Artificial Intelligence (AI)-Based Segmentations vs Manually Adjusted AI-Based Segmentations as a Pre-Processing Step in Whole-Body PET Dosimetry Calculations

Pleun de Vries

S3936694

Nuclear Medicine and Molecular Imaging UMCG

Period: 17/04/2023 - 01/07/2023

Bachelor's Project

1st Examiner: prof. dr. ir. Charalampos Tsoumpas

Quantification in Molecular Diagnostics & Radionuclide Therapy

2nd Examiner: dr. Joyce van Sluis

Nuclear Medicine and Medical Oncology

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
CD8+ T-cell	Cytotoxic T Cell
CNN	Convolutional Neural Network
CSV	Comma Separated Values
CT	Computed Tomography
DFO	Deferoxamine B
EARL	European Association of Nuclear Medicine Research Ltd.
FDG	Fluorodeoxyglucose
ICC	Intraclass Correlation Coefficient
ICRP	International Commission on Radiological Protection
mAb	Monoclonal Antibody
NIfTI	Neuroimaging Informatics Technology Initiative
OLINDA/EXM	Organ Level INternal Dose Assessment/ Exponential Modelling
PET	Positron Emission Tomography
SUV	Standardised Uptake Value
VOI	Volume of Interest

Abstract

Positron emission tomography (PET) with radiolabelled monoclonal antibodies, known as immunoPET, is a promising method for non-invasive tumour detection. The evaluation of newly developed immunoPET tracers, like [89]ZED88082A, is necessary to assess radiation doses and ensure safety. Currently, organ segmentation is carried out manually, which is not only time-consuming but also subject to variability. The current study aims to compare AI-based segmentations to manually adjusted AI-based segmentations, aiming to reduce analysis time and variability in dosimetry assessments. Organ volumes and estimated effective doses were obtained using the inhouse developed Biodistribution tool, Residence Time Calculator and OLINDA/EXM. The impact of segmented organ volumes on estimated effective doses was evaluated, showing a mean absolute percentage error of 2.51% for organ volumes and 0.80% for estimated effective doses, indicating slight variations between the two methods. The analysis indicated that smaller organs exhibited relatively higher errors, which was supported by the average Jaccard indexes calculated per organ. Notably, although the Wilcoxon signed rank test revealed significant differences in brain volumes, these differences were inconsistent with the mean absolute percentage error and the average Jaccard index of the brain. These findings highlight the challenges and potential of AI-based segmentation methods in clinical settings, where accurate organ segmentation is crucial for dose estimation and radiation protection. While the findings support the reliability and efficiency of the AI-tool, ongoing research and improvements are essential to optimize its performance and expand its practical applications in clinical settings.

Introduction

ImmunoPET

Positron emission tomography (PET) with radiolabelled monoclonal antibodies, also called "immunoPET", is a promising method for non-invasive tumour detection. This technique combines the high sensitivity of PET with the high antigen specificity of monoclonal antibodies (mAbs) (Zhang et al., 2011). CD8+ T-cells, which are part of the adaptive immune system, serve as highly potent effectors in the immune response against cancer and form the foundation of cancer immunotherapy (Raskov et al., 2020). The understanding of the biodistribution patterns and variability of CD8+ T-cell levels among different tumour types is limited (Kist de Ruijter et al., 2022). To tackle this problem, Kist de Ruijter et al., (2022) developed a new PET tracer, containing the [89]Zr isotope linked to CD8+-specific one-armed antibody [89]ZED88082A.

PET imaging often involves radioisotopes with short half-lives, such as ^{18}F with a half-life of only 110 minutes. These isotopes offer the advantage of minimal radiation exposure. However, they are not the most suitable choice for long-circulating probes like mAbs (Yoon et al., 2020). ImmunoPET imaging requires specific PET isotopes that exhibit excellent *in vivo* stability when combined with a mAb. Additionally, the decay half-life of the PET isotope needs to align with the pharmacokinetics of the mAb (Zhang et al., 2011). Using [89]Zr for immunoPET imaging has a few advantages over using other positron emitters (Yoon et al., 2020). Firstly, the decay half-life of [89]Zr is 3.3 days, i.e., 78.41 h, which aligns with the duration required to attain optimal tumour-to-background ratios for intact mAbs. Secondly, [89]Zr has a relatively short positron range by emitting low-energy β+ rays ($E_{β+,ave}$ = 396 KeV), enabling high-resolution PET imaging (Yoon et al., 2020; Zhang et al., 2011). Additionally, when using [89]Zr as a metallo-radionuclide, it remains stably bound to the mAb using deferoxamine B (DFO) as a chelator (Yoon et al., 2020). Due to the relatively long decay half-life of [89]Zr, it has the disadvantage of high radiation exposure. To keep radiation exposure within acceptable limits, the European guidelines allow a maximum of 37 MBq of [89]Zr administration.

The use of antibodies labelled with [89]Zr enables non-invasive whole-body visualisation through PET imaging, as highlighted by Kist de Ruijter et al., 2022. Using this approach, it becomes possible to assess whether the antibody, bound to the specific receptor, effectively targets the desired tissues, ideally the tumour. The degree of PET tracer uptake is influenced by the expression of the receptor within the tumour, which can be measured using the standardised uptake value (SUV); the higher the SUV, the higher the contrast intensity, and the higher the degree of PET tracer uptake. While tissue collection for pathology testing remains the preferred method in oncology, non-invasive evaluation, including post-therapy scans, can be employed over time (Van De Donk et al., 2022). By monitoring changes in the SUV over time, response to therapy can be assessed, e.g., a decrease in SUV can suggest a potential reduction in tumour aggressiveness or size unless further differentiation has occurred.

To determine the most effective protein dose of [89]ZED88082A, Kist de Ruijter et al. (2022) conducted a dose-finding study. Patients participating in the study received a tracer injection comprising 37 MBq (1.2-1.5 mg) of [89]ZED88082A, along with additional unlabelled CED88004S, until a protein dose of either 4 mg or 10 mg was achieved (Kist de Ruijter et al., 2022). During the initial hour after injection, the spleen exhibited significant uptake in the case of the 4 mg dose of [89]ZED88082A, likely due to its high perfusion. However, to achieve

saturation in the spleen and enable uptake in other organs, a higher dose of the unlabelled CED88004S antibody was required. Consequently, the study concluded that a protein dose of 10 mg, comprising a combination of ^{89}ZED88082A and unlabelled DFO-conjugated one-armed antibody CED88004S, yielded optimal outcomes by reducing and stabilising uptake in the spleen. This finding was attributed to the higher concentration of unlabelled CED88004S present in the 10 mg dose (Kist de Ruijter et al., 2022). The administration of a protein dose of 10 mg successfully visualised tumour lesions and lymphoid tissues, with the highest uptake observed on days 2 and 4 (Kist de Ruijter et al., 2022).

Biodistribution Assessment and Radiation Dose Estimates

Ensuring safety and efficacy in the development of new PET tracers requires an accurate assessment of their biodistribution and reliable estimates of radiation doses (van Sluis et al., 2023). While the average dose for the entire organ with uniform organ uptake is suitable for diagnostic purposes, a more comprehensive dosimetric approach is routinely employed for new drug applications and clinical evaluations involving therapeutic radiopharmaceuticals (Stabin et al., 2012). This approach allows for a more precise evaluation of the distribution of radiation doses and their implications for treatment or evaluation. This analysis involves the use of PET/CT imaging, where the computed tomography (CT) component serves two important purposes. Firstly, it enables accurate anatomical localisation of regions identified on the PET tracer uptake images. Secondly, it is utilised for attenuation correction of the PET emission data, which is a crucial step for achieving quantitative PET imaging (Griffeth, 2017). By acquiring several PET/CT scans at different time points after injection of the PET tracer, the absorbed dose, i.e., deposited radiation energy, in various organs and tissues can be determined (Kaushik et al., 2015).

Currently, the accurate assessment of biodistribution and dose estimates for newly developed PET tracers requires manual segmentation. Manual segmentation is performed on PET/CT images to delineate organs accurately. Tissue time activity curves are generated to assess the amount of radioactivity that is present in the organs of interest at each time point after administration of the radiotracer. The area under the curve for each organ of interest represents the total amount of radioactivity that has accumulated in the organ over time. Next, the total number of disintegrations over a longer period, i.e., spontaneous decay and emission of radiation by the radioisotope is estimated (Hindorf et al., 2010). The tissue time activity curves are therefore extrapolated to infinity (van Sluis et al., 2023). By extrapolating the tissue time activity to infinity, it is assumed that all of the radioisotope in the body has decayed and is no longer emitting radiation. This allows researchers and medical professionals to estimate the total number of disintegrations, that will occur over the entire period that the radioisotope is present in the body. The result of this calculation is the cumulated activity or residence time, which represents the total number of radioactive disintegrations that would occur in a tissue or organ per unit of administered activity. The mean absorbed dose of the organs of interest can be calculated next, using Organ Level INternal Dose Assessment/ Exponential Modelling (OLINDA/EXM) software, which converts the cumulated activity into dose using dose conversion factors (Stabin, 2023; van Sluis et al., 2023). The effective dose is calculated considering tissue weight factors as described by the International Commission on Radiological Protection (ICRP) Publication 103 (Clarke et al., 2007).

Challenges with Manual Segmentation for Whole-Body Dosimetry

Manual segmentation performed for radiation dosimetry analysis has several drawbacks. Firstly, it is an extremely time-consuming process. Secondly, manual segmentations have high rates of inter- and intra-observer variability (Covert et al., 2022). A faster, (semi-) automated whole-organ segmentation method for dosimetry purposes would therefore be beneficial and has been emphasised and explored before (Makris et al., 2016; Makris et al., 2014; Schmidt et al., 2016). In a previous pilot study by van Sluis et al. (2023) an online readily available fully automated artificial intelligence (AI)-based whole-organ segmentation tool was tested as a pre-processing step for calculating organ and whole-body absorbed doses, aiming to reduce analysis time with approximately 4 hours per scan. The effect of differences in manual segmentations and AI-based segmentations on the calculated whole-body effective dose estimates was evaluated and showed promising results (van Sluis et al., 2023). Hereafter, Boellaard et al. (2022) developed their own AI method consisting of three 2D U-net-shaped convolutional neural networks (CNNs), and one CNN for each slice orientation (axial, coronal, sagittal). The CNNs were trained using low-dose CT images from 30 fluorodeoxyglucose (FDG) PET/CT studies as input images with manually delineated organs (brain, lungs, liver, spleen, kidneys, bladder). Results from each trained CNN were combined using a majority vote approach to obtain the final organ segmentation (Boellaard et al., 2022).

The Objective of the Current Study

The current study aims to compare whole-organ segmentations on low-dose CT obtained using the AI-based method described by Boellaard et al. (2022) to manually adjusted (improved/finetuned) whole-organ segmentations. The study investigates the impact of variations in segmented organ volumes on the estimated effective doses by comparing AI-based segmentations with manually adjusted AI-based segmentations. Comparisons between the two segmentation methods are made to evaluate the performance of the AI-based network developed by Boellaard et al. (2022).

Methodology

Data Collection

PET/CT images obtained by Kist de Ruijter et al. (2022) were used for this study. A total of 9 patients, including 4 females and 5 males, were involved in this study. They received an intravenous administration of 37 MBq ^{89}ZED88082A. PET scans were acquired with low-dose CT for correcting attenuation and localising anatomy. The PET/CT cameras used were the Biograph mCT 40-slice, Biograph mCT 64-slice (Rausch et al., 2015) or Biograph Vision 128-slice (Van Sluis et al., 2019), all from Siemens, with software versions VG70B/VG70C/VG60C/CG70C/VG76A/VG80A. The PET scan acquisition method varied depending on the camera model: a total body mode (from skull to feet) with up to 15 bed positions for Biograph mCTs, or a total of four passes for Vision (Kist de Ruijter et al., 2022). PET/CT scans were performed at 1 hour, 2 days, 4 days, and 7 days after the PET tracer injection. All patients provided written informed consent for the imaging procedure. The studies conducted by Kist de Ruijter et al. (2022) obtained approval from both the Medical Ethical Committee of the University Medical Center Groningen and the Central Committee on Research Involving Human Subjects (Kist de Ruijter et al., 2022).

Image Pre-Processing

The PET reconstruction process followed harmonisation procedures that were in line with the EARL1 PET/CT accreditation and European Association of Nuclear Medicine guidelines (Makris, Boellaard, et al., 2014). Visual evaluation of the PET images was conducted using Syngo.via software, version VB_40.02. The Accurate tool, specifically versions .08072019, .22042020, and .14082020, was utilised for the analysis of the PET images (Boellaard, 2018). For the Biograph mCT system, EARL1 reconstruction settings were 3D TOF OP-OSEM with 3 iterations 21 subsets, resolution modelling and a Gaussian filter of 6.5 mm, for a resulting image matrix of 256x256 with a voxel size of 3.2 x 3.2 x 2 mm. In the case of Vision reconstruction, EARL1 was performed using OP-OSEM 3D-iterative algorithm with 4 iterations 5 subsets, applying TOF, into an image matrix size of 220 x 220, resulting in a voxel size of 3.3 x 3.3 x 1.6 mm. A 7 mm Gaussian filter was used (Van Sluis et al., 2019).

Manually Adjusted AI-Based Segmentation and AI-Based Segmentation

The PET/CT data described above served as input for the AI-based segmentation method developed by Boellaard et al. (2022). Manual adjustments to the segmented VOIs were made using the inhouse developed Biodistribution tool (MVOIWBDOSIMETRY_v19122021). The organs of interest, including the brain, lungs, kidneys, spleen, liver, and bladder, were analysed in this study. Specifically, the right and left lung and kidney were segmented separately. Segmentations of the fat tissue, muscle tissue (quadriceps) and red bone marrow tissue were manually added in order to obtain whole-body dosimetry. Manual adjustments to VOIs were performed by the author PV with 8 weeks of experience in image segmentation. These manual adjustments were assessed, if necessary, by the second examiner JS.

Before saving the organ activities of the VOIs after injection, the generic volumes were converted to OLINDA format. This conversion step is essential in order to use OLINDA/EXM at the following step. As a result of this conversion, the right and left lung volumes, as well as the right and left kidney volumes, are combined and treated as the body's overall lungs and kidneys. The organ activities of the VOIs after injection (day 0, 2, 4 and 7) were then saved

and used as input files for the inhouse developed Residence Time Calculator (ResidenceTimeCalculatorV02072020). This was repeated for all 9 patients. The following parameters were defined in Residence Time Calculator:

- Isotope: ^{89}Zr
- Organ: Spleen
- Data type: Activity Conc
- Volume: Reference
- Decay correction: OFF
- Interpolation: Ll (t0=t1)
- Extrapolation: Radioactivity Decay
- Gender: M/F
- Scan duration: 45 min
- Scale factor organ masses: 1.00

The residence times were then calculated and saved as CSV files. Next, the information obtained by the Residence Time Calculator was used as input for OLINDA/EXM to perform dosimetry calculations (Stabin, 2023). First, the ^{89}Zr isotope was selected in Nuclide Input Form in OLINDA/EXM. Second, the gender was selected (Adult Male/Adult Female) in the Model Input Form. Then, the residence times of the organs of interest obtained from Residence Time Calculator were transferred to the Kinetic Input Form. The effective dose per organ and the whole-body effective dose (mSv/MBq) were calculated, considering the tissue weight factors as outlined in the ICRP publication 103 (Clarke et al., 2007).

After completing these steps for the manually adjusted VOIs, the same steps were repeated for the non-adjusted VOIs. It should be noted that the CNNs utilised in the study did not directly segment fat, muscle, and red bone marrow tissue. As a result, the output file generated by the Biodistribution tool, which provides organ activities, contained 0 values for these specific tissues. However, for a more comprehensive whole-body dosimetry calculation, information pertaining to the tissue activity of fat, muscle, and red bone marrow was added through manual segmentation. To ensure consistency in volumetric measurements of these tissues between the two segmentation methods, the values obtained from the manually adjusted segmentations were transferred to the AI-based segmentations. With this approach, no differences in volume were observed for the aforementioned tissues, thereby focusing solely on identifying differences in the effective dose within the VOIs, i.e., brain, lungs, liver, spleen, kidneys and bladder.

Data Analysis

In order to assess the effectiveness of the AI-based network developed by Boellaard et al. (2022), several comparisons were made between the two segmentation methods. Firstly, the Biodistribution tool provided organ activity measurements per scan, which included volumes (in mL) of the organs of interest. These volumes served as the basis for comparing the manually adjusted AI-based segmentations with the AI-based segmentations. To accomplish this, the mean absolute percentage errors per organ and within patients were calculated using eq. (1).

$$Mean\ absolute\ \%\ error = \frac{|\,V_1 - V_2\,|}{V_1} \times 100$$

(1)

Where V_1 = volume (in mL) obtained by the manually adjusted AI-based segmentation, V_2 = volume (in mL) obtained by the AI-based segmentation.

Next, the radiation dose estimates in terms of the estimated whole-body effective dose (mSv/MBq), calculated using OLINDA/EXM, were compared to assess the differences in estimated effective dose between the two segmentation methods. This was done by calculating the mean percentage error as described in eq. (2).

$$Mean\ absolute\ \%\ error = \frac{|\,ED_1 - ED_2|}{ED_1} \times 100$$

(2)

Where ED_1 = estimated whole-body effective dose obtained by the manually adjusted AI-based segmentation, ED_2 = estimated whole-body effective dose obtained by the AI-based segmentation.

Additionally, the segmentations per organ, obtained by the Biodistribution tool in NIfTI format, were employed to evaluate the degree of spatial overlap between the segmentations generated by the two methods. The Jaccard index was used for this and is defined in eq. (3) (Shi et al., 2014).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

(3)

The Jaccard index is a measure of similarity between two sets. $|A \cap B|$ represents the intersection of set A and B. The intersection is the set of elements that are common to both sets. $|A \cup B|$ represents the union of set A and set B (*Jaccard Similarity – LearnDataSci*, n.d.). The union is the set of all elements present in either Set A or Set B, or both. It quantifies the overlap between the elements of two sets and provides a value between 0 and 1, where 0 represents no similarity and 1 represents complete similarity. The segmentations (NIfTI files) of the organs of interest obtained from one segmentation method can be denoted as Set A, while the segmentations (NIfTI files) acquired from the other segmentation method can be represented as Set B.

For statistical analysis, both the Wilcoxon signed-rank test and the Intraclass correlation coefficient (ICC) test were utilised. The Wilcoxon signed-rank test, a non-parametric statistical test, was used to determine if there was a statistically significant difference between the manually adjusted AI-based segmented volumes and the AI-based segmented volumes. Given the small sample size in the current study, it is recommended by standard statistical guidelines to employ nonparametric methods for analysing the collected data (Dwivedi et al., 2017). It assessed the ranks of the differences between the paired data. The null hypothesis assumes that the median difference between the paired observations is zero. The alternative hypothesis states

that the median difference is non-zero. A *p*-value less than 0.05 was considered statistically significant. The ICC is a statistical measure used to assess the agreement or consistency between multiple methods used for measuring a variable. It is a commonly employed measure of reliability in analyses of test-retest, intrarater and interrater reliability (Bobak et al., 2018). The two-way mixed effects model was applied, and the absolute agreement type was chosen to measure the extent of perfect consistency. This test was performed for the mean volumes of the brain, lungs, kidneys, liver, spleen, and bladder per patient. Meaning that the volumes over 4 scan days were averaged for each organ. The resulting value ranges between 0 and 1, with values closer to 1 representing stronger reliability (Koo & Li, 2016). The formula for calculating ICC is defined in eq. (4) (Bobak et al., 2018).

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

(4)

Where σ_b^2 = the variance between encounters, σ_b^2 = the variance of the raters within encounters (Bobak et al., 2018).

Results

All 9 patients included in this study received an intravenous injection of 37 MBq [89]ZED88082A. The mAb protein dose for the patients was 10 mg as described by Kist de Ruijter et al. (2022). The scan duration was equal to 45 minutes. Approximately 40 minutes per scan were required to manually adjust the AI-based segmentations.

Figures 1 and 2 provide an illustrative example of the manual adjustment process applied to the AI-generated VOIs developed by Boellaard et al. (2022). The liver of patient 4 is showcased in these figures, demonstrating the whole-organ segmentation achieved using the Biodistribution tool. In Figure 1, the CT (left column) and PET (right column) images of patient 4 are depicted. The AI-generated segmented VOI created by Boellaard et al. (2022), represented in red, corresponds to the liver. Figure 2 depicts the same patient, including the CT (left column) and PET (right column) images. The AI-generated segmented VOI in Figure 2, highlighted in red to represent the liver, has undergone manual adjustments. An important observation made from these liver segmentations is that the AI-tool (Figure 1) has mistakenly included an extra volume above the liver. Since the intended VOI is meant to specifically only include the liver, the additional lung volume is subsequently removed, as can be observed in Figure 2.
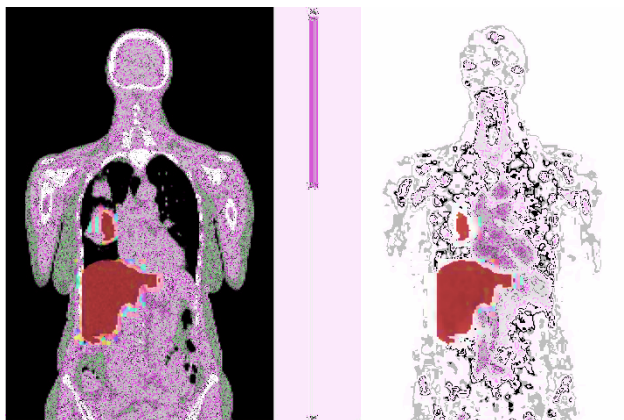

Figure 1 Low-dose CT and PET example patient 4 images in coronal view. The AI-based volume segmentation represented in red, corresponds to the liver.
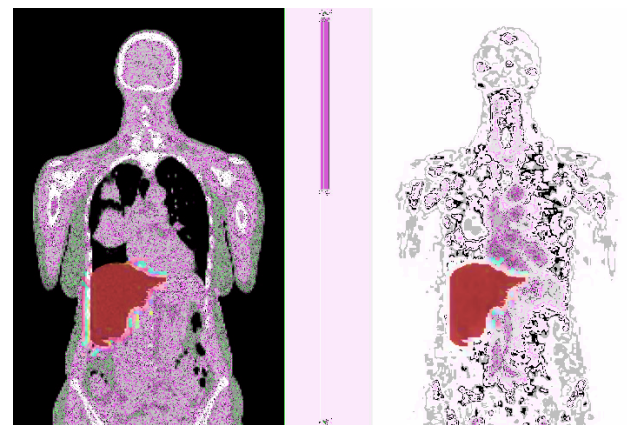
Figure 2 Low-dose CT and PET example patient 4 images in coronal view. The manually adjusted AI-based volume segmentation represented in red, corresponds to the liver.

Obtained Organ Volumes

In Figure 3, scatter plots visually present a comparison of organ volumes obtained from all patients using the two different segmentation methods. As can be observed, the data points lie along the line of identity, indicating a close correspondence between the organ volumes determined by the two segmentation methods.

Figure 3 Scatter plot of segmented whole organ volumes obtained using manually adjusted AI-based segmentation (x-axis) and AI-based segmentation (y-axis). Both axes are on a logarithmic scale.
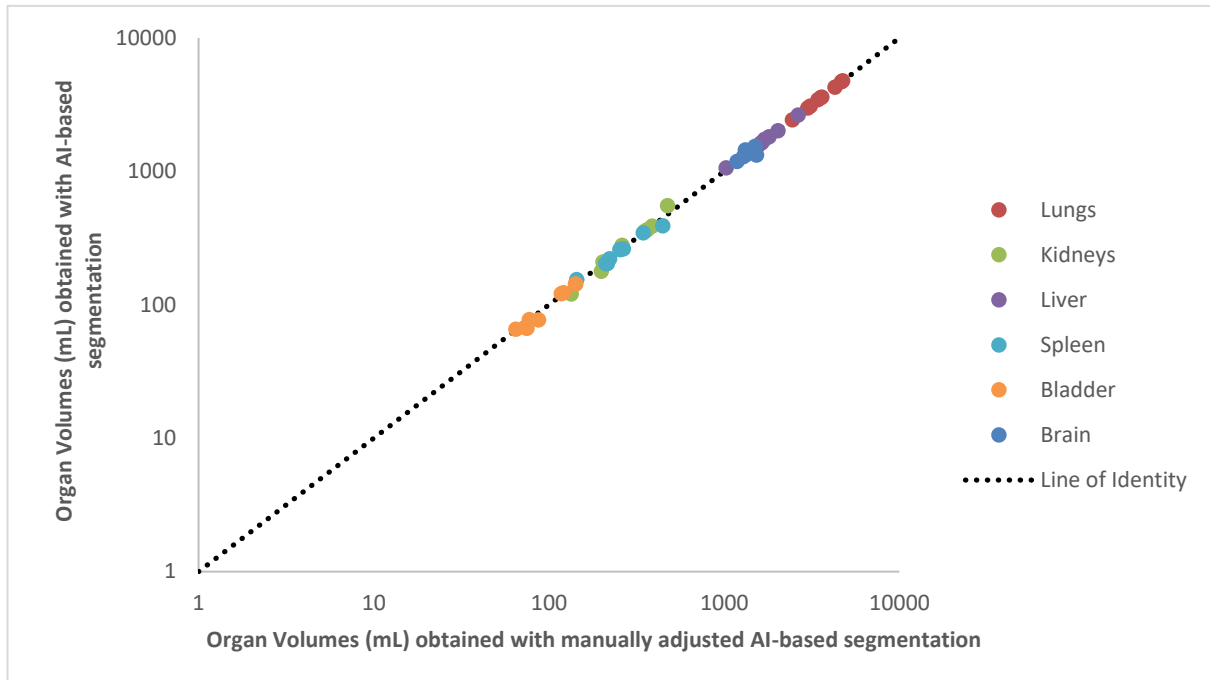
Table 1 presents the mean absolute percentage error of the volumes for each organ, aiming to provide a more detailed examination of the variations in volumes per organ. The average volumes per organ across all nine patients exhibited variations within a mean absolute percentage error of 2.45%, ranging from 0.04% to 5.71%. The total lung volume exhibited the smallest difference, while the total kidney volume showed the greatest difference.

Table 1 The mean absolute percentage error of organ volumes, averaged across all patients for each organ.

VOI	Mean Absolute Percentage Error (%)
Brain	0.10
Lungs	0.04
Kidneys	5.71
Liver	0.84
Spleen	3.81
Bladder	4.18

In Table 2, the mean absolute percentage error between organ volumes per patient is presented for both segmentation methods. The calculation involved determining the mean absolute percentage error for each organ within each patient and then computing the average of these values for each patient. The overall average difference in organ volume across all patients was found to be 2.51%, ranging from 0.19% to 5.20%. Patient 8 exhibited the smallest percentage difference, whereas patient 2 displayed the largest difference in organ volumes.

Table 2 Mean absolute percentage error of organ volumes per patient.

Patient	Error (%)
1	4.16
2	5.20
3	4.41
4	3.81
5	1.47
6	2.15
7	0.62
8	0.19
9	0.58

Table 3 presents the *p*-values obtained from the Wilcoxon signed rank test. These *p*-values evaluate the volumetric differences between the two segmentation methods for each organ. Statistically significant differences were observed in the segmented volumes of the brain (*P*=0.011) when comparing the manually adjusted AI-based segmentation and AI-based segmentation. Since the *p*-value is below the significance level of *p*=0.05, the null hypothesis is rejected. The bladder showed the highest *p*-value (*P*=0.859), suggesting that there was a minimal statistically significant difference between the segmented volumes obtained from the two segmentation methods. The null hypothesis is retained for *p*-values above 0.05.

Table 3 The *p*-values resulting from the Wilcoxon signed rank test, which assesses the volumetric differences per organ (significance level of 0.05).

VOI	*p*-value	Null Hypothesis
Brain	0.011	rejected
Lungs	0.139	retained
Kidneys	0.401	retained
Liver	0.484	retained
Spleen	0.484	retained
Bladder	0.859	retained

Since differences in volume alone do not provide information about the spatial overlap between the segmentations, the Jaccard indexes for each patient, scan and organ are calculated and illustrated in Appendix 1.

Obtained Effective Doses

Figure 4 presents a scatter plot that showcases a comparison between the calculated effective dose (in mSv/MBq) per organ, obtained through both manual and AI-based segmentation methods. As depicted in Figure 4, the data points align with the line of identity, indicating a close correspondence between the effective doses determined by the two segmentation methods.
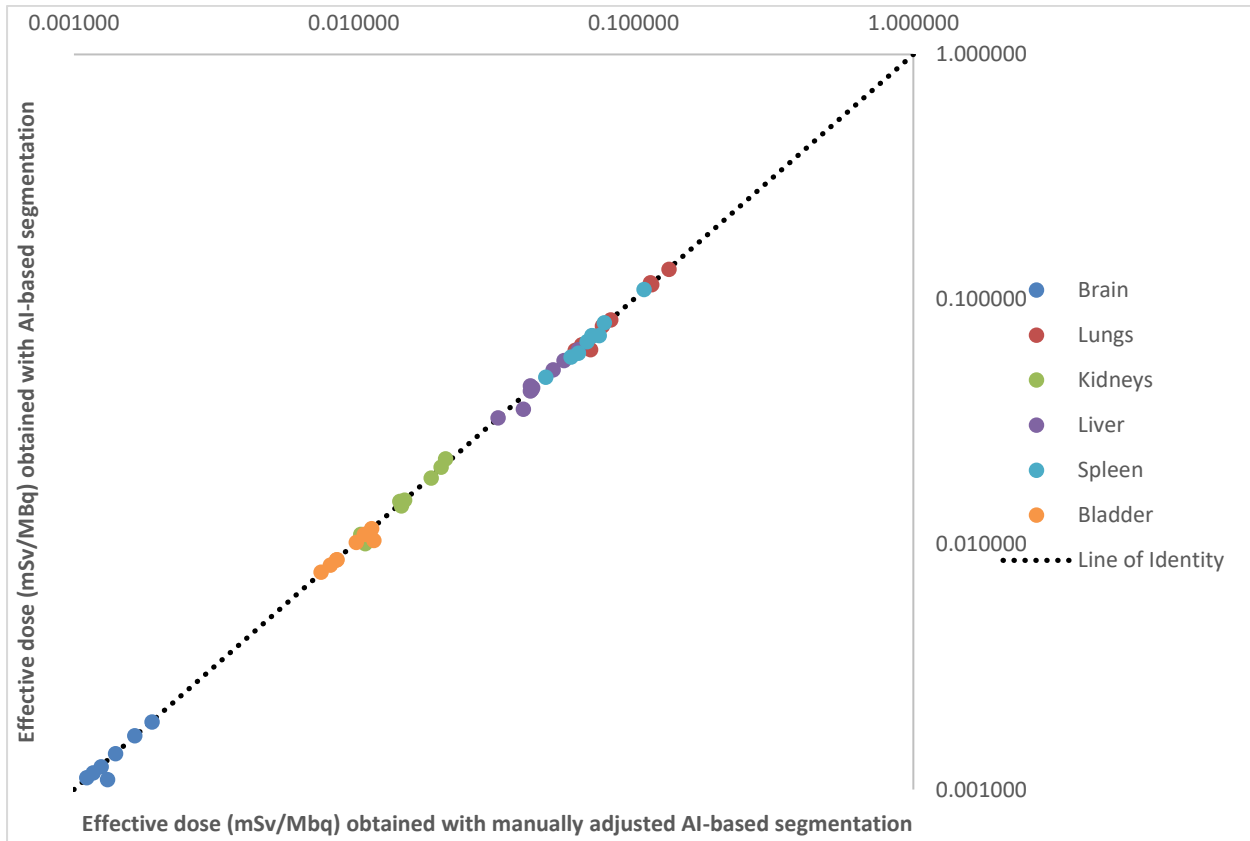
Figure 4 Scatter plot of calculated effective dose per organ obtained using manually adjusted AI-based segmentation (x-axis) and AI-based segmentation (y-axis). Axes are on a logarithmic scale.

Whole-body effective doses are shown in Table 4, aiming to provide a more detailed examination of the variations in whole-body effective doses per patient. Calculated whole-body effective doses differed minimally for patients 1-9 with a mean absolute percentage error of 0.80% (range 0.00%-1.62%). Patient 2 showed the largest mean absolute percentage error (1.62%). Patient 4 showed the smallest mean absolute percentage error (0.00%).

Table 4 Calculated whole-body effective doses (mSv/MBq) considering the tissue weight factors as outlined in the ICRP.

Patient	Effective dose (mSv/MBq): AI-based	Effective dose (mSv/MBq): Manually adjusted AI-based	Error (%)
1	3.97E-01	3.95E-01	0.45
2	6.18E-01	6.28E-01	1.62
3	4.55E-01	4.59E-01	0.88
4	4.14E-01	4.14E-01	0.00
5	6.60E-01	6.53E-01	1.06
6	4.56E-01	4.52E-01	0.88
7	5.19E-01	5.24E-01	0.96
8	4.15E-01	4.17E-01	0.48
9	4.70E-01	4.66E-01	0.85

ICC test

Table 5 presents the results of the ICC test, aiming to provide insight into the reliability between the two segmentation methods for the observed organ volume. The highest coefficient is found for the brain (1.00) and the lowest coefficient is found for the liver and bladder (0.93).

Table 5 ICC scores for the VOIs between the two segmentation methods.

VOI	ICC Average
Brain	1,00
Lungs	0,99
Kidneys	0,99
Liver	0,93
Spleen	0,98
Bladder	0,93

Discussion

The aim of the current study is to compare whole-organ segmentations on low-dose CT scans using an AI-based method described by Boellaard et al. (2022) to manually adjusted (improved/finetuned) AI-based whole-organ segmentations. The primary focus of this investigation is to assess the impact of variations in segmented organ volumes on the estimated effective doses by comparing the AI-based segmentations with the manually adjusted AI-based segmentations.

In recent years, AI-based methods have shown great potential in automating and improving medical image analysis, including organ segmentation. The study by van Sluis et al. (2023) showed promising results in obtaining whole-organ segmentation by a readily available AI-based segmentation tool. The method developed by Boellaard et al. (2022) enables automated segmentation of organs on low-dose CT. It is important to note that, similar to manual segmentation, AI-based segmentations may exhibit variability influenced by factors such as the quality and quantity of the manual segmentation training data employed during network training. Nonetheless, it is crucial to thoroughly evaluate and validate the performance of such AI-based networks by comparing them to manual adjustments, which are typically executed by experts with extensive anatomical knowledge. By comparing the AI-based segmentations with manually adjusted AI-based segmentations, this study aims to provide insights into the performance and reliability of the AI-based network developed by Boellaard et al. (2022).

Figures 1 and 2 display the liver segmentation of patient 4 using both segmentation methods. This particular example provides insight into the possibility of the AI-tool incorrectly segmenting certain parts of the organ. The ensuing discussion points will shed light on the overall performance of the AI-tool across all nine patients.

Upon analysing the organ volumes and estimated effective doses acquired from the two segmentation methods in Figures 3 and 4, it becomes evident that all data points align closely with the line of identity. This suggests minimal differences between the segmentation methods regarding the volume and estimated effective dose. However, upon reviewing the mean absolute percentage errors presented in Tables 1,2, and 4, it becomes evident that there are differences in both the organ volume and estimated effective dose between the two segmentation methods.

Firstly, the average volumes per organ across all nine patients exhibited variations within a mean absolute percentage error of 2.45%, ranging from 0.04% (for the lungs) to 5.71% (for the kidneys). The reason behind the kidneys having a relatively high percentage error and the lungs having a low percentage error can be attributed to the findings of Trägårdh et al. (2020), which revealed that smaller organs present greater difficulties in terms of automated segmentation. Specifically, differentiating voxels at the edges of the organ and distinguishing neighbouring organs becomes particularly difficult. This difficulty is amplified in cases where organs have low contrast compared to the surrounding tissue, as observed in the dataset utilised in this study, specifically in relation to the kidneys. Consequently, it was anticipated that organs with larger and more distinctive characteristics, such as the lungs, would exhibit a lower mean absolute percentage error and thereby a higher Jaccard index, i.e., more overlap. Analysing the Jaccard indexes for the organs in Appendix I reveals that the lungs displayed a high average Jaccard index (0.99), while the kidneys showed the lowest average Jaccard index (0.86). These findings

appear to align with the conclusions drawn by Trägårdh et al. (2020) and hold relevance in the context of the present study.

Nonetheless, upon reviewing the *p*-values resulting from the Wilcoxon signed rank test in Table 3, a significant difference in brain volume was found (*P*=0.011), representing the lowest p-value observed. On the other hand, the bladder displayed the highest *p*-value (*P*=0.859). These findings indicate that there were significant differences in brain volumes and the least significant differences in bladder volumes. Surprisingly, upon analysing the average Jaccard indexes of the brain and bladder in Table 6 (obtained from Appendix I), the situation is reversed. A higher Jaccard index, i.e., a higher degree of overlap, was found for the brain than for the bladder, indicating that the degree of overlap for the brain volume was higher than for the bladder volume. Table 1 further supports this observation, as it reveals a mean absolute percentage difference of 0.10% for the brain and 4.18% for the bladder.

Table 6 *P*-values and calculated Jaccard indexes of the brain and bladder, retrieved from Appendix I.

	Brain	**Bladder**
***P*-value**	0.011	0.859
Average Jaccard Index	1.00	0.90

The statistically significant difference in brain volume found by the Wilcoxon signed rank test may be caused by the fact that the Wilcoxon signed rank test takes positive and negative differences into account. As the AI-tool underestimates the volume of the brain 8 out of 9 times, the Wilcoxon signed rank test assigns this as statistically significantly different. Hence, the results of the Wilcoxon signed rank test do not provide reliable information regarding the absolute differences between the two segmentation methods. Therefore a statistically significantly different does not necessarily mean the most absolute volumetric difference. Another factor contributing to this statistically significant difference might be that the average volume of the bladder is smaller than the average volume of the brain. The bigger average volume of the brain results in a higher likelihood of spatial overlap. However, since the brain has a larger volume, there may be more significant differences between manually adjusted AI-based and AI-based volumes. Additionally, due to the larger numbers involved, these differences can become statistically significant more quickly.

Upon comparing Tables 2 and 4, it becomes clear that a lower mean absolute percentage error among the 9 patients was found for the estimated effective dose than for the organ volumes. The average mean absolute percentage error among the 9 patients' organ volumes (Table 2) was found to be 2.51%. Additionally, the average mean absolute percentage error among the 9 patients' estimated effective dose was found to be 0.80% (Table 4). The findings reveal that there were more variations in organ volumes compared to the effective dose, despite the fact that organ volumes were required to calculate the effective dose. The difference between these mean absolute percentage errors is 1.71%. This error can be attributed to the fact that whole-body dosimetry involves combining weighted dose values from various organs and structures. Subsequently, it has been observed, consistent with the findings of van Sluis et al. (2023), that a slight difference in organ volume has minimal influence on the effective dose.

Consequently, the calculation of the effective dose is less influenced by the variabilities stemming from the selection of segmentation methods. This clarifies why patient 8 in Table 2 exhibits the lowest mean absolute percentage error for organ volumes (0.19%), while patient 4 in Table 4 demonstrates the lowest mean absolute percentage error for estimated effective dose (0.00%). The findings of this study indicate that the estimation of the effective dose remains accurate (within 0.80%), regardless of little variations in manually adjusted or AI-based organ segmentation. The tracer used in this study, ^{89}ZED88082A, contains a one-armed antibody. Due to the smaller size of this one-armed antibody, it exhibits a faster clearance rate through the kidneys. As a result, there is a possibility of a lower absorbed dose throughout the whole body compared to what is typically reported in the literature (Kist de Ruijter et al., 2022).

The ICC outcomes shown in Table 5 indicate that the two segmentation methods exhibit good reliability for the analysed organs, with the brain demonstrating the highest level of agreement. All ICC values are above 0.90 and an ICC score above 0.90 indicates excellent reliability (Koo & Li, 2016). This signifies strong agreement and reliability between the two segmentation methods, indicating that the volumes per organ are highly consistent and reproducible.

One limitation of the present study includes the low-dose CT dataset comprising various PET/CT cameras, namely the Biograph mCT 64-slice, Biograph mCT 40-slice, and Biograph Vision 128-slice. The dataset exhibited variations in image quality, specifically in terms of noise, as depicted in Figures 5 and 6. Noise in CT is influenced by tube current (mAs), slice thickness, patient size and reconstruction algorithm (Bell & Murphy, 2017) The difference in image quality in Figures 5 and 6 may be attributed to the utilisation of different PET/CT camera types, with different slice thickness. Another factor that may influence the amount of noise is patient size, as patient 5 (93.0 kg) in Figure 5 is larger than patient 9 (82.8 kg) in Figure 6 in terms of kilograms. The absorption of radiation is greater in larger patients, resulting in a decrease in the number of photons that reach the detector. Consequently, this leads to a reduction in the signal-to-noise ratio. (Bell & Murphy, 2017).

Figure 5 Patient 5 low-dose CT scan.

Figure 6 Patient 9 low-dose CT scan.

As a result, the depiction of organs in Figure 5 for patient 5 appears less distinct compared to the organs depicted in Figure 6 for patient 9. This limitation could hinder the AI tool's ability to accurately segment VOIs. Surprisingly, the Jaccard indexes presented in Table 7 for patient 5 are relatively high. The limited visibility of organs in Figure 5, resulting from low contrast, constrained the extent of adjustments that could be made to enhance the segmentation made by CNNs. As a result, the Jaccard indexes for the organs of patient 5 are high. Given the uncertainty surrounding the accuracy of the segmentations, it is important to exercise caution and not automatically assume the reliability of the findings for patient 5.

Table 7 Jaccard Indexes (retrieved from Appendix I) of patient 5 & 9. Values <0.90 appear red.

Patient 5				Patient 9			
0	2	4	7	0	2	4	7
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99
0.91	0.91	0.96	0.99	0.93	0.82	0.72	0.81
0.99	0.99	1.00	0.97	1.00	1.00	1.00	1.00
0.95	0.95	0.93	1.00	0.99	0.98	1.00	0.94
1.00	1.00	0.98	1.00	0.82	0.95	1.00	0.74

The current study has an additional limitation wherein the manual adjustments to the volumes of interest (VOIs) were conducted by a rater, PV, with only 8 weeks of experience in image segmentation and limited expertise in anatomy. Not all of the manual adjustments were evaluated by the second examiner, JS. As a result, the utilisation of low-dose CT for image segmentation introduces uncertainty concerning the accuracy of the manually segmented VOIs.

Conclusion

The primary aim of the current study was to investigate the impact of variations in segmented organ volumes on estimated effective doses by comparing AI-based segmentations with manually adjusted AI-based segmentations. The findings demonstrate that the AI tool performs effectively, achieving an accuracy of 0.80% for the estimated effective dose compared to manually adjusted AI-based segmentation. This indicates the potential of the AI-based tool as a reliable tool in clinical settings. Additionally, the utilisation of the AI tool significantly reduces analysis time compared to manual segmentation, as supported by the findings of van Sluis et al. (2022). This time-saving advantage further highlights the value and practicality of the AI-based approach.

However, it is important to acknowledge the limitations identified in the study. Further research and refinement are needed to address these limitations and enhance the applicability of the AI-based tool. Future investigations should focus on addressing factors such as variations in image quality and the training data used for network development. Furthermore, occasional differences were observed in both the volume and location of organs derived from low-dose CT and PET images, particularly in the case of the spleen. These differences might have arisen due to the utilisation of the CT image obtained before the emission data collection (PET image) for attenuation correction. However, due to the scanning process of the PET scanner, there is typically a time interval of 20 or 30 minutes between the attenuation scan (CT scan) and the final emission scan (PET scan). During this timeframe, significant motion can occur, creating substantial opportunities for the occurrence of motion artefacts (Griffeth, 2017). One specific and crucial aspect to consider is the effect of respiratory motion. Since the patient continues to breathe during the PET scan, it is essential to conduct the CT scan in a way that closely aligns with the positioning of the diaphragm and nearby organs, like the spleen. This approach aims to optimize the registration of the two data sets for accurate alignment. To improve on this limitation, future research needs to investigate the potential impact of PET image-based segmentations on the estimation of effective dose, particularly for organs with high uptake such as the spleen.

In summary, this study contributes to the understanding of the potential benefits and limitations of utilising AI-based methods for organ segmentation and dose estimation. While the findings support the reliability and efficiency of the AI-tool, ongoing research and improvements are essential to optimize its performance and expand its practical applications in clinical settings.

References

Bell, D., & Murphy, A. (2017). Noise (CT). *Radiopaedia.Org*. https://doi.org/10.53347/RID-51832

Bobak, C. A., Barr, P. J., & O'Malley, A. J. (2018). Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Medical Research Methodology*, *18*(1). https://doi.org/10.1186/S12874-018-0550-6

Boellaard, R. (2018). Quantitative oncology molecular analysis suite: ACCURATE. *Journal of Nuclear Medicine*, *59*(supplement 1).

Boellaard, R., de Vries, B., van Sluis, J., Pieplenbosch, S., Wiegers, S., de Hollander, M., Zijlstra, J., & Zwezerijnen, B. (2022). SEMI-AUTOMATED AI BASED ORGAN DELINEATION ON LOW DOSE CT TO FACILITATE PET RADIOTRACER BIODISTRIBUTION MEASUREMENTS. *Physica Medica*, *104*, S126. https://doi.org/10.1016/S1120-1797(22)02415-2

Clarke González Y Sasaki RM Alexakhin L-E Holm C Streffer JD Boice jr FA Mettler jr A Sugier, R. A., Holm J-K Lee Sasaki JD Boice jr H Menzel, L. Y., & Pentreath Sugier AJ González RJ Preston, R. A. (2007). The 2007 Recommendations of the International Commission on Radiological Protection. ICRP publication 103. *Annals of the ICRP*, *37*(2–4), 1–332. https://doi.org/10.1016/J.ICRP.2007.10.003/ASSET/J.ICRP.2007.10.003.FP.PNG_V03

Covert, E. C., Fitzpatrick, K., Mikell, J., Kaza, R. K., Millet, J. D., Barkmeier, D., Gemmete, J., Christensen, J., Schipper, M. J., & Dewaraja, Y. K. (2022). Intra- and inter-operator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry. *EJNMMI Physics*, *9*(1), 1–16. https://doi.org/10.1186/S40658-022-00515-6/FIGURES/5

Dwivedi, A. K., Mallawaarachchi, I., & Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine*, *36*(14), 2187–2205. https://doi.org/10.1002/SIM.7263

Griffeth, L. K. (2017). Use of Pet/Ct Scanning in Cancer Patients: Technical and Practical Considerations. *Https://Doi.Org/10.1080/08998280.2005.11928089*, *18*(4), 321–330. https://doi.org/10.1080/08998280.2005.11928089

Hindorf, C., Glatting, G., Chiesa, C., Lindén, O., & Flux, G. (2010). EANM dosimetry committee guidelines for bone marrow and whole-body dosimetry. *European Journal of Nuclear Medicine and Molecular Imaging*, *37*(6), 1238–1250. https://doi.org/10.1007/S00259-010-1422-4/TABLES/4

Jaccard Similarity – LearnDataSci. (n.d.). Retrieved June 21, 2023, from https://www.learndatasci.com/glossary/jaccard-similarity/

Kaushik, A., Jaimini, A., Tripathi, M., D'Souza, M., Sharma, R., Mondal, A., Mishra, A. K., & Dwarakanath, B. S. (2015). Estimation of radiation dose to patients from 18FDG whole body PET/CT investigations using dynamic PET scan protocol. *The Indian Journal of Medical Research*, *142*(6), 721. https://doi.org/10.4103/0971-5916.174563

Kist de Ruijter, L., van de Donk, P. P., Hooiveld-Noeken, J. S., Giesen, D., Elias, S. G., Lub-de Hooge, M. N., Oosting, S. F., Jalving, M., Timens, W., Brouwers, A. H., Kwee, T. C.,

Gietema, J. A., Fehrmann, R. S. N., Fine, B. M., Sanabria Bohórquez, S. M., Yadav, M., Koeppen, H., Jing, J., Guelman, S., … de Vries, E. G. E. (2022). Whole-body CD8+ T cell visualization before and during cancer immunotherapy: a phase 1/2 trial. *Nature Medicine 2022 28:12*, *28*(12), 2601–2610. https://doi.org/10.1038/s41591-022-02084-8

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

Makris, N. E., Boellaard, R., Menke, C. W., Lammertsma, A. A., & Huisman, M. C. (2016). An automatic delineation method for bone marrow absorbed dose estimation in 89Zr PET/CT studies. *EJNMMI Physics*, *3*(1), 1–9. https://doi.org/10.1186/S40658-016-0149-0/TABLES/1

Makris, N. E., Boellaard, R., Visser, E. P., De Jong, J. R., Vanderlinden, B., Wierts, R., Van Der Veen, B. J., Greuter, H. J. N. M., Vugts, D. J., Van Dongen, G. A. M. S., Lammertsma, A. A., & Huisman, M. C. (2014). Multicenter Harmonization of 89Zr PET/CT Performance. *Journal of Nuclear Medicine*, *55*(2), 264–267. https://doi.org/10.2967/JNUMED.113.130112

Makris, N. E., Van Velden, F. H. P., Huisman, M. C., Menke, C. W., Lammertsma, A. A., & Boellaard, R. (2014). Validation of simplified dosimetry approaches in 89 Zr-PET/CT: The use of manual versus semi-automatic delineation methods to estimate organ absorbed doses. *Medical Physics*, *41*(10), 102503. https://doi.org/10.1118/1.4895973

Raskov, H., Orhan, A., Christensen, J. P., & Gögenur, I. (2020). Cytotoxic CD8+ T cells in cancer and cancer immunotherapy. *British Journal of Cancer 2020 124:2*, *124*(2), 359–367. https://doi.org/10.1038/s41416-020-01048-4

Rausch, I., Cal-González, J., Dapra, D., Gallowitsch, H. J., Lind, P., Beyer, T., & Minear, G. (2015). Performance evaluation of the Biograph mCT Flow PET/CT system according to the NEMA NU2-2012 standard. *EJNMMI Physics*, *2*(1), 1–17. https://doi.org/10.1186/S40658-015-0132-1/TABLES/5

Schmidt, T. G., Wang, A. S., Coradi, T., Haas, B., & Star-Lack, J. (2016). Accuracy of patient-specific organ dose estimates obtained using an automated image segmentation algorithm. *Https://Doi.Org/10.1117/1.JMI.3.4.043502*, *3*(4), 043502. https://doi.org/10.1117/1.JMI.3.4.043502

Shi, R., Ngan, K. N., & Li, S. (2014). Jaccard index compensation for object segmentation evaluation. *2014 IEEE International Conference on Image Processing, ICIP 2014*, 4457–4461. https://doi.org/10.1109/ICIP.2014.7025904

Stabin, M. G. (2023). OLINDA/EXM 2-The Next-generation Personal Computer Software for Internal Dose Assessment in Nuclear Medicine. *Health Physics*, *124*(5). https://doi.org/10.1097/HP.0000000000001682

Stabin, M. G., Xu, X. G., Emmons, M. A., Segars, W. P., Shi, C., & Fernald, M. J. (2012). RADAR Reference Adult, Pediatric, and Pregnant Female Phantom Series for Internal and External Dosimetry. *Journal of Nuclear Medicine*, *53*(11), 1807–1813. https://doi.org/10.2967/JNUMED.112.106138

Trägårdh, E., Borrelli, P., Kaboteh, R., Gillberg, T., Ulén, J., Enqvist, O., & Edenbrandt, L. (2020). RECOMIA—a cloud-based platform for artificial intelligence research in nuclear medicine and radiology. *EJNMMI Physics*, *7*(1), 1–12. https://doi.org/10.1186/S40658-020-00316-9/TABLES/4

Van De Donk, P. P., Oosting, S. F., Knapen, D. G., Van Der Wekken, A. J., Brouwers, A. H., Lub-De Hooge, M. N., De Groot, D. J. A., & De Vries, E. G. E. (2022). Molecular imaging to support cancer immunotherapy. *Journal for ImmunoTherapy of Cancer*, *10*(8), e004949. https://doi.org/10.1136/JITC-2022-004949

Van Sluis, J., De Jong, J., Schaar, J., Noordzij, W., Van Snick, P., Dierckx, R., Borra, R., Willemsen, A., & Boellaard, R. (2019). Performance Characteristics of the Digital Biograph Vision PET/CT System. *Journal of Nuclear Medicine*, *60*(7), 1031–1036. https://doi.org/10.2967/JNUMED.118.215418

van Sluis, J., Noordzij, W., de Vries, E. G. E., Kok, I. C., de Groot, D. J. A., Jalving, M., Lub-de Hooge, M. N., Brouwers, A. H., & Boellaard, R. (2023). Manual Versus Artificial Intelligence-Based Segmentations as a Pre-processing Step in Whole-body PET Dosimetry Calculations. *Molecular Imaging and Biology*, *25*(2), 435–441. https://doi.org/10.1007/S11307-022-01775-5/FIGURES/4

Yoon, J. K., Park, B. N., Ryu, E. K., An, Y. S., & Lee, S. J. (2020). Current Perspectives on 89Zr-PET Imaging. *International Journal of Molecular Sciences 2020, Vol. 21, Page 4309*, *21*(12), 4309. https://doi.org/10.3390/IJMS21124309

Zhang, Y., Hong, H., & Cai, W. (2011). PET Tracers Based on Zirconium-89. *Current Radiopharmaceuticals*, *4*(2), 131. https://doi.org/10.2174/1874471011104020131

Appendix I

Scan day	Patient 1				Patient 2				Patient 3				Patient 4				Patient 5				Patient 6				Patient 7				Patient 8				Patient 9			
	0	2	4	7	0	2	4	7	0	2	4	7	0	2	4	7	0	2	4	7	0	2	4	7	0	2	4	7	0	2	4	7	0	2	4	7
Brain	0.99	1.00	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Lungs	1.00	0.99	0.99	0.99	0.95	0.99	0.99	0.99	0.99	0.99	0.98	0.99	1.00	0.98	0.98	0.99	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00
Kidneys	0.61	0.92	0.80	0.87	0.66	0.73	0.76	0.76	0.77	0.64	0.76	0.81	0.91	0.90	0.81	0.97	0.91	0.91	0.96	0.99	0.93	0.82	0.72	0.81	0.91	0.87	0.81	0.88	0.88	1.00	1.00	1.00	1.00	0.98	0.96	0.91
Liver	0.95	1.00	0.98	0.99	0.96	0.95	0.95	0.96	0.99	0.97	0.94	0.96	0.95	0.92	0.97	0.98	0.99	0.99	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
Spleen	0.77	1.00	0.92	0.99	0.79	0.91	0.96	0.79	1.00	0.99	0.97	0.97	0.95	0.90	0.87	0.97	0.95	0.95	0.93	1.00	0.99	0.98	1.00	0.94	0.99	0.98	0.99	0.99	0.99	1.00	1.00	0.90	1.00	1.00	0.97	
Bladder	0.79	0.88	0.84	0.85	0.90	0.80	0.89	0.79	0.95	0.76	0.57	0.65	0.81	1.00	0.75	0.99	1.00	1.00	0.98	1.00	0.82	0.95	1.00	0.74	1.00	0.99	1.00	0.98	0.98	0.97	0.96	0.96	0.95	0.75	1.00	1.00