabdominal LNs													
$PET_{100\%}$	5	5	0	5	5	0	1.0	0	0	0	N/A	N/A	N/A
PET_{AI}	6	5	1	5	5	0	N/A	N/A	N/A	N/A	1	0	1
Bone													
$PET_{100\%}$	17	16	1	10	10	0	1.0	7	6	1	N/A	N/A	N/A
PET_{AI}	12	11	1	10	9	1	N/A	N/A	N/A	N/A	2	2	0
Other													
$PET_{100\%}$	5	5	0	4	4	0	1.0	1	1	0	N/A	N/A	N/A
PET_{AI}	5	5	0	4	4	0	N/A	N/A	N/A	N/A	1	1	0
Overall													
$PET_{100\%}$	118	108	10	78	77	1	0.5	40	31	9	N/A	N/A	N/A
PET_{AI}	115	100	15	78	77	1	N/A	N/A	N/A	N/A	37	23	14

*P values calculated using McNemar test for paired proportions of distribution of C_1 and C_0 lesion scores in TP lesions on $PET_{100\%}$ vs. PET_{AI}.

Abdominal = intestines, intraabdominal carcinosis, and intraabdominal lymph nodes (LNs); N/A = not applicable; other = brain (1), ovary (1), thyroid or parathyroid (1), and skin (2). Analysis is performed on patient subset for clinical image analysis consisting of patients with ≤20 lesions per organ (n = 33).

of the low-dose PET acquisition or reconstruction regime before running the AI algorithm may improve the performance.

It could be argued that a more comprehensive evaluation of the performance of the denoising algorithm, in terms of restoring lesion detection, could be obtained with a receiver-operating-characteristic analysis (*24*). For example, the detection of regional or organwise and overall ^{64}Cu-DOTATATE avid disease (yes or no), on a per-patient basis, could be performed with 5-point confidence scores (e.g., definitely normal, probably normal, unsure, probably malignant, or definitely malignant) for both $PET_{100\%}$ and PET_{AI}, using an external standard of truth for presence of disease, to compare disease detection performance as the areas under the receiver-operating-characteristic

TABLE 4
Characteristics of 33 Patients with NENs Based on Lesion Type

Parameter	TP-only or no lesions (n = 11)	FN (n = 16)	*P**	FP (n = 15)	*P**
Injected dose (MBq)	188 (181.5–201.5)	190.5 (184–198.9)	0.94	192.0 (184.0–195.6)	0.94
Weight (kg)	76.0 (67–81.5)	73.0 (64.3)	0.87	86.7 (74.0–97.5)	0.09
Dose/weight (MBq/kg)	2.5 (2.4–2.9)	2.6 (2.0–3.1)	0.82	2.3 (2.0–2.6)	0.07
Liver SUV_{mean}	5.0 (4.7–6.6)	5.0 (4.7–6.1)	0.79	6.1 (5.0–6.9)	0.22

*Mann–Whitney *U* test for comparison with reference (TP-only or no lesions group).

Data are shown as medians with interquartile range in parentheses. Analysis is performed on patient subset for clinical image analysis consisting of patients with ≤20 lesions per organ (n = 33). n refers to number of patients in each group. Patients may appear in both FN and FP groups if they have both FP and FN lesions. Accordingly, total number of patients exceeds 33.

curves (*25*). For comparison of performance for the detection of multiple lesions per patients, the areas under the free-response operating characteristic curves, which take into account detection confidence and the location of lesions, could be compared for $PET_{100\%}$ and PET_{AI} using an external standard of truth (*26*). However, we consider $PET_{100\%}$ as the standard of truth to be the most relevant reference in our case, because PET_{AI} is directly derived from the corresponding full-dose images through low-dose simulation and denoising through the AI algorithm. We find the 2-point confidence score (C_1 or C_0) to be representative of the clinical reading situation: the reader either is confident that a lesion is present (C_1) or has some uncertainty and raises a flag (C_0) such that special attention can be drawn to the suspicious area on prior or subsequent scans. Furthermore, we find that the 2-point confidence score sufficiently underscores concerns about using PET_{AI} for lesion detection, because 23 of the 37 FP cases were considered definite lesions (and thus given a C_1 score). Thus, selecting C_1 as the threshold for lesions still provides alarmingly high lesion detection FDR of 23% (23/100) and low lesion detection sensitivity of 71% (77/108).

The Likert scale–defined image quality used in this paper represents the readers' overall subjective assessment of how the images compare to standard ^{64}Cu-DOTATATE PET images seen in the clinical setting. Other definitions of image quality for assessment of AI imaging methods include objective task-based evaluations of the image quality, e.g., lesion detection like in our study, for objective assessment of image quality (*27*). The distinction between the subjective image quality and the objective lesion detection task is important, because the PET_{AI} Likert scale-defined image quality generally were rated as good or excellent; that is, to the reader, the PET_{AI} images overall appear similar to high-quality ^{64}Cu-DOTATATE PET images seen in a clinical setting, whereas the objective lesion detection task demonstrated that the PET_{AI} images were inadequate.

CONCLUSION

We implemented a state-of-the-art PT-WGAN denoising AI algorithm on simulated low-dose ^{64}Cu-DOTATATE PET images from patients with NENs of a suboptimal standard to test whether the image quality and lesion detection rate could be restored. The algorithm improved the image similarity metrics, and the perceived Likert scale–defined image quality of PET_{AI} was equivalent to the full-dose PET images. However, application of the denoising algorithm resulted in FN lesions and FP lesions when compared with full-dose PET in a clinical analysis. The discrepancies highlight the need for relevant clinical validation of AI algorithms.

DISCLOSURE

This project received funding from the European Union's Horizon 2020 research and innovation program under grant agreements 670261 (ERC Advanced Grant) and 668532 (Click-It), the Lundbeck Foundation, the Novo Nordisk Foundation, the Innovation Fund Denmark, the Neuroendocrine Tumor Research Foundation, the Danish Cancer Society, Sygeforsikringen "Danmark," the Arvid Nilsson Foundation, the Neye Foundation, the Research Foundation of Rigshospitalet, PERSIMUNE through the Danish National Research Foundation (grant 126), the Research Council of the Capital Region of Denmark, the Danish Health Authority, the John and Birthe Meyer Foundation, the Danish Agency for Digitization (20196164), and the Research Council for Independent Research. Andreas Kjaer is a Lundbeck Foundation Professor. Ulrich Knigge and Andreas Kjaer are inventors of or hold intellectual property rights on a patent covering ^{64}Cu-DOTATATE. No other potential conflict of interest relevant to this article was reported.

KEY POINTS

QUESTION: Can the image quality and lesion detection rate of low-dose (<50 MBq) ^{64}Cu-DOTATATE PET scans from patients with NENs be restored using state-of-the-art AI for image denoising?

PERTINENT FINDINGS: The denoising AI algorithm performed well on standard image fidelity-based comparison metrics, and the perceived Likert scale–defined image quality was restored. However, clinical assessment showed that more than half of the lesions found on the denoised low-dose ^{64}Cu-DOTATATE PET were FP or FN compared with the full-dose scans.

IMPLICATIONS FOR PATIENT CARE: The study highlights the importance of assessing clinically relevant endpoints when evaluating AI algorithms in nuclear medicine in accordance with the RELAINCE guidelines.

REFERENCES

1. Bozkurt MF, Virgolini I, Balogova S, et al. Guideline for PET/CT imaging of neuroendocrine neoplasms with ^{68}Ga-DOTA-conjugated somatostatin receptor targeting peptides and ^{18}F-DOPA. *Eur J Nucl Med Mol Imaging.* 2017;44:1588–1601.
2. Janson ET, Knigge U, Dam G, et al. Nordic guidelines 2021 for diagnosis and treatment of gastroenteropancreatic neuroendocrine neoplasms. *Acta Oncol.* 2021; 60:931–941.
3. Knigge U, Capdevila J, Bartsch DK, et al. ENETS consensus recommendations for the standards of care in neuroendocrine neoplasms: follow-up and documentation. *Neuroendocrinology.* 2017;105:310–319.
4. Strosberg JR, Halfdanarson TR, Bellizzi AM, et al. The North American Neuroendocrine Tumor Society consensus guidelines for surveillance and medical management of midgut neuroendocrine tumors. *Pancreas.* 2017;46:707–714.
5. International Commission on Radiological Protection. The 2007 Recommendations of the International Commission on Radiological Protection. ICRP publication 103. *Ann ICRP.* 2007;37:1–332.
6. Detectnet label. U.S. Food and Drug Administration website. https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/213227s000lbl.pdf. Updated December 2021. Accessed August 5, 2022.
7. Loft M, Carlsen EA, Johnbeck CB, et al. Activity dose reduction in ^{64}Cu-DOTATATE PET in patients with neuroendocrine neoplasms: impact on image quality and lesion detection ability. *Mol Imaging Biol.* 2022;24:600–611.
8. Wang T, Lei Y, Fu Y, et al. Machine learning in quantitative PET: a review of attenuation correction and low-count image reconstruction methods. *Phys Med.* 2020;76:294–306.
9. 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.
10. Schaefferkoetter J, Yan J, Ortega C, et al. Convolutional neural networks for improving image quality with noisy PET data. *EJNMMI Res.* 2020;10:105.
11. Yu Z, Rahman MA, Jha AK. Investigating the limited performance of a deep-learning-based SPECT denoising approach: an observer-study-based characterization. *Proc SPIE Int Soc Opt Eng.* 2022;12035.
12. 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(suppl 1):575.
13. 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.* 2020;3:e200137.
14. Jha AK, Bradshaw TJ, Buvat I, et al. Nuclear medicine and artificial intelligence: best practices for evaluation (the RELAINCE guidelines). *J Nucl Med.* 2022;63:1288–1299.
15. Pfeifer A, Knigge U, Mortensen J, et al. Clinical PET of neuroendocrine tumors using ^{64}Cu-DOTATATE: first-in-humans study. *J Nucl Med.* 2012;53:1207–1215.
16. Pfeifer A, Knigge U, Binderup T, et al. ^{64}Cu-DOTATATE PET for neuroendocrine tumors: a prospective head-to-head comparison with ^{111}In-DTPA-octreotide in 112 patients. *J Nucl Med.* 2015;56:847–854.
17. Loft M, Carlsen EA, Johnbeck CB, et al. ^{64}Cu-DOTATATE PET in patients with neuroendocrine neoplasms: prospective, head-to-head comparison of imaging at 1 hour and 3 hours after injection. *J Nucl Med.* 2021;62:73–80.

18. Johnbeck CB, Knigge U, Loft A, et al. Head-to-head comparison of ^{64}Cu-DOTATATE and ^{68}Ga-DOTATOC PET/CT: a prospective study of 59 patients with neuroendocrine tumors. *J Nucl Med.* 2017;58:451–457.
19. Gong Y, Shan H, Teng Y, et al. Parameter-transferred Wasserstein generative adversarial network (PT-WGAN) for low-dose PET image denoising. *IEEE Trans Radiat Plasma Med Sci.* 2021;5:213–223.
20. van der Walt S, Schonberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ.* 2014;2:e453.
21. Xing Y, Qiao W, Wang T, et al. Deep learning-assisted PET imaging achieves fast scan/low-dose examination. *EJNMMI Phys.* 2022;9:7.
22. Sanaat A, Shiri I, Arabi H, Mainta I, Nkoulou R, Zaidi H. Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging. *Eur J Nucl Med Mol Imaging.* 2021;48:2405–2415.
23. Zhou B, Tsai YJ, Chen X, Duncan JS, Liu C. MDPET: a unified motion correction and denoising adversarial network for low-dose gated PET. *IEEE Trans Med Imaging.* 2021;40:3154–3164.
24. Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol.* 2006;3:413–422.
25. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* 2010;5:1315–1316.
26. Thompson JD, Manning DJ, Hogg P. The value of observer performance studies in dose optimization: a focus on free-response receiver operating characteristic methods. *J Nucl Med Technol.* 2013;41:57–64.
27. Jha AK, Myers KJ, Obuchowski NA, et al. Objective task-based evaluation of artificial intelligence-based medical imaging methods: framework, strategies, and role of the physician. *PET Clin.* 2021;16:493–511.