Validation of low-dose lung cancer PET-CT protocol and PET image improvement using machine learning

Ying-Hwey Nai[1], Josh Schaefferkoetter[1,2,3], Daniel Fakhry-Darian[1], Sophie O'Doherty[1], John J. Totman[1], Maurizio Conti[3], David W. Townsend[1,4], Arvind K. Sinha[5], Teng-Hwee Tan[6], Ivan Tham[7], Daniel C. Alexander[1,8], and Anthonin Reilhac[1]

[1] Clinical Imaging Research Centre, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[2] Joint Department of Medical Imaging, University Health Network, Toronto, Canada

[3] Siemens Medical Solutions USA, Inc., Molecular Imaging, Knoxville, TN USA

[4] Department of Diagnostic Radiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[5] Department of Diagnostic Imaging, National University Hospital, Singapore

[6] Department of Radiation Oncology, National University Cancer Institute, Singapore

[7] Radiation Oncology Centre, Mt Elizabeth Novena Hospital, Singapore

[8] Centre for Medical Image Computing and Department of Computer Science, University College London, UK

Corresponding author

Name: Ying-Hwey Nai, Ph.D.

Email: mednyh@nus.edu.sg, yinghweynai@yahoo.com

Phone: +65- 65167412

Address (Current): MD6, Centre for Translational Medicine, National University of Singapore, 14 Medical Drive, #B1-01, Singapore 117599

ORCID: 0000-0002-6634-8785

Abstract (250 / 250 words)

Purpose: To conduct a simplified lesion-detection task of a low-dose (LD) PET-CT protocol for frequent lung screening using 30% of the effective PETCT dose and to investigate the feasibility of increasing clinical value of low-statistics scans using machine learning.

Methods: We acquired 33 SD PET images, of which 13 had actual LD (ALD) PET, and simulated LD (SLD) PET images at seven different count levels from the SD PET scans. We employed image quality transfer (IQT), a machine learning algorithm that performs patch-regression to map parameters from low-quality to high-quality images. At each count level, patches extracted from 23 pairs of SD/SLD PET images were used to train three IQT models – global linear, single tree, and random forest regressions with cubic patch sizes of 3 and 5 voxels. The models were then used to estimate SD images from LD images at each count level for 10 unseen subjects. Lesion-detection task was carried out on matched lesion-present and lesion-absent images.

Results: LD PET-CT protocol yielded lesion detectability with sensitivity of 0.98 and specificity of 1. Random forest algorithm with cubic patch size of 5 allowed further 11.7% reduction in the effective PETCT dose without compromising lesion detectability, but underestimated SUV by 30%.

Conclusion: LD PET-CT protocol was validated for lesion detection using ALD PET scans. Substantial image quality improvement or additional dose reduction while preserving clinical values can be achieved using machine learning methods though *SUV* quantification may be biased and adjustment of our research protocol is required for clinical use.

Keywords

Positron emission tomography (PET); Lung cancer; Machine learning; Lesion detection

Abbreviations

SD: Standard Dose, LD: Low-Dose, ALD: Actual Low-Dose, SLD: Simulated Low-Dose, IQT: Image Quality Transfer, PET: Positron Emission Tomography, CT: Computed Tomography, VOI: Volume of Interest, GL: Global Linear regression, ST: non-linear Single Tree regression, RF: non-linear Random Forest regression

SUV: Standardized Uptake Values, *NMSE*: Normalized Mean-Square-Error, *NMedSE*: Normalized Median-Square-Error, *STD*: Standard Deviation, *CNR*: Contrast-to-Noise Ratio, *AUROC*: Area Under Receiver Operating Characteristics Curve (Calculated parameters are italicized for ease of differentiation between names of images and metrics)

1. Introduction

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer death worldwide with 2.09 million cases in 2018 [1]. As early lung cancer is typically asymptomatic, it is often diagnosed only at an advanced stage [2,3]. The systematic screening of subjects at risk using low-dose (LD) computed tomography (CT) can potentially lead to a reduction in lung cancer mortality by about 20% as demonstrated in large screening trials [3,4]. However, screening using CT has high sensitivity for detecting small cancerous lesions [4] but at the cost of false-positive rates over 90% [5]. This leads to over-diagnosis resulting in unnecessary complications of incurring excessive cost, anxiety, and inconvenience associated with unnecessary investigation of subjects without lung cancer. Positron Emission Tomography (PET) imaging with [^{18}F]fluorodeoxyglucose ([^{18}F]FDG) provides crucial metabolic information, which allows for better differentiation between malignant tumors and benign tumors, and facilitates the detection of nodal and distant metastases, particularly in the bone [2]. PET was shown to yield high sensitivity and specificity, which, when combined with CT anatomical information, overcomes the shortcoming of CT screening [3]. However, a PET-CT scan adds a median radiation exposure of 4.0 (1.2 – 28.8) mSv to patients with a 4 MBq/kg injection of [^{18}F]FDG, compared to that from LD CT of about 1.0 mSv [6]. Thus, the widespread use of PET-CT for the frequent screening of patients at risk, and disease surveillance after treatment, would require LD PET-CT protocols to minimize radiation exposure [7].

While many studies investigated lesion detectability and staging capability with LD CT, little effort has gone into investigating the clinical use of LD PET protocol for lung cancer screening [7]. Reducing the radiation dose of PET increases the statistical noise in the reconstructed images, which can affect the ability to detect lung nodules, especially in early stages where they are often of limited size. We previously investigated the diagnostic value of LD PET images for lung lesion detection. Our results showed that PET dose could be reduced to 18.5 MBq leading to an effective patient dose of 0.4 mSv [6-8] without losing clinical significance. However, in this previous work, the assessment was performed using simulated LD PET (SLD) images that were generated by discarding detected counts from the actual standard dose (SD) PET scans. Simulated and actual low-dose (ALD) PET images may exhibit different image quality due to differences in the number of randoms, the level of deadtime of the system, and the quality of the attenuation correction. With this regard, we recently demonstrated that count decimation was a quantitatively accurate technique for emulating low-count scans [9]. In this work, we found no significant measured differences, between ALD and SLD images, in the standard metrics used to quantify regions of focal uptake such as mean and max *SUV*. In addition, we showed that despite the difference in random counts levels, ALD and SLD reconstructed images exhibited similar noise properties. Nevertheless, these metrics or image noise properties do not necessarily correlate with visual lesion detection performance as intrinsic variability of human observer's performance is not only dependent on image noise and their performance consistency decreased at low signal-to-noise ratio (SNR) [10]. Therefore, for its clinical acceptance, it is essential to validate the LD-PET protocol using actual lesion detection task.

Finally, in this previous work, no attempt was made to investigate the applicability of advanced image processing techniques in the context of lesion detection with the goal to increase the clinical value of PET images reconstructed from low-statistics scans. Machine Learning (ML) has recently gained immense credits in medical applications and has for instance proven to be an efficient avenue for mapping suboptimal image quality to high image quality with richer image content and lower noise. Albeit already well-accepted in structural imaging using

CT and Magnetic Resonance Imaging (MRI), ML for image quality improvement is still burgeoning for functional imaging such as PET. In this area, a few authors have nevertheless recently demonstrated the superiority of ML-based approaches over several standard denoising methods, including Gaussian filter, anatomical-guided non-local mean filter, and maximum a posterior reconstruction with Quadratic prior and relative difference prior, in terms of image quality and trade-off between noise and bias [11,12]. In these studies, classical metrics such as image error, lesions contrast to noise ratio, or difference in lesion *SUV* were employed for the assessment of the proposed methods and, to our knowledge, the advantage of ML-based approaches in the context of visual lesion detection task remains to be demonstrated.

The first objective of this study was to conduct evaluation and validation of the LD PET-CT protocol using a simplified lesion-detection task by comparing, for the first time, on the same cohort, SD and ALD PET-CT. The second objective was to assess the image quality and detection performance obtained with a ML algorithm called Image Quality Transfer (IQT) [13] and to determine to which extent this technique allows to increase the clinical value of LD PET images or to further reduce the scan statistics without trading the detection accuracy.

2. Methods and Materials

2.1. Image acquisition

Thirty-three subjects (weight: 62.6 ± 14.0 kg) with biopsy-proven malignancy in the lungs were recruited for this study. Written informed consent was obtained before their recruitment into the study and ethics approval was given by the Institutional Review Board. The PET data were acquired on the Biograph mCT (Siemens Healthcare Molecular Imaging, USA) in list-mode format. Thirteen subjects had LD PET scans followed by SD PET scans, while the remaining twenty subjects had only SD PET scans.

The LD PET-CT protocol consisted of 1 (3 subjects only) or 2 bed-positions of 10 min each, covering the entire lungs using our previously described imaging protocol [9], 60 min post-injection of [^{18}F]FDG (27.0 ± 4.00 MBq). For the LD protocol, the CT parameters were 120 kVp, 40 mAs (careDose) with a pitch of 1. The use of LD-CT protocol has been demonstrated to retain the accuracy of CT-derived attenuation correction of SD CT protocol [9,14]. The SD PET-CT acquisitions followed the same protocol except for the injection of a SD of [^{18}F]FDG (221 ± 5.68 MBq) as well as the use of 140 mAs (careDose) and a pitch 1.5 for the CT acquisition. Subjects were instructed to void the bladder before being positioned on the scanning bed. Also, an additional 30 min waiting time was imposed before the SD injection for subjects who underwent LD PET scans to limit the contribution of the remaining activity into the acquired SD scans. We estimated that activity from the 1st scan accounted for only about 6% of the total activity at the time of the SD scan [9]. The average estimated total effective dose is 2.39 ± 1.83 mSv and 5.34 ± 2.32 mSv for all LD-CT and SD-CT scans.

2.2. Low-dose simulation and image reconstruction

A total of 7 count levels of SLD with expected net trues count levels (prompts – randoms events) of 0.25, 0.5, 1.0, 2.0, 5.0, 7.5 and 10 million (M) were generated from the SD PET data by randomly discarding events from the list-mode using our previously published method. For each of these count levels, 5 independent realizations were generated per subject from their SD PET scan. The number of true counts, average percentage of true counts in the SLD images compared to that of SD images, and the corresponding estimated injected dose and effective dose are reported in Table 1 for each of the 7 count levels, as well as that for ALD.

Count levels	**0.25**	**0.5**	**1**	**2**	**5**	**7.5**	**10**	**ALD***	**SD**
Counts ($x10^6$)	0.253 ± 0.002	0.506 ± 0.005	1.012 ± 0.010	2.025 ± 0.019	5.062 ± 0.048	7.593 ± 0.072	10.123 ± 0.096	17.261 ± 8.257	128.264 ± 47.120
% counts / full counts	0.20	0.39	0.79	1.58	3.95	5.92	7.89	12.2	100
Dose (MBq)	0.44	0.87	1.74	3.49	8.72	13.1	17.4	27.0	221
Effective PET dose (mSv)	0.01	0.02	0.03	0.07	0.17	0.25	0.33	0.51	4.20
Effective PETCT dose (mSv)	2.40	2.41	2.42	2.46	2.56	2.64	2.72	2.90	9.54
% Reduction in overall dose	74.9	74.8	74.6	74.2	73.2	72.3	71.5	69.6	100

Table 1: Actual counts (average ± standard deviation) of the simulated low-dose (SLD) realizations at each count level and actual low-dose (ALD), and average percentage of true counts to the true counts of standard dose (SD) PET scan. The scans were acquired with 1-2 bed positions of 10 min each. The average dose of SLD at each count level was approximated by multiplying the fractional true counts at each level with the averaged standard dose of 221 MBq injected to all 33 subjects. The effective dose was estimated with reference to ICRP Publication 128. * Actual counts and injected dose of 13 subjects only.

All PET scans were reconstructed using JSRecon and e7tools with the constructor implementation of the ordinary Poisson ordered-subset expectation-maximization (OP-OSEM) algorithm, with all corrections including time-of-flight information, and system resolution modeling with 2 iterations and 21 subsets. Each reconstructed image consisted of a 400 × 400 × 171 matrix with voxel size of 2.04 mm × 2.04 mm × 2.03 mm. The PET images were converted to NIFTI format and normalized by injected dose and subject's weight to obtain the standardized uptake value (*SUV*) SD, ALD, and SLD PET images.

2.3. IQT training

IQT is a machine learning technique that was originally proposed to enhance information content and resolution of low-quality diffusion MR images [13]. In this work, the patch-regression process that learns the correction mapping from matched pairs of low-quality and high-quality PET images was adapted to identify similarities between pairs obtained at the same spatial resolution but reconstructed from scans with different statistics; a LD scan and its corresponding SD scan. The mapping was learned independently for each voxel from its direct neighbors within 3D patches. Three regression models: Global linear (GL), non-linear single tree (ST), and random forest (RF) models were implemented and trained. Overfitting was controlled by using half of the training data for validation. The learned mappings were then used to map low-quality, LD PET images to high-quality, SD images, and the training performance was evaluated using the root median squared error (*RMedSE*) between the estimated and ground-truth SD images. IQT was written in Matlab and the training and output estimation steps were carried out on Matlab version R2017b (The MathWorks, Inc., Natick, Massachusetts, US).

The general process is illustrated in figure 1. To allow comparison with ALD, data of the last 10 subjects who had both ALD and SD scans were selected for evaluation while the remaining twenty-three subjects were used for training (Fig. 1, step 1). Masks were generated for all subjects covering all the lung space and body on the SD PET images but excluding the top and bottom 10 transverse slices, which contained higher noise due to their proximity to the end of the field of view (Fig. 1, step 2). For each count level, 5 realizations of SLD were generated from the 23 SD training images and 3 from the 10 SD evaluation images, leading to 115 and 30 matched pairs of SD/SLD dataset for training and evaluation (Fig.1, step 3). The IQT algorithms were trained to map SLD to SD images, with patch regression performed using only voxels within the mask and with cubic patch sizes of 3 and 5

voxels (Fig. 1, step 4). For each count level, each trained model was finally used to estimate the SD estimate from the SLD images of the ten unseen subjects (Fig. 1, step 5).

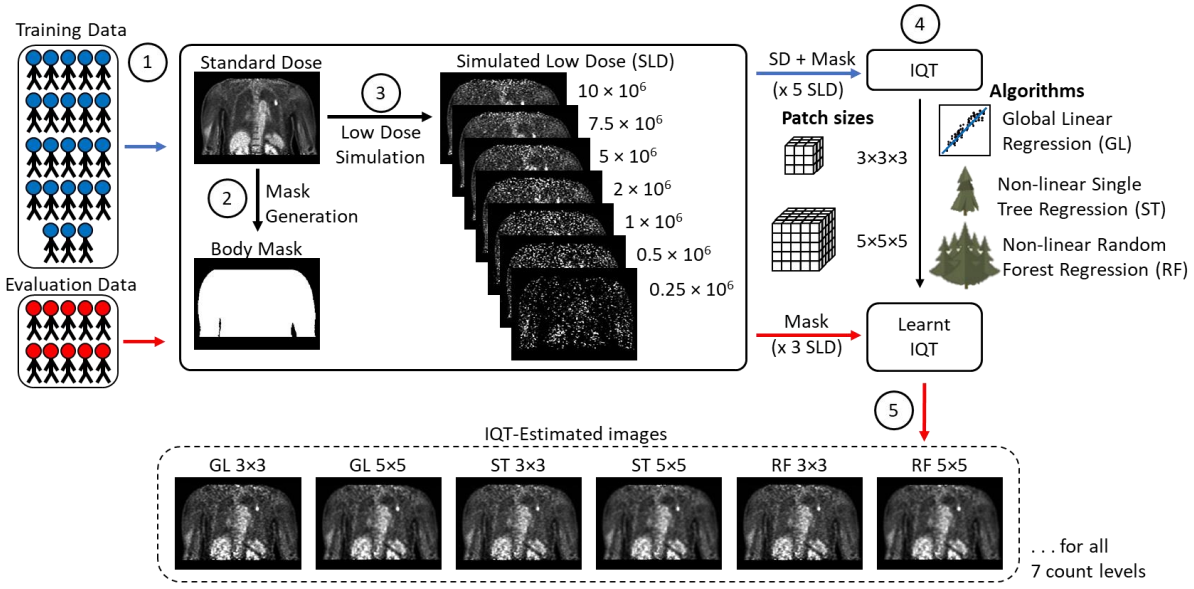


Fig. 1: Overview of steps taken to (1) split the dataset into training and evaluation datasets, (2) generate mask from standard dose (SD) PET images, (3) simulate the low-dose (SLD) images from SD PET images, (4) train the IQT algorithms using a set of SD, mask and SLD images for each subject at each count level and (5) estimate SD images from SLD images using the trained IQT algorithms – GL, ST, and RF regressions, with 3 and 5 cubic patch sizes, for all SLD of each evaluation subject for all 7 count levels.

2.4. Image analysis

A total of 19 isolated lesions of various sizes and contrasts were identified within and near lung regions, including bone and lymph node metastasis from the 10 evaluation subjects using their SD image. The presence and location of true isolated lesions were confirmed by an experienced nuclear medicine physician. ALD images were registered to the SLD images for ease of visual comparison of the same lesion. Lesion volume of interest (VOI) was obtained by seeding on the highest value of each suspected lesion on the SD and ALD PET images with a 40% threshold. The lesion's volume was defined as the number of voxels in the lesion multiplied by the PET image's voxel volume of 8.45 mm^3 [7]. The SD-VOIs were used to obtain the *SUV* of lesions in the SD and SLD images, while the ALD-VOIs were applied only to ALD images. The length of the lesion was defined on the major axes in pixels then multiplied by the voxel size of 2.04 cm. *SUV*s were also measured for each subject in normal tissues of the lung air spaces, muscles, and liver using cubic VOIs of 32.6 x 32.6 x 32.5 mm^3 that were defined from the SD image.

Normalized mean-square-error (*NMSE*) and normalized median-square-error (*NMedSE*) of the SLD and IQT-estimated images from the original SD images were calculated within the masked region [11]:

$$NMSE = \frac{\|I_{\underline{SLD}}^{IQT} - I_{SD}\|_2^2}{\|I_{SD}\|_2^2} \quad (1)$$

Where I_{SD} refers to the entire reference image, in this case, the SD image and $I_{SLD/IQT}$ refers to SLD or IQT-estimated images. Mean is replaced with the median to obtain *NMedSE*.

Contrast-to-noise ratio (*CNR*) of the lesion to background was determined using the following equation [15]:

$$CNR = \frac{SUV_{MEAN_{Lesion}} - SUV_{Mean_{Background}}}{SUV_{STD_{Background}}} \quad (2)$$

where SUV_{STD} refers to the standard deviation of *SUV* measured in the lung tissue.

2.5. Lesion Detection Task

A lesion-detection task was carried out with two distinct objectives: 1) to validate using the ALD scans, the LD PET-CT protocol for lesion detection and 2) to investigate performance improvement in the lesion detection obtained with IQT machine learning. The best IQT algorithm that yielded the most reliable and accurate estimated images were determined using *NMSE* and *CNR*, together with visual assessment. The detection task was performed by two experienced post-doctoral researchers in PET and one medical physicist (MSc) with 10 years of lesion detection and image segmentation experience.

Two sets of 266 (19 lesions x 7 count levels x 2 (paired lesion-present and lesion-absent)) images each were generated, showing the transverse, coronal and sagittal sections of SLD images and their corresponding images, estimated using the best algorithm in IQT. All the images were set to the same image intensity range of 0-5. The images were stratified by image-type (e.g. simulated vs. IQT-estimated) and randomly presented regardless of count levels to avoid bias [16]. Cubic VOIs with dimensions of 32.6 x 32.6 x 32.5 mm were placed over each lesion on the SLD and corresponding IQT-estimated images centered at voxel with the highest *SUV* value. Similarly, VOIs of the same dimension were placed in the lung tissues, away from surrounding lesions and other tissues (e.g. bone), thus generating a population of matched lesion-present (Fig. 2a) and lesion-absent images (Fig. 2b). Besides, high-quality estimates of ALD images (IQT-ALD) were produced using the best IQT model trained using the 10M count level SLD dataset, which corresponds to the closest statistics. Raters were given the SLD dataset on the first day followed by the IQT-estimated dataset on the next day. The ALD dataset was given a few days later with the IQT-ALD dataset on the following day, using a set of 19 matched lesion-present and lesion-absent images. This was done to avoid bias in rating based on image type and for the observers to remember the images.

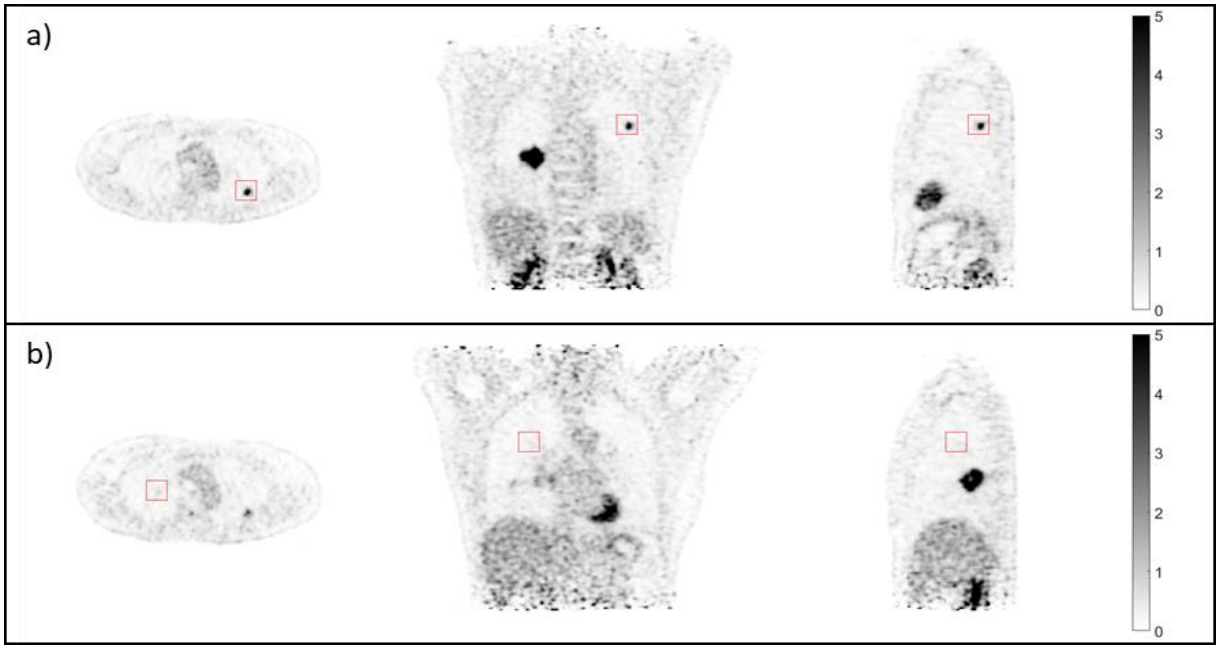


Fig. 2: Three orthogonal views (from left to right: transverse, coronal, and sagittal) of SLD or IQT-estimated images shown in lesion detection task for (a) lesion-present and (b) lesion-absent in the cubic volume, highlighted in red. The example used here shows the SLD image of a subject at 7.5M count level.

Human observers were asked to use a five-point rating scale to report their confidence in identifying the lesion, with 1 representing 100% confidence of no lesion and 5 representing 100% confidence of a lesion [7]. A lesion was assumed to be detectable by the observer if it was given a confidence rating of 3 or more, to mimic the clinical situation of detecting all suspicious lesions. They were given the SLD dataset on the first day followed by the IQT-estimated dataset on the next day. The ALD dataset was given a few days later with the IQT-ALD dataset on the following day. This was done to avoid bias in rating based on image type and for the observers to remember the images. For each observer, the average confidence ratings for lesion-present and lesion-absent of SLD and IQT-estimated images were determined at each count level, as well as for ALD and IQT-ALD images. The area under the receiver operating characteristics curve (*AUROC*) was estimated non-parametrically for the performances of human observers on ALD, SLD, and IQT-estimated images.

3. Results

3.1. Lesion Delineation

A total of 19 isolated lesions were delineated in 10 subjects as shown in Table 2. The volumes of the lesions generated using SD and ALD images were 0.92 ± 1.69 cm^3 (Median, [1st quartile, 3rd quartile]: 0.27 [0.18, 0.62]) and 1.07 ± 1.82 cm^3 (0.41 [0.18, 0.49]). The SUV_{MEAN} of SD and ALD images were 4.46 ± 2.63 (3.64 [2.54, 5.85]) and 4.08 ± 2.52 (3.78 [2.03, 5.45]), while the SUV_{MAX} were 7.17 ± 4.56 (6.14 [3.42, 9.45]) and 6.59 ± 4.28 (5.97 [2.90, 8.26]). Significant difference was only found in SUV_{MEAN} of the lesions generated using SD and ALD images (Wilcoxon signed-rank, $p < 0.05$). However, after excluding 3 lesions in subjects that moved during the ALD scans, no significant difference was observed.

Lesion	Subject	Weight (kg)	Histology	TNM Stage	Lesion Type	SD PET				ALD PET			
						Vol	Length	SUV_{MEAN}	SUV_{MAX}	Vol	Length	SUV_{MEAN}	SUV_{MAX}
1	1	44.1	ADC	IV	Lung nodule	0.55	1.04	4.06	8.22	0.50	1.00	3.50	6.33
2	2	55.7	ADC	IIIA	Lung nodule	1.75	1.61	4.01	7.97	1.70	1.63	3.79	8.21

3	3	66.6	ADC	IIIB	Lung nodule	0.40	1.28	1.68	2.27	0.43	1.41	1.43	2.19
4					Lymph node	0.18	0.70	2.74	3.71	0.47	1.14	1.49	2.04
5					Lung nodule	0.19	0.97	6.01	7.95	0.18	1.12	6.19	8.31
6	4	33.8	ADC	*	Bone	0.13	0.83	2.36	3.14	0.11	0.64	2.26	3.12
7					Lymph node	0.27	1.05	1.95	2.62	0.08	0.77	4.04	5.53
8					Bone	0.16	0.71	2.90	3.96	0.18	1.17	2.33	3.14
9	5	77.0	RCC	IV	Lymph node	0.69	1.13	3.64	4.94	0.41	1.05	3.78	5.20
10					Lung nodule	0.24	0.91	1.88	2.55	0.45	1.24	1.33	1.84
11	6	64.6	ADC	IV	Lung nodule	0.13	0.83	4.79	6.14	0.13	0.84	5.10	7.17
12					Lung nodule	0.10	0.69	10.36	14.19	0.13	0.71	9.10	12.22
13					Lung nodule	0.23	0.85	9.35	17.23	0.24	0.88	8.66	14.48
14					Bone	0.43	1.32	5.70	10.68	0.48	1.39	5.80	11.22
15	7[#]	73.9	ADC	*	Lung nodule	0.33	1.24	3.58	4.77	0.30	1.43	1.48	2.12
16					Lung nodule	0.24	0.84	6.10	11.62	0.38	1.00	4.39	7.78
17	8	59.8	ADC	IV	Lung nodule	3.30	2.61	2.71	6.46	3.17	2.53	2.69	5.97
18	9	45.1	ADC	IA3	Lung nodule	7.12	2.50	8.80	14.66	7.25	2.55	8.33	15.62
19	10[#]	74.8	ADC	*	Lymph node	1.06	1.84	2.18	3.05	3.79	5.00	1.81	2.69

Table 2: Characteristics of 19 identified lesions in the selected 10 evaluation subjects with SD and ALD PET images. Vol refers to the volume of the lesion (cm^3). Length refers to the length of the major axes of the lesion (cm). ADC = Adenocarcinoma, RCC = Renal cell carcinoma. *Staging was not carried out. [#]Lesions shifted due to subject motion in ALD.

3.2. Image Error – NMedSE and *NMSE*

Fig. 3 shows the average *NMSE* and *NMedSE* measured between the SLD or IQT-estimated images and the SD images from the 10 unseen subjects. As expected, *NMSE* and *NMedSE* increased with decreasing true counts for both SLD and IQT-estimated images. However, we can appreciate the substantial reduction of the *NMSE* and *NMedSE* that was obtained using IQT. Moreover, the patch size of 5 yielded more accurate estimates than the patch size of 3. Overall, RF regression yielded the best performances across all count levels. IQT was able to estimate high-quality images with *NMSE* of less than 0.2 and *NMedSE* of less than 0.05 down to 5M count levels, beyond which the errors increased more steeply. The root mean square error and root median square error resulting from the IQT correction process are shown in supplementary figure 1.

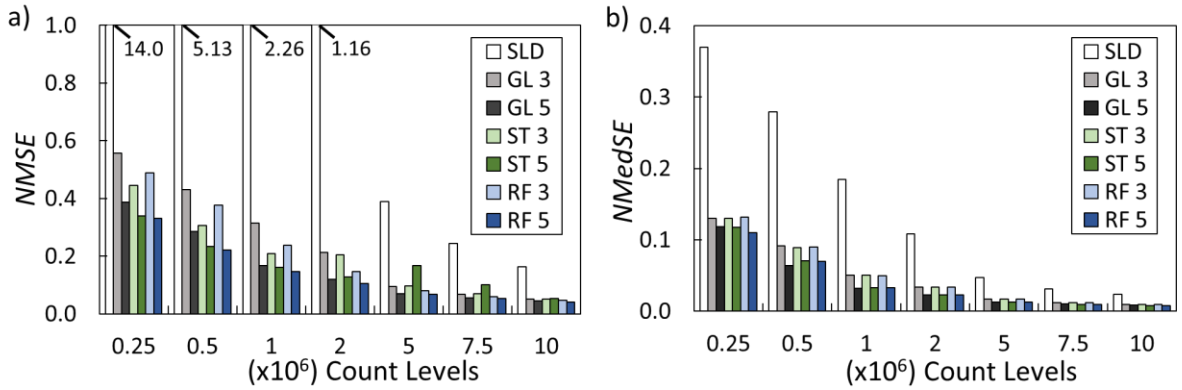


Fig. 3: (a) Average normalized mean-square error (*NMSE*) and (b) average normalized median-square-error (*NMedSE*) of the SLD images (white), and the IQT-estimated images, using GL (gray), ST (green) and RF (blue) regression algorithms with patch sizes of 3 (light-shade) and 5 (dark-shade) with respect to the SD images of 10 evaluation subjects within the masked regions. Each evaluation subject has 3 SLD and 3 corresponding IQT-estimated images.

3.3. Lesion-to-background contrast

The *CNR* increased steadily with increasing true counts for both SLD and IQT-estimated images (Fig. 4). The proposed LD protocol yielded ALD images with a *CNR* that is half of the SD images while reducing by roughly 8 times the subject exposure (Table 1). All IQT models improved the *CNR* of the estimated images across all

count levels, however with a noticeable advantage for the ST and RF models using a patch size of 5. Larger patch size of 7 or more may improve the results, but we are limited by our computational capabilities. The relative increase in *CNR* obtained with IQT compared to the unprocessed SLD images decreases slightly with increasing count levels, with for instance a relative increase of 121% at 0.25M count level and 82% at 10M count level observed with RF5. We can note that RF5 and ST5 could estimate images from the 7.5M count level SLD images with *CNR* higher than the proposed low-dose protocol (ALD), potentially allowing an additional reduction by 50% of the injected dose from 27.0 MBq with the current protocol down to 13.1 MBq.

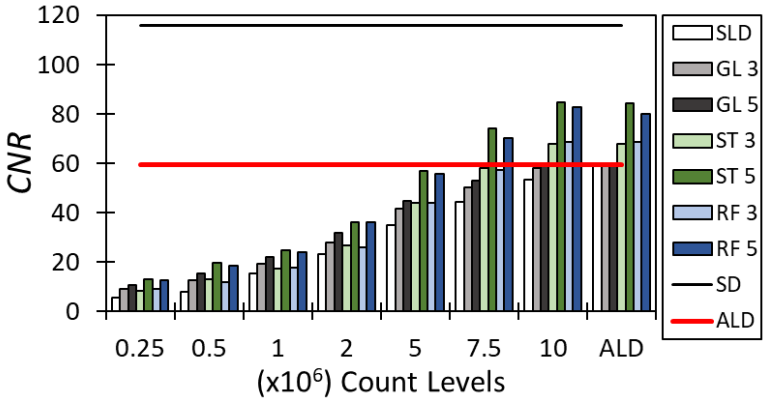


Fig. 4: Average contrast-to-noise (*CNR*) of SLD/ALD (white), and the IQT-estimated images, using GL (gray), ST (green) and RF (blue) algorithms with patch sizes of 3 (light-shade) and 5 (dark-shade) with respect to the SD images for 10 evaluation subjects. Each evaluation subject has 3 SLD and 3 corresponding IQT-estimated images. The black and red lines indicate the *CNR* of SD and ALD images.

3.4. Bias in SUV

Fig. 5 shows the SUV_{MEAN} *and* SUV_{MAX} of the lesions, lung air spaces, muscles, and liver of the SD, SLD and IQT-estimated images. The bias in SUV_{MEAN} of 19 lesions resulting from the IQT correction process are shown in supplementary figure 2. We can note that the SUV_{MEAN} and SUV_{MAX} of the lesions were underestimated by over 30% even for ALD and SLD images of 10M count levels.

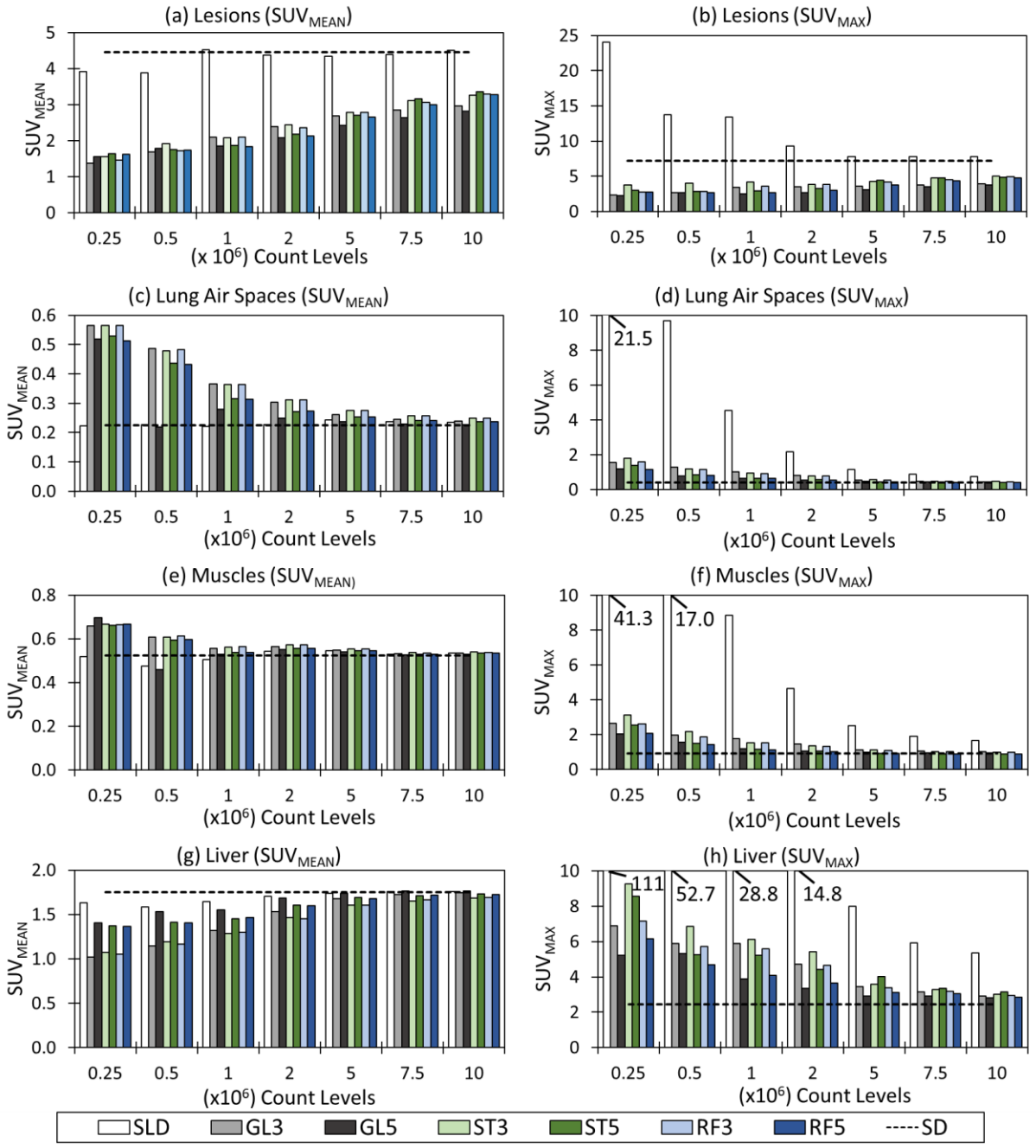


Fig. 5: (a) SUV_{MEAN} and (b) SUV_{MAX} of all 19 lesions, (c) SUV_{MEAN} and (d) SUV_{MAX} of lung air spaces, (e) SUV_{MEAN} and (f) SUV_{MAX} of muscles, and (g) SUV_{MEAN} and (h) SUV_{MAX} of liver. The values were averaged across 3 realizations of SLD (white) and their corresponding IQT-estimated images using global linear (GL, gray), non-linear single tree (ST, green) and non-linear random forest (RF, blue) regression algorithms, with 3×3×3 (light shade) and 5×5×5 (dark shade) patch sizes. The dotted line refers to the averaged values obtained using the SD images of 10 evaluation subjects. Significant differences in (a) SUV_{MEAN} and (b) SUV_{MAX} of lesions (Wilcoxon signed-rank test, $p < 0.05$) were generally found across different count levels within each image-type and across the different image types at each count level except between SLD and SD above 1M counts.

3.5. Lesion Detection Task

Fig. 6 shows the coronal views of the SLD and IQT-estimated PET images obtained from the 7 different count levels for one subject. The corresponding SD, ALD PET images as well as the mask generated from the SD image are also shown. This subject presents only one small lesion that can be observed with reasonable confidence from

both SLD and IQT-estimated images down to about 2M count level. Below this level, the SLD image demonstrated unclear tissue boundaries due to decreasing true counts. IQT-estimated images showed improved image quality across all count levels compared to the SLD images with an observable image superiority, in terms of clearer tissue boundaries and reduced noise, when estimated with ST and RF regression using a patch size of 5. However, the lesion can hardly be observed from 2M count level and below. As observed in Fig. 6, ST and RF regressions using a patch size of 5 produced image estimates from the 5M and 7.5M count levels of similar or higher quality than the ALD images.

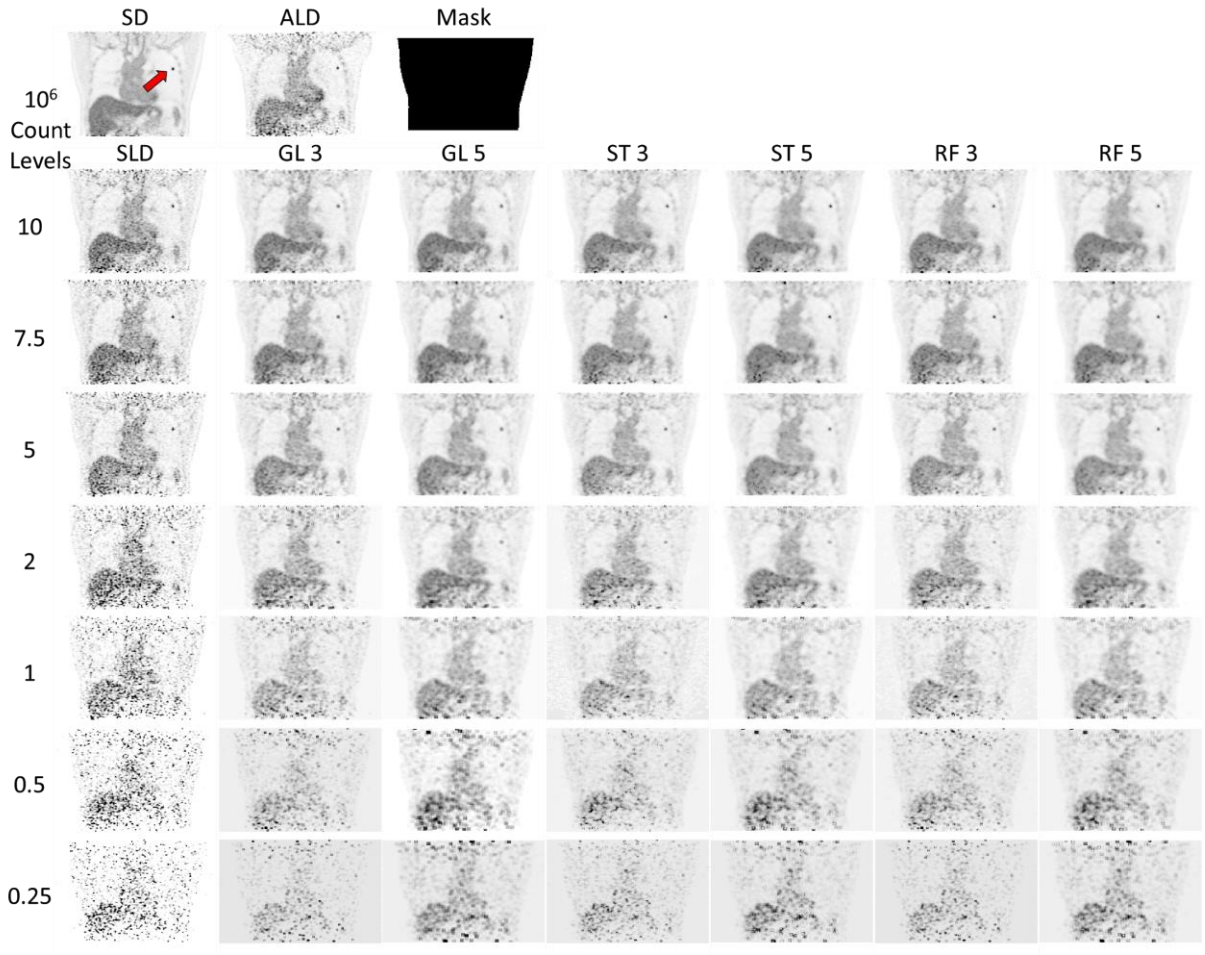


Fig. 6: Image qualities of the SD, ALD, and the SLD and IQT-estimated images at 7 different count levels of one subject with one small lesion (highlighted by the red arrow) visible in the selected coronal slice. IQT-estimated images were estimated using GL, ST, and RF regression algorithms, with cubic patch sizes of 3 and 5. All the images are viewed at the same image intensity range at the same slice. The body mask was generated using the SD image for IQT-training.

Although ST with cubic patch size of 5 generally yielded the smallest biases in SUV_{MEAN} *and* SUV_{MAX} (refer to supplementary fig. 2) and highest *CNR* (Fig. 4), RF with cubic patch size of 5 generally resulted in smallest image error and the quality of the estimated images appeared higher during the visual assessment. As such, the images estimated using RF with cubic patch size of 5 were selected for further evaluation in the lesion detection task.

Fig. 7 shows the average confidence level rated by the three observers (a), *sensitivity* (b), *specificity* (c), and the average *AUROC* (d) for ALD, SLD, and for the corresponding IQT-estimated images obtained with RF5. The confidence rating of true lesions shown in Fig. 7a steadily increased with increasing count levels and eventually

plateaued at 4.7 from 7.5M count levels, indicating that not all lesions were identified with 100% confidence. We can observe that IQT did not increase the confidence for the detection of true lesion except at very low counts (< 1M). This higher confidence level at low counts did not translate in an increase of the sensitivity (Fig. 7b). However, non-lesions were rated with higher confidence in IQT-estimated images than on the SLD images and with a much smaller variance across observers. The corresponding reduction in the false-positive rate led to a substantial increase in the specificity (Fig. 7c) and of the AUROC (Fig. 7d). The *AUROC* also showed that without IQT, only ALD images yielded AUROC close to 1, providing further validation evidence for this LD PET protocol. However, the use of IQT allowed a further reduction of the dose down to 7.5M count levels while preserving true/false lesion discrimination power. On the whole, IQT substantially improved the specificity across all count levels than SLD images.

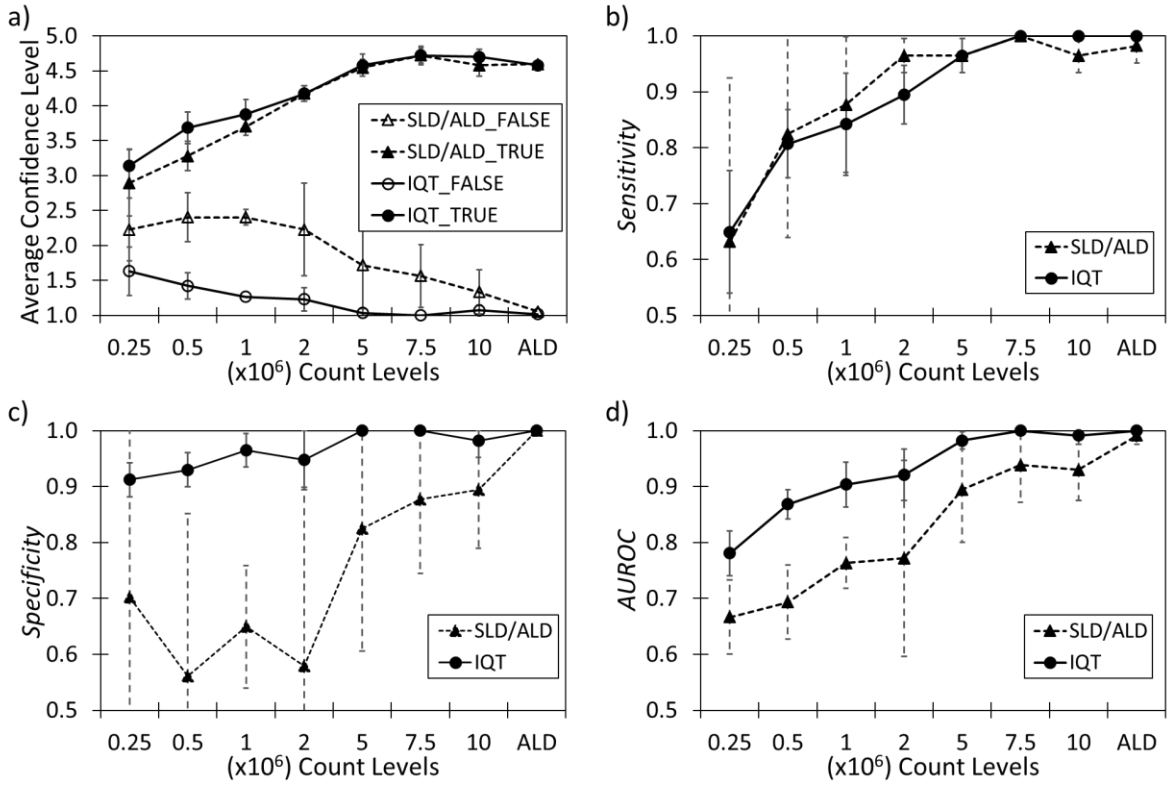


Fig. 7: Results of lesion detection tasks carried out by 3 human observers: (a) confidence rating of SLD or ALD (Δ--) and IQT-estimated (○-) images of true / lesions (filled marker) and false / non-lesion regions (unfilled marker), and the (b) *Sensitivity*, (c) *Specificity* and (d) area under the receiver operating curve (*AUROC*), in lesion detection for SLD (▲--) and IQT-estimated (●-) images. The parameters were averaged over the 3 human observers and the STD of their performance are shown in the error bar across the different count levels of SLD images and ALD images, each consisting of 19 matched lesion-present and lesion-absent images.

4. Discussion

In this study, we evaluated the image quality of LD PET images and validated their clinical value for lesion detectability. The feasibility to use IQT, a machine learning regression method, to increase the image quality of low-statistics scans or to further reduce the dose while maintaining the image quality for lesion detections was also investigated.

4.1. Validation of the low-dose PET-CT protocol

We first conducted a validation of a LD PET-CT protocol for frequent lung screening. This was developed from previous work [9] that showed using simulated PET data, the PET dose could be reduced to 8% of a SD (roughly 18.5 MBq) without losing clinical significance, leading to an effective patient dose of 0.4 mSv. However, as pointed out in this preliminary work, it is preferable to use a dose slightly higher than this lower limit as the benefit of improved accuracy from PET would far outweigh the small risk associated with these levels of radiation. Under the LD PET-CT protocol in this study, the subjects received an average dose of 27 MBq of [^{18}F]FDG, which corresponds to 12% of a standard PET dose, with a 55% reduction in CT dose. The results showed that although lesion contrast was reduced by about half in LD PET, there was no significant difference in lesion volume, length, $SUV_{MEAN,}$ and SUV_{MAX}. We then assessed whether there was any loss in lesion detectability with the LD PET-CT protocol as compared to SD PET-CT. A lesion detection task was carried out using a matched dataset of lesion-present and lesion-absent images instead of asking the human observers to locate and detect all the lesions. This was done to ensure a more consistent and error-free method of comparing the ability to detect the same lesion and differentiate lesion from non-lesion region for a large number of datasets.

The outcome of this task was only one missed lesion by one rater, leading to a sensitivity of 0.98. The missed lesion is close to the lung wall, and is small (see lesion 10 in Table 2) making it hard to distinguish it from the background noise and texture (refer to supplementary fig. 3). No false positive was observed with the ALD images leading to a specificity of 1. Note that with the recommended dose based on our previous work (SLD at 10M count level), the task resulted in a total of 6 false positives and 2 false negatives among three raters, leading to a sensitivity of 0.965 and a specificity 0.895 and supporting our choice to slightly increase the dose. We can also observe from Fig.7 that, at least down to 5M count levels, decreasing the image quality did not impact much the sensitivity but had a significant effect on the specificity, meaning that under higher noise conditions we are tempted to see more lesions. With this regard, as discussed in the next section, IQT played a major role and led to improved image quality allowing raters to better distinguish true lesions from background noise.

4.2. Feasibility of LD PET-CT protocol and use of AI in improving image quality

In recent years, artificial intelligence has shown potential in recovering image quality of LD PET images for oncology and neurology applications [11,17-19]. In addition, computer-aided detection (CAD) systems were also developed to support the detection of lung nodules using machine learning or deep learning. Some studies showed the use of automated detection of lesions at ultralow PET-CT dose of 0.11 mSv using deep learning [17,20]. The use of AI showed great potential even at ultralow dose, but the final diagnosis still relies on the radiologist's decision to agree with the CAD's results. Therefore, the image quality of the LD PET-CT images must still be good to ensure the detection of all lesions while ensuring no false positives. These previous studies showed that images enhanced using ML approaches are characterized by higher CNR and SNR mainly due to noise removal.

Our results showed indeed reduced noise in IQT-estimated images resulting in higher *CNR* (Fig. 4) and lower image error (Fig. 3) with *NMedSE* below 0.1 and *NMSE* below 0.2 even down to 5M count level (3.95% counts of SD images). These results are relatively similar to those obtained by Lu et al [11] with LD of 10% counts of SD images. However, we had higher biases in SUV_{MEAN} of about 30% compared to Lu et al [11] who obtained about 15% bias in SUV_{MEAN} using deep learning, with 10 lesions of fairly similar length (0.85 ± 0.52 mm) to the lesions used in our study. This may indicate that deep learning was able to better estimate the uptake in lesions.

We intend to implement the more recent IQT techniques that utilize deep learning via convolutional neural networks, which might lead to better estimation of SD PET images from LD images [20,21].

ML-enhanced images might also exhibit different textures and noise granularity, whereas lower than regular PET images. This might add to confound raters that are more familiar to assess classical PET images, even at low counts. To our knowledge, this work reports for the first time lesion detection task outcomes from ML-enhanced PET images in the context of LD PET-CT lung screening. Our results showed that the increased image quality obtained with the IQT approach translated in a substantial reduction of false-positive across all count levels. This leads to a much higher specificity while maintaining high sensitivity, at least down to 5M count levels. The detection task using the ALD images that were enhanced with IQT resulted in the detection of all the lesions with no false positives. Overall, better agreement between readers was obtained with IQT with smaller variabilities than with SLD. Visually, the estimated images showed better body and lesion contours but appeared more pixelated probably due to "spaces" within SLD images (Fig. 6). Thus, the estimation of high-quality images using IQT may be limited to LD images above 5M true counts. Although IQT is limited by the number of true counts available for reliable estimation of SD PET images, our results showed that IQT was able to improve the image quality, in terms of lesion contrast (Fig. 3) and small lesions were detected with higher accuracy and confidence levels than SLD PET images down to 5M count levels (Fig. 7). This supports the use of IQT in improving low-quality, LD PET images, which would allow the use of injected doses somewhere between 9 MBq and 18 MBq (Table 1).

As already mentioned, reducing the dose to these ultralow levels is somehow debatable due to the small risk associated with the accompanying level of radiation. However, achieving the optimal LD image whose quality can be subsequently enriched with ML technology as we demonstrated with the ALD or even with the 10M count level images, is a more relevant application of IQT. Moreover, this developed IQT technology is applicable to many situations where PET images are reconstructed from low counting statistics scans. This is particularly the case in dynamic and/or gated PET acquisitions where total counts are split between each frame and gate. This is also the case in scanning protocols with shorter acquisition time in order to reduce motion artifacts with children or patients with conditions and may have difficulties in remaining still in the scanner. IQT may also be combined with motion correction approaches that rely on estimating the rigid motion parameters or motion fields from a first dynamic and/or gated reconstruction. Reducing the frame duration or increasing the number of gates to increase the temporal frequency of motion estimates leads to image degradation. IQT could help to increase the image quality of these short time frame images enabling either an increase in accuracy of the motion parameters and/or motion fields or an increase of the temporal frequency of the motion parameters.

4.3. Study Limitations

Presented results are based on a 10 min acquisition for lung screening using the Siemens mCT scanner. With an axial field of view of 22.1 cm, this scanner offers a sensitivity of 10 cps/kBq [22]. Acquisition times and/or injected activities should be proportionally adjusted to obtain the counting statistics presented in table 1 with scanners less sensitive. However, due to faster electronics and increased axial field of view, most of the modern scanners offer similar or higher performance with sensitivities ranging from 13.3 cps/kBq for the Siemens synchronous PET-MR scanner and the GE DMI PET/CT system [23], to more than 20 cps/kBq for the SIGNATM [24] and the DIQ PET/CT system [25] from General Electric. Higher sensitivities would potentially allow further

reduction of the acquisition time and/or of the injected dose while preserving the image quality. The reported dose reduction obtained with the ALD protocol is based on a 10 min acquisition research protocol for lung screening using the Siemens mCT scanner, which does not translate directly to standard clinical protocols. However, same scan statistics can be reached using a 3 min bed acquisition as in standard whole-body PET imaging, but with a corresponding injected dose 3.33 times higher than our proposed injected LD in order to compensate for the shorter acquisition time. This would lead to an effective PET dose corresponding to 40% of the effective dose obtained with a SD PET. When combined with the dose induced by the CT, the overall effective dose of a LD 3 min bed acquisition would be 43% of that of a SD. Based on the results obtained with the 7.5 M counts scans, IQT technology would allow further reductions leading to an overall effective dose of 34% of that of a standard whole-body PET scan. However, background tissue activity and inhomogeneity are other determinant factors impacting our capacity to detect lesions, and the translation of the lung lesion detection performance presented here is hardly translatable to whole-body screening.

Significant difference was obtained for SUV_{MEAN} only when all lesions were included, despite a small difference in SUV_{MEAN} and SUV_{MAX} between SD and ALD of -7.39% and -5.59%. However, no significance was observed after excluding 3 lesions in subjects that moved, indicating that our study may be underpowered. In addition, SUV_{MAX} from ALD was unexpectedly lower than that from SD, though the difference is expected to be very small (Fig. 5b). ALD and SD are coming from different acquisitions, leading to the inclusion of other sources of variability, which might explain this surprising finding (and contrary to SLD and SD). Motion in ALD scans and other sources of variabilities may have contributed to the need for larger samples. Images were provided to raters using 2D views in 3 orthogonal directions for visual assessment to shorten the assessment time, and to ensure consistency in assessment. The visual assessment might have been more accurate using a 3D visualization program allowing human observers to scroll through the slices. However, this is time-consuming and the need to load many images may lead to errors in classification. We made every attempt to standardize the observer task by selecting lesions that can be isolated within cubic VOIs, and small in volume (not exceeding 25% of the cubic volume) to avoid confusion of reading multiple lesions per image and have an experienced medical physicist to verify the selected lesions before we started the task. Notwithstanding this, there is always the possibility for bias in a study such as this as all the lesions were identified by only one nuclear medicine physician once, and different physicians may not agree with identified lesions and even within themselves.

5. Conclusions

Only one small lesion out of 19 was missed by a single rater out of 3 using the proposed LD PET-CT protocol, which allows a reduction of the exposure to patients to 43% of the standard protocol. In addition, although lesion contrast was reduced by about half in LD PET, there was no significant difference in lesion volume, length, $SUV_{MEAN,}$ and SUV_{MAX}. These results obtained validated the use of our proposed LD PET-CT protocol for frequent lung cancer screening. RF algorithm with a cubic patch size of 5 yielded the lowest *NMedSE* and *NMSE* and the estimated images appeared better during the manual visual assessment. IQT-estimated images yielded higher specificity and lesion contrast than SLD images, indicating an improvement in visual image quality and accuracy in lesion detection. Estimation of high-quality images using IQT can be achieved using SLD images with about 5-7.5M true counts without any loss in lesion detection, potentially enabling further reduction in radiation

exposure to 34% of SD PET-CT protocol. However, enriching the image content of LD PET images obtained with our current protocol is a more relevant application of IQT.

Journal policies detailed in this guide have been reviewed.

Acknowledgment

Funding: This study was funded by the National University Cancer Institute, Singapore Centre Grant Seed Funding Program.

Declarations of interest: none

Ethical approval: This study was approved by the Domain Specific Review Board (DSRB) of the National University Hospital Singapore. The study was conducted following the 1964 Helsinki declaration.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Research data for this article:

The raw data would remain confidential and would not be shared.

Authorships

YHN is responsible for data simulation, data analysis, and drafting of the manuscript. **DF** supported the data analysis and testing of the machine learning program. JS developed the low-dose simulation and other related programs. **SOD**, **YHN,** and **AR** carried out the lesion detection task. **SOD** also verified the locations of the lesions. **DWT**, **JJT**, **IT**, and **MC** are involved in the conception and design of the low-dose PET study. **THT** carried out clinical data support. **AK**S is the main nuclear physician responsible for identifying the lesions on the standard dose PET images. **DCA** developed the machine learning program for improving image quality. **AR** is the principal investigator of this study and is responsible for the conception and design of the study, overseeing the project development, and revising the manuscript critically. All authors are responsible for the approval of the final version of the manuscript for submission.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424. https://doi.org/10.3322/caac.21492.

[2] Volpi S, Ali JM, Tasker A, Peryt A, Aresu G, Coonar AS. The role of positron emission tomography in the diagnosis, staging and response assessment of non-small cell lung cancer. Ann Transl Med 2018;6:95–95. https://doi.org/10.21037/atm.2018.01.25.

[3] Liew CJY, Leong LCH, Teo LLS, Ong CC, Cheah FK, Tham WP, et al. A practical and adaptive approach to lung cancer screening: A review of international evidence and position on CT lung cancer screening in the Singaporean population by the College of Radiologists Singapore. Singapore Med J 2019;60:554–9. https://doi.org/10.11622/smedj.2019145.

[4] Ostrowski M, Marjański T, Rzyman W. Low-dose computed tomography screening reduces lung cancer mortality. Adv Med Sci 2018;63:230–6. https://doi.org/10.1016/j.advms.2017.12.002.

[5] Shields LBE, Wilkett Barnes JG, Buckley C, Mikos GJ, Rogers KN, Hamm JT, et al. Multidisciplinary approach to low-dose CT screening for lung cancer in a metropolitan community. Fam Pract 2020;37:25–9. https://doi.org/10.1093/fampra/cmz036.

[6] Rampinelli C, De Marco P, Origgi D, Maisonneuve P, Casiraghi M, Veronesi G, et al. Exposure to low dose computed tomography for lung cancer screening and risk of cancer: Secondary analysis of trial data and risk-benefit analysis. BMJ 2017;356:1–6. https://doi.org/10.1136/bmj.j347.

[7] Schaefferkoetter JD, Yan J, Sjöholm T, Townsend DW, Conti M, Tam JKC, et al. Quantitative accuracy and lesion detectability of low-dose 18F-FDG PET for lung cancer screening. J Nucl Med 2017;58:399–405. https://doi.org/10.2967/jnumed.116.177592.

[8] Kubo T, Ohno Y, Takenaka D, Nishino M, Gautam S, Sugimura K, et al. Standard-dose vs. low-dose CT protocols in the evaluation of localized lung lesions: Capability for lesion characterization-iLEAD study. Eur J Radiol Open 2016;3:67–73. https://doi.org/10.1016/j.ejro.2016.03.002.

[9] Schaefferkoetter J, Nai Y-H, Reilhac A, Townsend DW, Eriksson L, Conti M. Low dose positron emission tomography emulation from decimated high statistics: A clinical validation study. Med Phys 2019. https://doi.org/10.1002/mp.13517.

[10] Burgess AE, Colborne B. Visual signal detection IV Observer inconsistency. J Opt Soc Am A 1988;5:617. https://doi.org/10.1364/josaa.5.000617.

[11] Lu W, Onofrey JA, Lu Y, Shi L, Ma T, Liu Y, et al. An investigation of quantitative accuracy for deep learning based denoising in oncological PET. Phys Med Biol 2019;64. https://doi.org/10.1088/1361-6560/ab3242.

[12] Kaplan S, Zhu YM. Full-Dose PET Image Estimation from Low-Dose PET Image Using Deep Learning: a Pilot Study. J Digit Imaging 2019;32:773–8. https://doi.org/10.1007/s10278-018-0150-3.

[13] Alexander DC, Zikic D, Ghosh A, Tanno R, Wottschel V, Zhang J, et al. Image quality transfer and applications in diffusion MRI. Neuroimage 2017;152:283–98. https://doi.org/10.1016/j.neuroimage.2017.02.089.

[14] Sjoholm T, Schaefferkoetter J, Townsend D. Ultralow dose CT attenuation correction for lung cancer PET/CT screening - a phantom evaluation study. J Nucl Med 2016;57:308–308.

[15] Yan J, Schaefferkoette J, Conti M, Townsend D. A method to assess image quality for Low-dose PET: Analysis of SNR, CNR, bias and image noise. Cancer Imaging 2016;16:1–12. https://doi.org/10.1186/s40644-016-0086-0.

[16] Metz CE. Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC Studies. Invest Radiol 1989;24:234–45. https://doi.org/10.1097/00004424-198903000-00012.

[17] Retico A, Fantacci ME. The potential contribution of artificial intelligence to dose reduction in diagnostic imaging of lung cancer. J Med Artif Intell 2019;2:6–6. https://doi.org/10.21037/jmai.2019.03.03.

[18] Xiang L, Qiao Y, Nie D, An L, Lin W, Wang Q, et al. Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. Neurocomputing 2017;267:406–16. https://doi.org/10.1016/j.neucom.2017.06.048.

[19] Chen KT, Gong E, de Carvalho Macruz FB, Xu J, Boumis A, Khalighi M, et al. Ultra–Low-Dose 18 F-Florbetaben Amyloid PET Imaging Using Deep Learning with Multi-Contrast MRI Inputs. Radiology 2019;290:649–56. https://doi.org/10.1148/radiol.2018180940.

[20] Schwyzer M, Ferraro DA, Muehlematter UJ, Curioni-Fontecedro A, Huellner MW, von Schulthess GK, et al. Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks – Initial results. Lung Cancer 2018;126:170–3. https://doi.org/10.1016/j.lungcan.2018.11.001.

[21] Blumberg SB, Tanno R, Kokkinos I AD. Deeper Image Quality Transfer: Training Low-Memory Neural Networks for 3D Images 2018:118–25. Tanno, R., Ghosh, A., Grussu, F., Kaden, E., Criminisi, A., & Alexander, D. C. *Bayesian Image Quality Transfer*. 2016;(pp. 265–273). https://doi.org/10.1007/978-3-319-46723-8_31

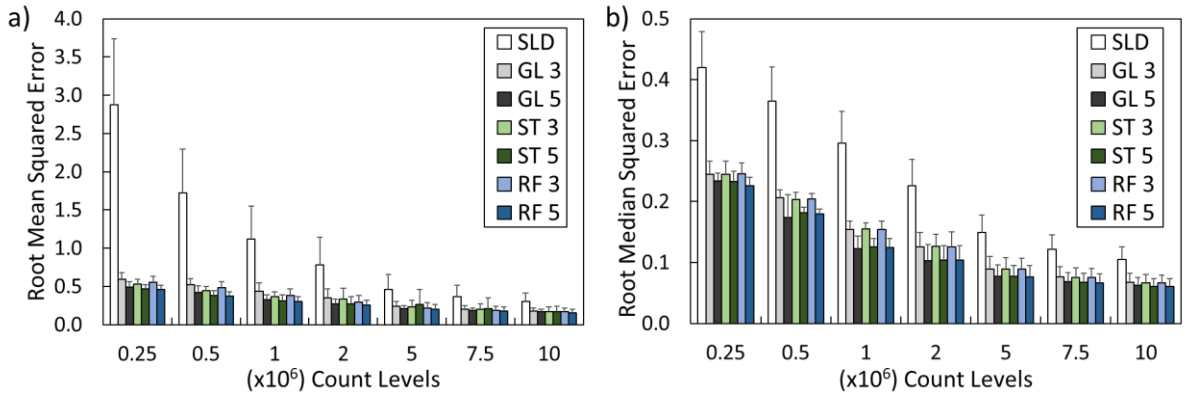
[22] Karlberg AM, Sæther O, Eikenes L, Goa PE. Quantitative comparison of PET performance—siemens biograph mCT and mMR. EJNMMI Phys 2016;3. https://doi.org/10.1186/s40658-016-0142-7.

[23] Chicheportiche A, Marciano R, Orevi M. Comparison of NEMA characterizations for Discovery MI and Discovery MI-DR TOF PET/CT systems at different sites and with other commercial PET/CT systems. EJNMMI Phys 2020;7:1–20. https://doi.org/10.1186/s40658-020-0271-x.

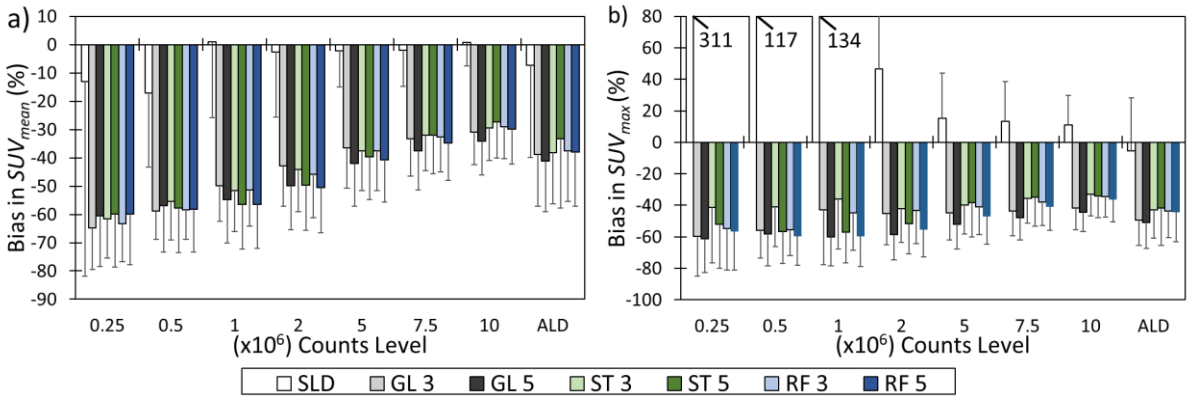
[24] Demir M, Toklu T, Abuqbeitah M, Çetin H, Sezgin HS, Yeyin N, et al. Evaluation of PET Scanner Performance in PET/MR and PET/CT Systems: NEMA Tests. Malecular Imaging Radionucl Ther 2018;27:10–8. https://doi.org/10.4274/mirt.97659.

[25] Jha A, Mithun S, Puranik A, Purandare N, Shah S, Agrawal A, et al. Performance characteristic evaluation of a bismuth germanate-based high-sensitivity 5-ring discovery image quality positron emission tomography/computed tomography system as per National Electrical Manufacturers Association NU 2-2012. World J Nucl Med 2019;18:351. https://doi.org/10.4103/wjnm.wjnm_72_18.

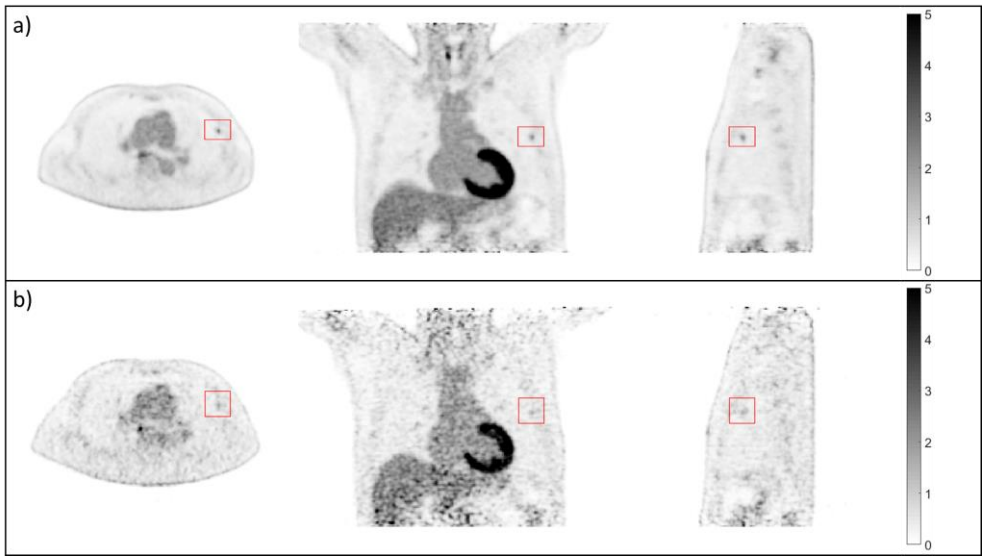
Supplementary Materials



Supplementary Fig. 1: (A) Average root-mean-square error (*RMSE*) and (B) average root-median-square-error (*RMedSE*)of the SLD images (white), and the IQT-estimated images, using global linear (GL, gray), non-linear single tree (ST, green) and random forest (RF, blue) regression algorithms with patch sizes of 3 (light-shade) and 5 (dark-shade) with respect to the SD images of 10 evaluation subjects within the masked regions. Each evaluation subject has 3 SLD and 3 corresponding IQT-estimated images.



Supplementary Fig. 2: Average bias in (a) SUV_{MEAN} and (b) SUV_{MAX} between the SLD images (white), and the IQT-estimated images, using global linear (GL, gray), non-linear single tree (ST, green) and random forest (RF, blue) regression algorithms with patch sizes of 3 (light-shade) and 5 (dark-shade) with respect to the SD images of 10 evaluation subjects with a total of 19 lesions. Each evaluation subject has 3 SLD and 3 corresponding IQT-estimated images per algorithm, but one ALD, and only one IQT-estimated image per algorithm. Bias in *SUV* is determined as $\frac{\frac{SUV_{SLD}}{IQT} - SUV_{SD}}{SUV_{SD}} \times 100\%$.



Supplementary Fig. 3: Orthogonal views of (a) SD and (b) ALD images of a subject, with a small lesion in the cubic volume, highlighted in red, as displayed in the lesion detection task.