European Journal of Nuclear Medicine and Molecular Imaging
https://doi.org/10.1007/s00259-025-07672-7

ORIGINAL ARTICLE

Low-count whole-body PET denoising with deep learning in a multicenter, multi-tracer and externally validated study

Justine Maes[1] · Charles Carron[1] · Simon DeKeyser[2] · Tomas Brants[2] · Vicky De Ridder[2] · Amine Chaouki[3] · Camille Steenhout[4] · Stefaan Vandenberghe[5] · Yves D'Asseler[12] · Laurens Raes[6] · Azzam Abdalla Ibrahim[7] · Gerard Moulin-Romsee[8] · Isaac Kargar Samani[9] · Ludovic D'hulst[10] · Sezgin Ustmert[11] · Maarten Larmuseau[2]

Received: 29 July 2025 / Accepted: 5 November 2025

Abstract

Background Positron Emission Tomography (PET) is a powerful diagnostic tool, but its availability, high cost and radiation burden limit its accessibility. Deep learning–based denoising offers a potential solution by enabling low-count PET scans, reducing tracer dose or scan time without compromising diagnostic utility. However, clinical validation of such approaches across different scanner technologies and radiotracers remains limited.

Methods We conducted a multicenter, blinded evaluation of NUCLARITY, a deep learning–based denoising software, using PET data from three European hospitals. Data included 65 scans acquired with [1^8F]FDG, [1^8F]PSMA, [^{68} Ga]PSMA, and [^{68} Ga]DOTATATE on GE and Siemens systems not seen during model training. Low-count scans (50% simulated) were denoised and compared to full-count clinical scans. Image quality was assessed using RMSE, PSNR, and SSIM. Six nuclear physicians evaluated diagnostic image quality (DIQ), diagnostic confidence (DC), and lesion detection across six anatomical regions. Lesion quantification was compared using SUVmean, SUVmax, and MTV.

Results Low-count enhanced (LCE) scans showed improved quantitative image quality metrics compared to unenhanced low-count scans (higher PSNR/SSIM, lower RMSE). Across 243 lesions, SUVmean and SUVmax showed high concordance between standard-count (SC) and LCE scans (CCC = 1.00 and 0.99, respectively). Diagnostic image quality and confidence were slightly lower on LCE versus SC scans, but only one reader indicated a clear preference for SC. Sensitivity and specificity for lesion detection in LCE scans were 99% and 99%, respectively, with interscan agreement exceeding inter-reader variability.

Conclusions This is the first blinded, multicenter reader study evaluating a PET denoising algorithm in a European clinical setting across multiple tracers, incorporating unseen scanner technologies. The denoising algorithm demonstrated robust generalizability and preserved diagnostic accuracy on 50% count data. These findings support the clinical adoption of deep learning-based PET denoising to reduce dose or scan time for four commonly used tracers.

Keywords PET denoising · Deep learning · Diagnostic accuracy

Background

Over the past decades Positron Emission Tomography (PET) has become indispensable in the field of oncology due to its ability to provide detailed metabolic and functional information. Compared to other imaging technologies such as MRI and CT, which visualize anatomical changes, PET scans allow visualization of complementary functional activity, enabling earlier detection of disease [1]. Because of this desirable property, the demand for PET scans has systematically increased in most Western countries over the past years. However, the high diagnostic power of PET scans comes at a price, as the need for a radiotracer makes PET an expensive modality that involves more complex logistics

Justine Maes, Charles Carron and Simon DeKeyser contributed equally to this work.

Extended author information available on the last page of the article

Published online: 20 November 2025

and carries additional radiation risks for both patients and healthcare providers [2–4]. Especially in younger patients who require periodic surveillance scans, the cumulative radiation dose may be a concern, as the exposure results in a considerable higher life-time attributable risk of cancer [5, 6]. In addition, compared to other imaging modalities relatively long acquisition times are needed to obtain sufficient image quality. Moreover, as new and more expensive radiotracers find their way to clinical practice, the costs associated with PET imaging are a pressing concern [7–9].

The advent of advanced deep learning technologies has enabled post-hoc denoising of PET scans, offering the potential to reduce scan duration or lower the administered radiotracer dose. In recent years, substantial research has been dedicated to denoising low-count PET acquisitions using a variety of deep learning methodologies [9–11]. Most approaches rely on simulated low-count scans obtained by downsampling list-mode data, which are paired with their corresponding full-count clinical scans [12, 13]. Supervised learning models are then trained to reconstruct high-quality images from these low-count input images. Early models predominantly utilized convolutional neural networks (CNNs) [14, 15], whereas more recent architectures have incorporated attention mechanisms originally developed for large language models, such as Vision Transformers (ViTs) and Swin Transformers [16–18]. Each architecture presents specific advantages and limitations, as demonstrated in the benchmarking study by Wang et al., where they compared five architectures on low-count PET-MRI scans [19]. More recently, diffusion probabilistic models have demonstrated compelling performance, generating enhanced images that, in some cases, surpass the subjective quality of standard clinical scans [20, 21]. The clinical utility of these AI-based denoising models critically depends on their ability to generalize to unseen scans acquired on different scanner technologies, acquired using different scanning protocols, and possibly with alternative radiotracers. As pointed out by Liu et al., varying noise levels in input images represent a critical challenge that must be addressed to enable the successful clinical adoption of in silico PET denoising [22].

To measure and assess the ability of a deep learning model to generalize to unseen scans of varying noise levels, criteria are needed to gauge the quality of the denoised scan. As pointed out by Rogasch et al., image quality criteria can be subdivided into subjective assessments, such as those provided by certified readers, and objective quantitative measures, which, for example, evaluate SUV quantification in enhanced scans or in phantom datasets where a theoretical ground truth is available [23, 24]. Other metrics that are commonly used in literature are the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) [15, 22]. Ultimately, the primary objective of any diagnostic scan is to enable accurate identification of potential abnormal or malignant lesions by the interpreting physician [25, 26]. Consequently, any enhanced PET scan should not give rise to new, false positive lesions, while minimizing the number of missed or false negative lesions.

In this study, we present a comprehensive clinical validation of NUCLARITY, a PET image denoising software, within a European clinical setting. Validation was performed using data from three external nuclear medicine centers employing both contemporary and legacy scanner technologies distinct from those used during model training [27]. The study design follows a blinded evaluation protocol similar to that of Chaudhari et al. [28], and extends it by incorporating four radiotracers to evaluate the model's generalizability across different tracers. The analysis includes both quantitative and qualitative assessments of image quality, as well as lesion detectability, to rigorously evaluate NUCLARITY's denoising performance on previously unseen acquisition protocols, scanner types, and radiotracers.

Methods

Data collection

The denoising model was trained on 64 × 64x64 volumes sampled from over thousand scans acquired on a Siemens Biograph Vision 600 with a point spread function (PSF) + time of flight (TOF) 3i5s reconstruction method and a GE discovery 710 with a Q.Clear FX reconstruction method. Three tracers were included during model training, [^{18}F]FDG, [^{68} Ga]PSMA-11, and [^{68} Ga]DOTATATE.

For validation within this study, list-mode PET data were collected from three European hospitals, AZ Groeninge, VUB Brussels and Ghent University Hospital. This acquisition mode captures a time-stamped list of all detected radioactive events during the scan. While standard clinical reconstructions utilize 100% of the recorded events, simulating a 50% reduction, either in scan time or radiotracer dose, can be achieved by reconstructing images using only 50% of the list-mode data. Across the three institutions, matched clinical (100%) and simulated low-count (50%) scans were obtained for a total of 65 subjects. The study was conducted using fully anonymized data in compliance with the General Data Protection Regulation (GDPR), and the study protocols were approved by the respective institutional review boards. Eligible patients were adults undergoing routine whole-body PET/CT scans with [18F]FDG, [18F]PSMA, [68 Ga]PSMA, or [68 Ga]DOTATATE. Scans were randomly selected without consideration of the clinical indication, which was not disclosed for the purposes of this study. Three different scanners were included in the study,

Table 1 Demographics of subjects included in this study and the corresponding PET scanner specifications

Institution #	A	B	C
Number of patients	35	20	10
Radiotracer(s)	[^{18}F]FDG (20) [^{68} Ga]PSMA-11 (10) [^{68} Ga]DOTATATE (5)	[^{18}F]FDG (20)	[^{18}F]PSMA-1007 (10)
Age (years)	*	56±12 Range: 36—83	73±8 Range: 61—82
Sex (M/F)	[^{18}F]FDG: 10/10 [^{68} Ga]PSMA: 10/0 [^{68} Ga]DOTATATE: 2/3	8/12	10/0
Patient Weight (kg)	76.8±19.6	83.4±19.8	77.8±9.8
Dose (MBq)	[^{18}F]FDG: 155.3 [^{68} Ga]PSMA: 171.9 [^{68} Ga]DOTATATE: 152.1	284±61.9	161.8±27.8
Scanner model	GE Omni Legend	Siemens Biograph 128 mCT	GE Discovery MI
Axial length (cm)	32	21.6	15
Standard scan time (min)	[^{18}F]FDG: 8 [^{68} Ga]PSMA: 10 [^{68} Ga]DOTATATE: 10	[^{18}F]FDG: 16	[^{18}F]PSMA: 24
Reconstruction technique**	HPDL-Q.Clear HD	PSF+TOF 2i21s	Q.Clear FX

* Patient's age was unknown

** *HDPL* high performance deep learning, *PSF* point spread function, *TOF* time-of-flight

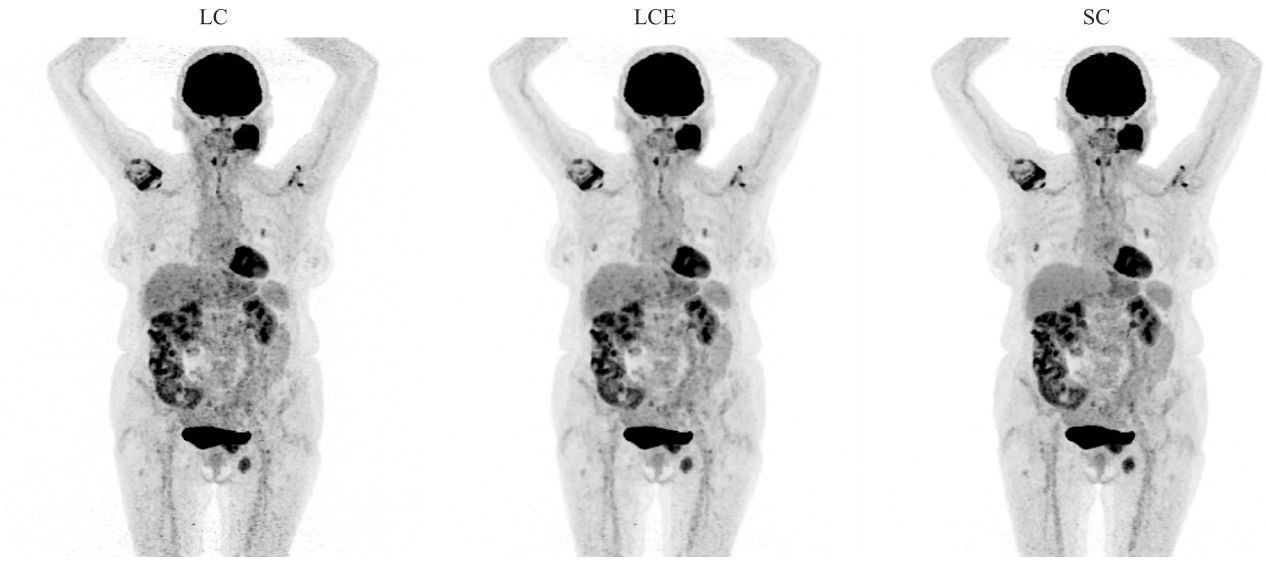

Fig. 1 A comparison of the low-count (LC), low-count enhanced (LCE) and standard count (SC) scans. The low-count scan was obtained using 50% of the measured counts of the standard-count scan, which is the scan used in routine clinical practice

an Omni and a Discovery MI scanner from GE Healthcare and the Biograph 128 mCT from Siemens Healthineers, which represent scanner technologies that were not part of the training data for the denoising model. Detailed demographic and technical information is summarized in Table 1, for each of the participating centers.

Low-count PET enhancement

The scans with 50% reduced counts were enhanced using NUCLARITY® v1.0.0, a deep learning algorithm developed by NUCLIVISION. An example of such an enhancement can be seen in Fig. 1. The NUCLARITY image processing software employs a CNN-based algorithm to enhance image quality by reducing noise. The CNN in NUCLARITY assesses and processes each voxel in relation to its neighbouring pixels, effectively learning optimal filters that capture the most relevant features with minimal pre-processing [14, 15, 22]. Additionally, the software utilizes a residual learning framework to facilitate the training of deeper networks. In residual learning, instead of attempting to learn an entire new output, the network learns

the difference (residual) between the input and a predicted output. This approach helps in distinguishing between noise components and structural components in the image [29]. By emphasizing these residuals, NUCLARITY effectively enhances structural details while concurrently diminishing noise, leading to clearer, more precise imaging outcomes [15, 29, 30].

The image enhancement is achieved by NUCLARITY through a dual UNet architecture, wherein each UNet is specialized to handle different noise levels in the imaging data. This methodology is similar to the approach proposed by Liu et al., which demonstrated superior image quality enhancement performance through the application of an ensemble of 3D UNet models, each trained at varying noise levels [22].

Reader study

Six independent nuclear medicine physicians with 1, 3, 3, 8, 9, and 16 years of experience, blindly assessed both the standard count (SC, i.e. the clinical scan at 100% of the counts) and the low-count enhanced (LCE) scans of the subjects. The readers did not have access to the clinical indications for the scan, nor to any other prior information on the patient. Three of the six readers reviewed the 100 ^{18}F-labeled scans, while the other three reviewed the 30 ^{68} Ga-labeled scans.

The readers were asked to indicate the number of abnormalities they could identify, for six different anatomical regions. The regions considered were lung, liver, lymph nodes, bone, spleen and muscle, in accordance with Chaudhari et al. [28]. Any number of lesions above 5 was set to 6. In addition, all readers were asked to give a diagnostic confidence (DC) score on a 3-point Likert scale (1 = not sure, 2 = confident, 3 = very confident), and a diagnostic image quality (DIQ) score on a 5-point Likert scale (1 = very poor, 2 = poor, 3 = acceptable, 4 = good, 5 = excellent).

Approximately 1 year after the first reading session, the two readers that obtained the highest interscan concordance, reader 2 and reader 4, performed the assessment again on the low-count scans. Enough time was left in between, to mitigate any recall bias. Unbeknownst to the readers, one third of the low-count scans was replaced with the original clinical images to assess intra-reader variability. In total, during the second assessment phase, reader 2 assessed 17 SC and 33 low-count (LC) scans, while reader 4 assessed 6 SC and 9 LC scans. and In addition, all false positive and false negative body regions were analysed retrospectively by C.C. and quantitative analysis was performed of the uptake values of lesions that were only seen in either the LCE or the SC scans. During this final review, C.C. could consult the all information and comments provided by all six readers.

Statistical analysis

The similarity between the low-count enhanced and clinical scans was assessed using three metrics, the Root Mean Squared Error (RMSE), the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) [15, 22]. For each of these metrics, the clinical image is taken as the reference image, and the similarity to this reference is compared for the low-count and the low-count enhanced scan. A better similarity results in a higher PSNR and SSIM, and a lower RMSE.

For the Diagnostic Confidence and Diagnostic Image Quality, a paired non-parametric Wilcoxon test for non-inferiority was performed with a 0.5 inferiority margin, similarly to Chaudhari et. al. [28].

To enable SUV quantification in the tumor lesions, semi-automated tumor annotations were performed by C. C. and J. M. using the MIM Encore®viewer software and the accompanying PET Edge® software for tumor contouring. Metabolic tumor volume (MTV) was calculated by considering all voxels in an annotated tumor and selecting only those that had an SUV within 40% of the SUVmax within that tumor. The volume was then calculated by multiplying the withheld voxels with the voxel spacing.

Results

Image quality

First, the image quality was assessed using a combination of technical metrics and subjective reader scores. Figure 2 shows the improvement in image quality using three quantitative metrics SSIM, RMSE and PSNR for the different radiotracers. The figure shows that the PSNR and the SSIM are higher for all the low-count enhanced scans compared to the original low-count scans, reflecting a higher similarity with the clinical scan. Similarly, the RMSE is systematically lower for the enhanced images, showing again a better correspondence to the ground truth images.

To further explore the differences in image quality, subjective readers scores on a 1–5 Likert scale were compared between the different readers. Two readers, readers 2 and 4, also assessed a subset of the LC scans. The upper row in Fig. 3 shows the results obtained for the different readers in terms of DIQ and DC ratings. Pooled, the DIQ scores for the standard-count (SC), low-count (LC) and low-count enhanced (LCE) scans were 4.1 ± 0.7, 3.2 ± 1.4, and 3.6 ± 0.7, respectively. The DC scores were 2.8 ± 0.3, 2.4 ± 0.6, and

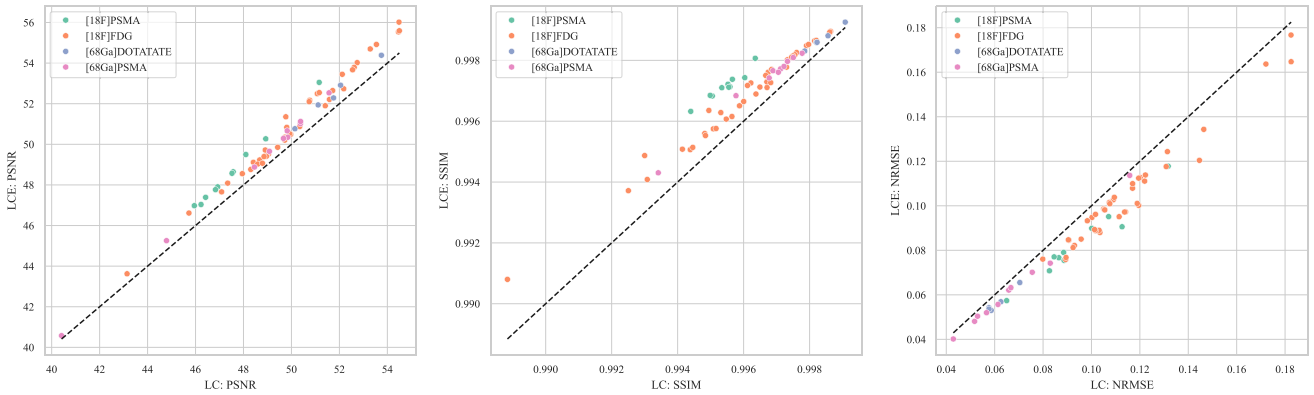

Fig. 2 Scatter plots showing the image quality improvement of the low-count enhanced (LCE) scans versus the low-count (LC) scans, assessed by PSNR, SSIM and NRMSE with respect to the standard-count (SC) scans for the different radiotracer types

2.5 ± 0.3 for the SC, LC, and LCE scans, respectively. For one reader, reader 3, the subjective image quality of the LCE scans seemed clearly inferior to the SC scans, when using a 0.5 non-inferiority margin as in Chaudhari et al. [28]. Notably, reader 3 is the most experienced reader, a similar finding to what was reported before by Schaefferkoetter et al., where the most experienced reader gave the lowest scores to AI-enhanced scans [14]. For reader 2 and 4 an improvement of the LCE to the LC scans could be observed, both for DIQ and DC, which can be seen in the upper row of Fig. 3.

Interestingly, when looking across the different tracers, no obvious non-inferiorities could be observed and subjective image quality scores were relatively best for [^{18}F]PSMA when compared against the standard count image, a tracer that was not seen during training. The same applies for scanner technologies, where the biggest discrepancy in subjective image quality was seen for the Omni scanner. In general, most readers felt that the scans from the Omni exhibited excellent image quality, as reflected in high subjective reader score for this particular scanner.

Quantitative assessment

In total, 243 hypermetabolic lesions were segmented on the SC scans by two independent physicians in residence (J.M. and C.C.), with lesion sizes varying from 21 to 119669 voxels, with a median of 330 voxels. The upper row of Fig. 4 shows a scatter plot comparing the SUV_{mean} and the SUV_{max} as measured in the segmented lesions of the SC and LCE scans. For both SUV_{mean} and SUV_{max} a strong correlation can be observed between the quantification in the SC and the LCE scan, resulting in a CCC of 1.00 and 0.99 for SUV_{mean} and SUV_{max}, respectively. Notice that for the SUV_{max} some outliers can be observed between the SC and LCE scans, in highly metabolically active regions where the $SUV_{max} > 20$ in the SC scan. These outliers originate from a pelvic and an iliac lymph node in two different patients. For these lesions, the relatively large deviation in SUV quantification is likely attributable to the combined effects of small lesion size, high SUV values, and the reported suboptimal image quality of these [18F]PSMA scans. Together, these factors may cause the model to be uncertain whether the observed signal reflects true physiological uptake or stochastic noise arising from the limited number of detected counts, resulting in considerable smoothing of the lesion.

These outliers stem from a pelvic and iliac lymph node from two distinct patients. For these lesions, the relatively large deviation in SUV quantification is presumably caused by a combination of small lesion size with relatively low image quality. a combination that for SUV values above 20 results in considerable smoothing. Next, we also wanted to know whether the enhanced scans alter the metabolic tumor volume (MTV) in comparison to the original standard-count scans. The bottom row of Fig. 4 shows the Bland–Altman plots for the low-count scans, when compared against the standard-count scans, and for the enhanced scans. Increasing the count level by 50% generally results in a lower MTV, which is most strikingly exemplified by three outliers in the MTV. After the enhancement the deviations observed in these outliers are strongly reduced, resulting in a lower mean difference.

Lesion detection

For the lesion detectability, we focused in first instance on whether a body region (i.e. Lung, Liver, Spleen, Bone, Muscle and Lymph node) was abnormal, i.e. contained at least one clinically significant lesion, or normal, meaning that no lesions could be identified. Figure 5 shows the number of readers that denoted a specific body region as abnormal (#lesions > 0), comparing the assessment on SC and LCE scans in a confusion matrix. Notice that on the SC scans only 313 body regions out of a total of 390 body regions receive consensus by all three readers, as can be seen by

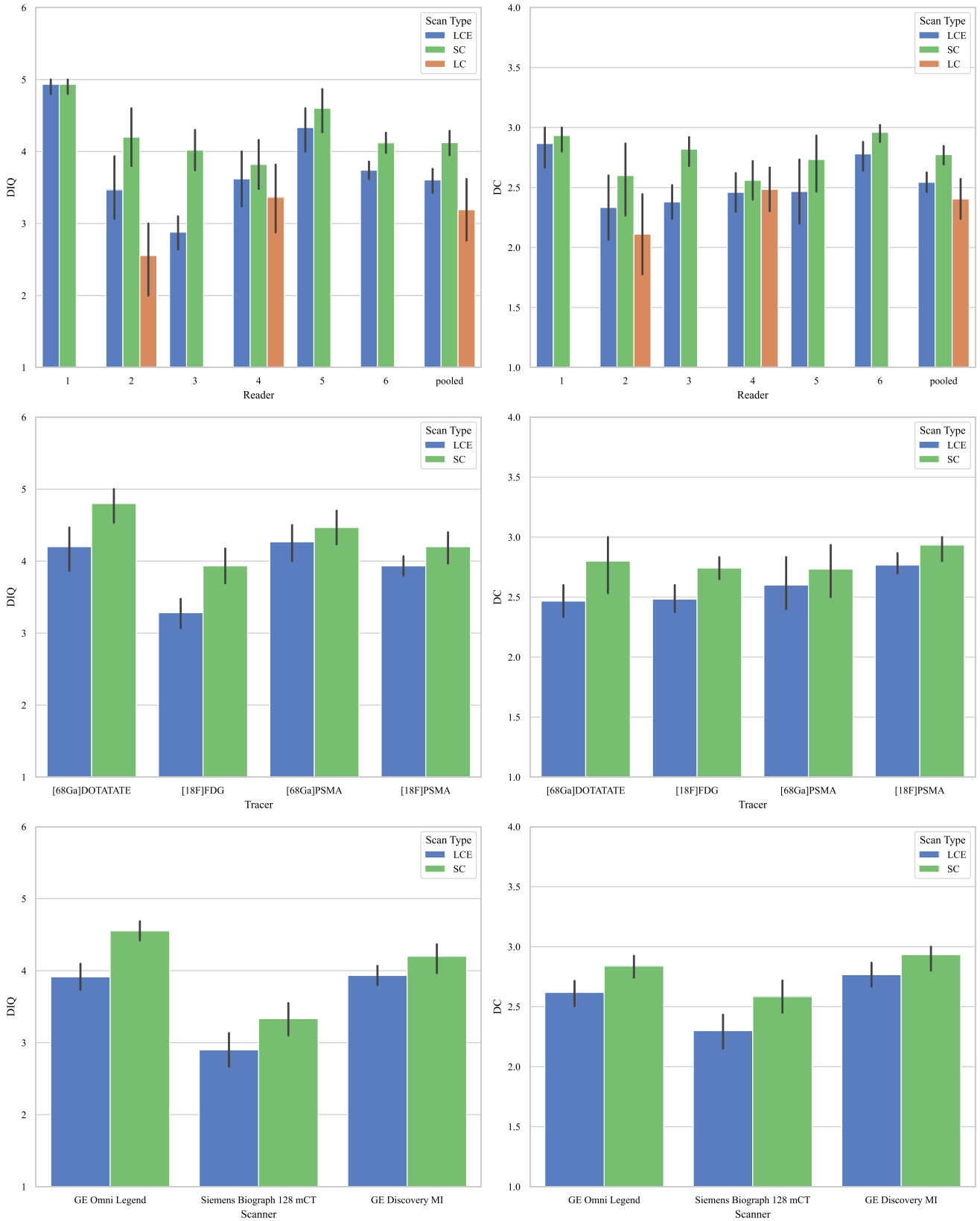

Fig. 3 Mean and standard deviation of the diagnostic image quality (DIQ) and diagnostic confidence (DC) for the standard-count (SC), low-count (LC) and low-count enhanced (LCE) scans, comparing between different readers (upper row), different tracer (middle row) and different scanner technologies (lower row). The scores for the LC scans were only obtained for a subset of the readers and could not be aggregated

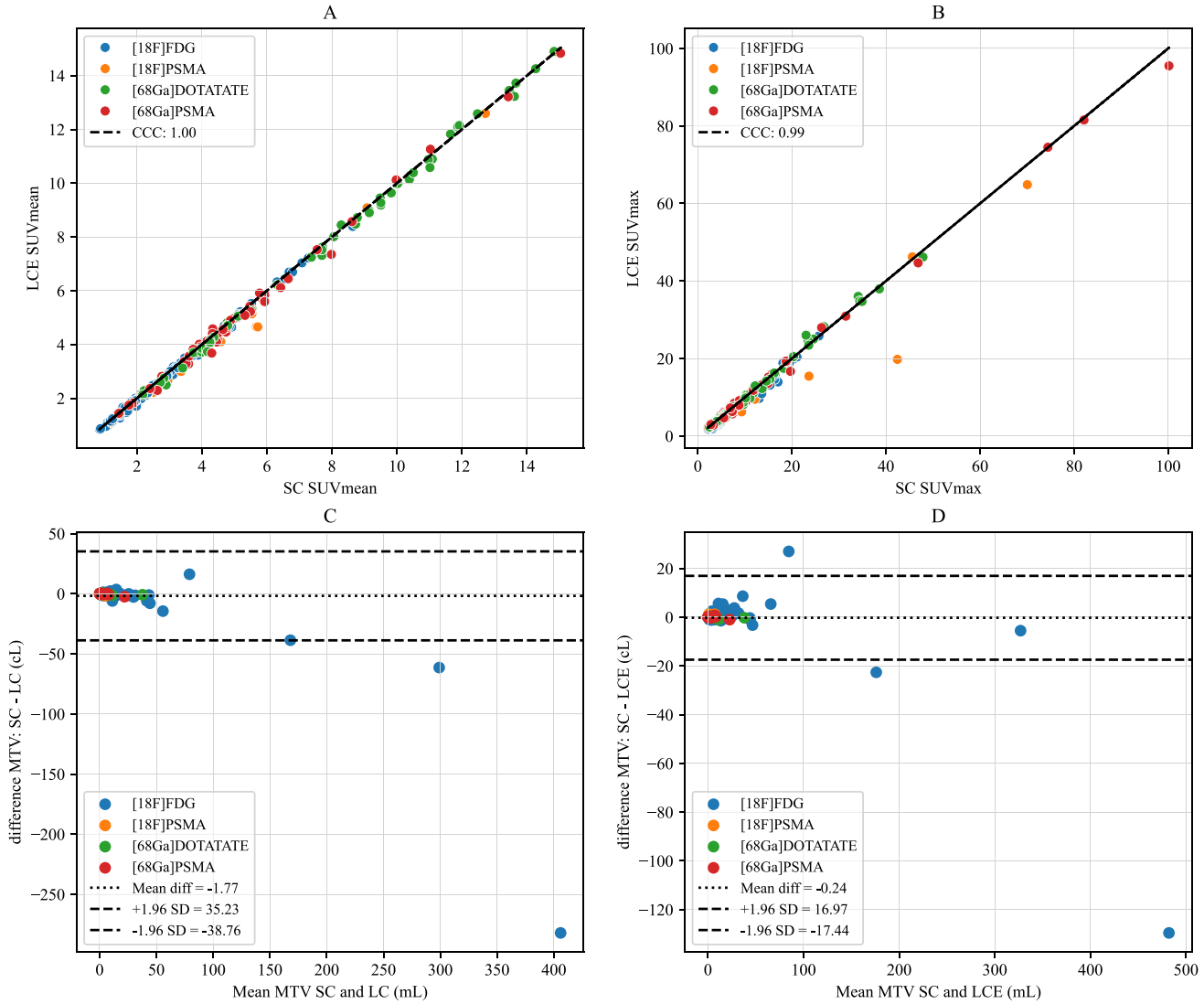

Fig. 4 Upper row: Scatter plots comparing the SUVmean (**A**) and SUVmax (**B**) between the standard count (SC) and the low-count enhanced (LCE) scans. Lower row: Bland–Altman plots comparing the MTV between the standard count and low-count scans (**C**) and between the standard count and enhanced scans (**D**)

summing the upper and lower row of the matrix, i.e. where the number of readers that find an abnormality is either 0 or 3. In total, 46 body regions were regarded as abnormal by all three readers on the SC scans, while on the LCE scans 44 regions could be identified by all three readers, as shown in the lowest row and last column, respectively. To calculate the sensitivity and specificity, SC and LCE scans were defined to be discordant if at least two readers did not agree on the label of a body region, between SC and LCE scans. From Fig. 5 it can be seen that 2 body regions were indicated as abnormal by two readers on the LCE only, whereas one body region was identified by 2 readers only on the SC scan. The corresponding sensitivity and specificity for each body region are listed in Table 2. Here, a region is denoted as abnormal when two or more readers find an abnormality on the SC scans, and false positives or false negatives are when two or more readers on LCE scan assign a different label than on the SC scan. Table 2 shows that the specificity and sensitivity are consistent among the different body regions, although the lower number of lesions for spleen, muscle, liver and lung results in larger variations of the sensitivity.

To gain a better understanding of the false positive and false negatives findings, we performed a more in-depth analysis of these scans, which are shown in Table 3. The mean intensity projections (MIPs) of these samples are shown in Supplementary Figs. 1–3. For Subject 66, a bone lesion close to the sigmoid colon was seen on the SC scan that could not be found on the LCE scan. However, during tumor annotations by independent reviewers, these bone lesions were also missed on the SC scan, such that the bone lesions

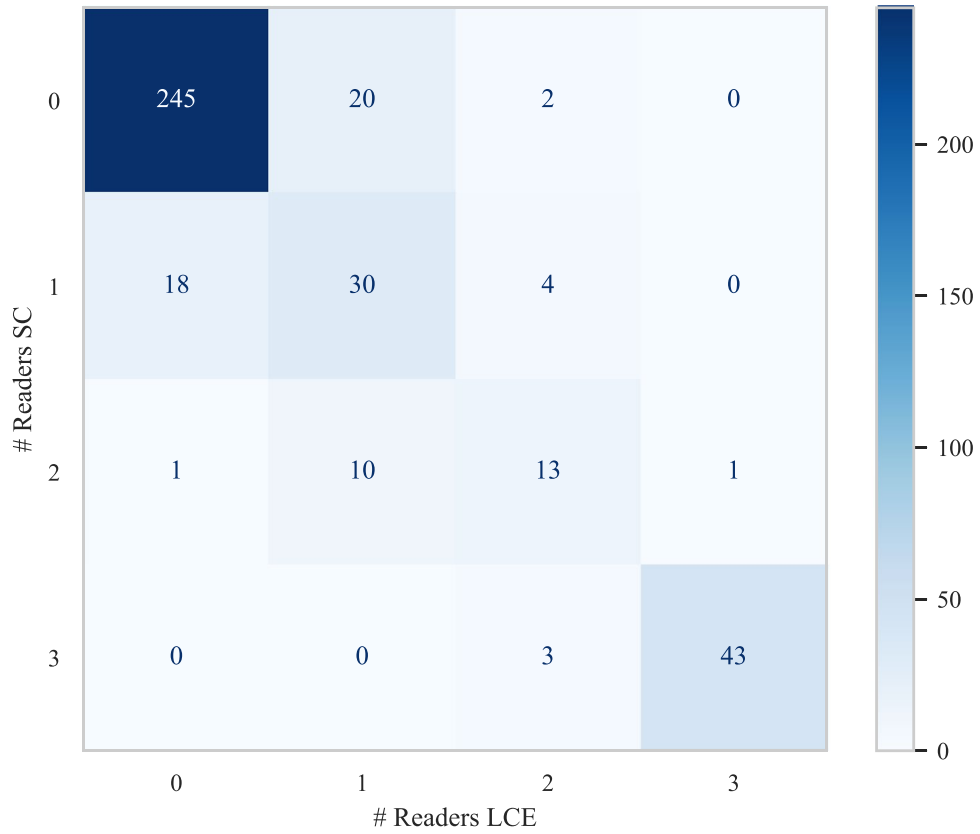

Fig. 5 Confusion matrix showing the number of readers who classified a body region as abnormal (#lesions > 0) in standard-count (SC) and low-count enhanced (LCE) scans

Table 2 Patient-level sensitivity, specificity and the corresponding F1-score together with the concordance correlation coefficient (CCC) of the hypermetabolic lesions detected by the readers (pooled) comparing low-count enhanced scans to the standard-count scans

Organ	Sensitivity	Specificity	F1	CCC (All lesions)
Bone	0.95 19/20	1.00 45/45	0.97	0.90
Lymph nodes	1.00 28/28	0.95 35/37	0.97	0.95
Liver	1.00 5/5	1.00 60/60	1.00	0.98
Lung	1.00 10/10	1.00 55/55	1.00	0.93
Muscle	1.00 6/6	1.00 59/59	1.00	0.90
Spleen	1.00 2/2	1.00 63/63	1.00	0.92
Overall	**0.99** **70/71**	**0.99** **317/319**	**0.98**	**0.95**

Table 3 Summary of body regions where reader classifications (normal/abnormal) in low-count enhanced (LCE) scans differed from those in standard-count (SC) scans. Diagnostic image quality (DIQ) and diagnostic confidence (DC) scores are given for the SC scan (pooled)

Subject ID	Tracer	DIQ (SC)	DC (SC)	Body region	Classification
66	[^{18}F]FDG	4.67	3.00	Bone	False negative
113	[^{68} Ga]PSMA	5.00	3.00	Lymph nodes	False positive
97	[^{18}F]PSMA	3.33	2.33	Lymph nodes	False positive

are probably non-pathological and therefore not indicated by the third reader. A line profile of this lesion, showed that the SUVmax of the lesion did decrease with approximately 25% on the enhanced scan, although the lesion itself remained visible (see Figure S4). For Subject 97, one reader commented that there are inguinal but non-pathological lymph nodes on the LCE, and marked this as an abnormality. The other reader that had a mismatch between SC and LCE commented that the image quality of the SC scan was too low to properly assess the scan. A final review by C.C. revealed no increased uptake in SUV on the LCE scans, but rather a decrease in SUVmax by 26%, such that the detection on LCE should probably be attributed to misclassification of benign uptake as pathological. For patient 113, two readers commented that they saw a 10R lymph node on the LCE scan, that was not seen on the SC by any of the readers. During the final review by C.C., no clear pathological uptake was observed in the lymph nodes on both SC and LCE and this was considered a misclassification.

The confusion matrix in Fig. 5 shows noticeable discrepancies between the different readers, both on the enhanced and standard-count scans. Therefore, we wanted to compare the inter-reader agreement, of different readers on the same

Table 4 Interreader agreement between pairs of readers, measured using Cohen's kappa, was assessed for lesion detection across six classes (1 lesion, 2 lesions, …, 5+lesions) per body region within the F-labeled and Ga-labeled radiotracer groups. Results are stratified by scan type: low-count enhanced (LCE), standard-count (SC), and overall

	Reader pairs (F-labeled, $N=600$)			Reader pairs (Ga-labeled, $N=180$)		
	1 & 2	1 & 3	2 & 3	4 & 5	4 & 6	5 & 6
LCE	0.53	0.40	0.34	0.49	0.35	0.50
SC	0.58	0.43	0.46	0.50	0.36	0.50
Overall	0.55	0.41	0.40	0.49	0.36	0.50

Table 5 Interscan agreement between the standard-count (SC) and low-count enhanced (LCE) scans, measured using Cohen's kappa, was assessed for lesion detection across six classes (1 lesion, 2 lesions, …, 6+lesions) per body region within the F-labeled and Ga-labeled radiotracer groups. Results are stratified by reader

	Scan pairs (F-labeled, $N=300$)	Scan pairs (Ga-labeled, $N=90$)
	LCE & SC	LCE & SC
Reader 1	0.79	0.61
Reader 2	0.77	0.84
Reader 3	0.55	0.54

scan, against the inter-scan agreement, of the same reader on the LCE versus SC scan. Tables 4 and 5 respectively show the inter-reader and inter-scan agreement as quantified by the Cohen's Kappa on the number of lesions identified by each reader in each body region. From the tables it can be seen that the number of lesions detected has a substantially stronger correlation between the LCE and SC scans than between different readers.

The same analysis was also performed at the body region level, considering the percent agreement on whether a region is normal or abnormal. Between each pair of readers (inter-reader) and between LCE and SC scans for a single reader (inter-scan), the agreement was calculated as the percentage of body regions that received the same label, divided by the total number of body regions. Figure 6 shows that the inter-scan agreement is higher than the inter-reader agreement and on par with the intra-reader agreement for reader 2 and 4. For these readers, the agreement between SC and LC is lower, although still above 90%. Interestingly, the reader that gave the lowest subjective image quality score to the LCE scans (reader 3) also displayed the lowest inter-scan agreement, but also showed reduced agreement with the other readers on the SC scans. Figure 6 also shows that, aside from the lower agreements for the one reader, findings are similar between the 18F- and 68 Ga labeled radiotracers.

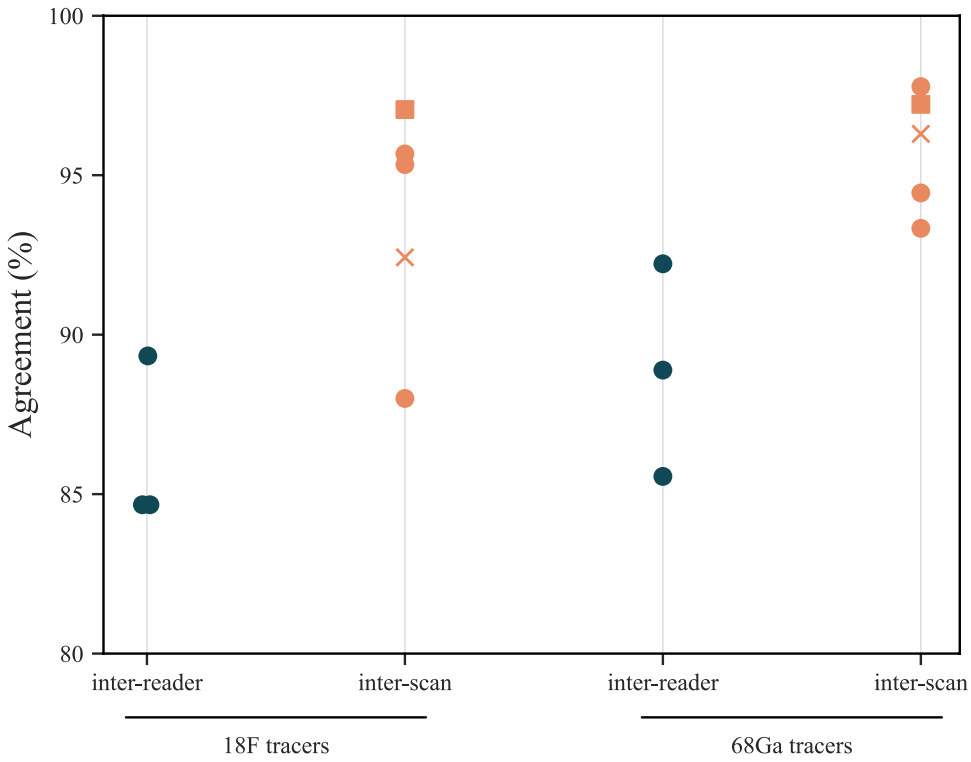

Fig. 6 Inter-reader and inter-scan agreement on whether a body region contains or does not contain a clinically relevant lesion. Orange dots denote the agreement for a single reader, when comparing LCE and SC scans, while blue dots compare the agreement between the different pairs of readers on the SC scans. An "x" denotes the agreement between the LC and SC scans, for the readers that obtained the highest inter-scan agreement (reader 2 and reader 4). The square denotes the intra-reader agreement on a subset of the SC scans, also for reader 2 and 4

Discussion

Although there has been significant scientific interest in using deep learning-based denoising to reduce tracer usage and scan time for PET scans, the number of real-world clinical validations remain limited. Here, a real-world validation of a denoising software was presented that focused on model generalizability by including three scanner technologies and one radiotracer, [18F]PSMA, that were not seen during model training. Given the recent rise in AI-powered models across many different scientific fields, the generalizability of these models to unseen settings remains a main concern and challenge. In this study we have replicated the blinded setting that was proposed in Chaudhari et al., to minimize potential bias across readers toward deep learning-based denoised images [28]. Overall, the use of enhanced scans did not lead to pronounced differences in patient treatment, with a comparable lesion detectability across the different readers. Most notably, it was shown that the variance in lesion detectability between different readers was larger compared to the variance between clinical scans and scans acquired using only 50% of the original clinical counts, enhanced with NUCLARITY. In their study, Chaudhari et al., claimed a 75% reduction in measured counts, but this was in American centers using clinical protocols that result in a significantly higher number of counts for the clinical scans [28]. In a European setting, it has been noted that such a strong reduction is not feasible [24, 26, 31], but none of these studies used a design with independent readers in a blinded setting and focused on a single tracer. Aside from the fact that the scanner technologies in this study were not included in the training data, the in-silico denoising was also performed on a wide range of scanner technologies, ranging from the PMT-based Biograph mCT from Siemens Healthineers to the SiPMT-based Omni from GE Healthcare. For the Omni, the HPDL reconstruction algorithm was used, which also relies on AI to improve image quality [27]. The results from this study seem to suggest that the denoising can be applied to scans that were already processed by another AI-powered algorithm, although more research is needed to assess the impact on SUV quantification in comparison to more standard reconstruction techniques, similarly to what was done in Dadgar et al. [32].

Although the study has tried to mimic the real-world clinical setting as much as possible, there are several limitations. First of all, readers were not aware of the clinical indications for the scan. While this allows a fully unbiased review of a PET-CT, it is possible that the absence of this knowledge results in a larger inter-reader and inter-scan variability compared to a real-world setting. Moreover, as readers came from independent centers, the majority of the readers were unfamiliar with the scanner technology and acquisition protocol used to acquire and reconstruct the scans. Additionally, to ensure consistent viewing between the different readers, readers could not view the images on their viewer of choice, which could also alter their assessment compared to a real-world clinical setting.

Conclusions

In this work we have presented the first blinded reader study of a PET denoising algorithm in a European setting. Results show that a deep learning based denoising algorithm generalizes well to unseen scanner technologies setting, across a range of older to more modern scanner technologies. Generalizability across tracers was also investigated and the denoising model seemed to perform well on the four different tracers considered, [^{18}F]FDG, [^{18}F]PSMA, [^{68} Ga]PSMA, and [^{68} Ga]DOTATATE. While clinical assessments were largely unaffected on enhanced scans acquired using 50% of the counts, one of the six readers had a pronounced preference for the original clinical scans over the enhanced scans in terms of subjective image quality. Importantly, we could show that the inter-reader agreement was well below the agreement between clinical and enhanced scans.

Abbreviations

CNN	Convolutional neural network
PET	Positron emission tomography
CT	Computed tomography
FDG	Fluorodeoxyglucose
PSMA	Prostate-specific membrane antigen
SC	Standard-count
LC	Low-count
LCE	Low-count enhanced
MTV	Metabolic tumor volume
SUV	Standard uptake value
DIQ	Diagnostic image quality
DC	Diagnostic confidence
SSIM	Structural similarity index
NRMSE	Normalized root mean squared error
PSNR	Peak signal-to-noise ratio
GDPR	General Data Protection Regulation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-025-07672-7.

Author's contributions M.L., T.B. and S.V. designed the study, M.L. and T.B. oversaw the study. S.DK. did oversaw data collection and statistical analysis. V.D.R. performed the MTV analysis. A.C., A.I., C.S., G.MR. I.K.S, L.D.H. and S.U. acted as readers in the study and provided feedback on the manuscript. Y.D., S.U. and L.R. collected the data and provided feedback on the manuscript and helped with the analysis. C.C. and J.M. performed the quantitative analysis of the lesions, and did a final review of the false positive and negative lesions.

C.C., J.M., M.L. and S.DK. wrote the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. A.C., A.I., C.S., G.MR. I.K.S, L.D.H. and S.U. received a small fee by Nuclivision for their participation as readers.

Data availability The data from this study are not publicly available in accordance to institutional requirements governing human subject privacy considerations. The data may be made available from the authors upon reasonable request subject to permission and approval from the participating centers.

Declarations

Ethical approval and consent to participate The study was conducted in accordance with the Declaration of Helsinki and used fully anonymized, retrospective data in compliance with the GDPR. Based on the nature of the data and existing data sharing agreements, the participating centers did not require formal ethics approval or informed consent under their institutional policies.

Consent for publication All authors have read and approved the manuscript.

Competing interests S.DK., T.B. and M.L. are currently employed at nuclivision and hold shares in the company. S.U. and S.V. act as advisors for nuclivision and hold shares in the company.

References

1. Geraldes CFGC. Introduction to infrared and raman-based biomedical molecular imaging and comparison with other modalities. Molecules. 2020. https://doi.org/10.3390/molecules25235547.
2. Li W, Fang L, Li J. Exposure doses to technologists working in 7 PET/CT departments'. Dose Response. 2020;18(3):1559325820938288. https://doi.org/10.1177/1559325820938288.
3. McCann A, Vintró LL, Cournane S, Lucey J. Assessment of occupational exposure from shielded and unshielded syringes for clinically relevant positron emission tomography (PET) isotopes-a Monte Carlo approach using EGSnrc. J Radiol Prot. 2021. https://doi.org/10.1088/1361-6498/ac0df5.
4. Farkas J, Martin M, Nielsen C, Jennings SG. The effects on technologist occupational exposure in PET/CT departments when working with students at various levels of supervision. J Nucl Med Technol. 2020;48(3):214–7. https://doi.org/10.2967/jnmt.119.241398.
5. Wen JC, Sai V, Straatsma BR, McCannel TA. Radiation-related cancer risk associated with surveillance imaging for metastasis from choroidal melanoma. JAMA Ophthalmol. 2013;131(1):56–61. https://doi.org/10.1001/JAMAOPHTHALMOL.2013.564.
6. Kessara A, Buyukcizmeci N, Gedik GK. Cancer risk estimation for patients undergoing whole-body PET/CT scans. Radiat Prot Dosimetry. 2023;199(6):509–18. https://doi.org/10.1093/RPD/NCAD040.
7. Subramanian K, et al. Complex implementation factors demonstrated when evaluating cost-effectiveness and monitoring racial disparities associated with [18F] DCFPyL PET/CT in prostate cancer men. Sci Rep. 2023;13(1):8321.
8. Mason C, Gimblet GR, Lapi SE, Lewis JS. Novel tracers and radionuclides in PET imaging. Radiol Clin North Am. 2021;59(5):887. https://doi.org/10.1016/J.RCL.2021.05.012.
9. Hashimoto F, Onishi Y, Ote K, Tashima H, Reader AJ, Yamaya T. Deep learning-based PET image denoising and reconstruction: a review. Radiol Phys Technol. 2024;17(1):24–46. https://doi.org/10.1007/s12194-024-00780-3.
10. Balaji V, Song TA, Malekzadeh M, Heidari P, Dutta J. Artificial intelligence for PET and SPECT image enhancement. J Nucl Med. 2024;65(1):4–12. https://doi.org/10.2967/jnumed.122.265000.
11. Bousse A, et al. A review on low-dose emission tomography post-reconstruction denoising with neural network approaches. IEEE Trans Radiat Plasma Med Sci. 2024;8(4):333–47. https://doi.org/10.1109/TRPMS.2023.3349194.
12. Seith F, et al. Simulation of tracer dose reduction in 18F-FDG PET/MRI: effects on oncologic reading, image quality, and artifacts. J Nucl Med. 2017;58(10):1699–705. https://doi.org/10.2967/jnumed.116.184440.
13. Gatidis S, Seith F, Schäfer JF, Christian la Fougère MD, Nikolaou K, Schwenzer NF. Towards tracer dose reduction in PET studies: simulation of dose reduction by retrospective randomized undersampling of list-mode data. Hell J Nucl Med. 2016;19(1):15–8.
14. Schaefferkoetter J, et al. Convolutional neural networks for improving image quality with noisy PET data. EJNMMI Res. 2020;10(1):105. https://doi.org/10.1186/S13550-020-00695-1.
15. Xu J, Gong E, Pauly J, Zaharchuk G. '200x low-dose PET reconstruction using deep learning'. arXiv preprint arXiv:1712.04119, 2017.
16 Jang SI, et al. Spach transformer: spatial and channel-wise transformer based on local and global self-attentions for PET image denoising. IEEE Trans Med Imaging. 2023;43(6):2036–49. https://doi.org/10.1109/TMI.2023.3336237.
17. Pan S, et al. Full-dose whole-body PET synthesis from low-dose PET using high-efficiency denoising diffusion probabilistic model: PET consistency model. Med Phys. 2024;51(8):5468–78. https://doi.org/10.1002/MP.17068.
18. Liu Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision. 2021. pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.
19. Wang YR, et al. Low-count whole-body PET/MRI restoration: an evaluation of dose reduction spectrum and five state-of-the-art artificial intelligence models. Eur J Nucl Med Mol Imaging. 2023;50(5):1337–50. https://doi.org/10.1007/S00259-022-06097-W.
20. Xie H, et al. Dose-aware diffusion model for 3D low-dose PET: multi-institutional validation with reader study and real low-dose data. arXiv preprint arXiv:2405.12996, 2024.

21. Yu B, et al. Robust whole-body PET image denoising using 3D diffusion models: evaluation across various scanners, tracers, and dose levels. Eur J Nucl Med Mol Imaging. 2025. https://doi.org/10.1007/S00259-025-07122-4.
22. Liu Q, et al. A personalized deep learning denoising strategy for low-count PET images. Phys Med Biol. 2022;67(14):145014.
23. Rogasch JMM, Hofheinz F, van Heek L, Voltin C-A, Boellaard R, Kobe C. Influences on PET quantification and interpretation. Diagnostics. 2022;12(2):451.
24. Bonardel G, et al. Clinical and phantom validation of a deep learning based denoising algorithm for F-18-FDG PET images from lower detection counting in comparison with the standard acquisition. EJNMMI Phys. 2022;9(1):36.
25. Quak E, Weyts K, Jaudet C, Prigent A, Foucras G, Lasnon C. Artificial intelligence-based 68Ga-DOTATOC PET denoising for optimizing 68Ge/68Ga generator use throughout its lifetime. Front Med. 2023. https://doi.org/10.3389/FMED.2023.1137514.
26. Weyts K, et al. Artificial intelligence-based PET denoising could allow a two-fold reduction in [18F]FDG PET acquisition time in digital PET/CT. Eur J Nucl Med Mol Imaging. 2022;49(11):3750–60. https://doi.org/10.1007/S00259-022-05800-1.
27. Mehranian A, et al. Deep learning–based time-of-flight (ToF) image enhancement of non-ToF PET scans. Eur J Nucl Med Mol Imaging. 2022;49(11):3740–9. https://doi.org/10.1007/S00259-022-05824-7/FIGURES/6.
28. Chaudhari AS, et al. Low-count whole-body PET with deep learning in a multicenter and externally validated study. NPJ Digit Med. 2021;4(1):127.
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015. pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
30 Singh G, Mittal A, Aggarwal N. ResDNN: deep residual learning for natural image denoising. IET Image Process. 2020;14(11):2425–34. https://doi.org/10.1049/iet-ipr.2019.0623.
31. Katsari K, et al. Artificial intelligence for reduced dose ^{18}F-FDG PET examinations: a real-world deployment through a standardized framework and business case assessment. EJNMMI Phys. 2021;8(1):25. https://doi.org/10.1186/S40658-021-00374-7.
32. Dadgar M, Verstraete A, Maebe J, D'Asseler Y, Vandenberghe S. Assessing the deep learning based image quality enhancements for the BGO based GE omni legend PET/CT. EJNMMI Phys. 2024;11(1):86.

Authors and Affiliations

Justine Maes[1] · Charles Carron[1] · Simon DeKeyser[2] · Tomas Brants[2] · Vicky De Ridder[2] · Amine Chaouki[3] · Camille Steenhout[4] · Stefaan Vandenberghe[5] · Yves D'Asseler[12] · Laurens Raes[6] · Azzam Abdalla Ibrahim[7] · Gerard Moulin-Romsee[8] · Isaac Kargar Samani[9] · Ludovic D'hulst[10] · Sezgin Ustmert[11] · Maarten Larmuseau[2]

✉ Maarten Larmuseau
maarten.larmuseau@nuclivision.com

1 Division of Nuclear Medicine, University Hospitals UZ Leuven, Louvain, Belgium

2 Nuclivision, Ghent, Belgium

3 Department of Nuclear Medicine, Chirec Hospital Group, Brussels, Belgium

4 Division of Nuclear Medicine and Oncological Imaging, CHU de Liège, Liège, Belgium

5 Department of Electronics and Information Systems, Medical Image and Signal Processing, Ghent University, Ghent, Belgium

6 Molecular Imaging and Therapy Research Group, Vrije Universiteit Brussel, Brussels, Belgium

7 Department of Nuclear Medicine, VieCurie Medisch Centrum, Venlo, The Netherlands

8 Department of Nuclear Medicine, AZ Monica, Antwerp, Belgium

9 Department of Nuclear Medicine, Centre Hospitalier EpiCURA, Baudour, Belgium

10 Department of Nuclear Medicine, AZ Sint-Lucas, Ghent, Belgium

11 Department of Nuclear Medicine, AZ Groeninge, Kortrijk, Belgium

12 Department of Nuclear Medicine, Ghent University Hospital, Ghent, Belgium