Optimizing SUV Analysis: A Multicenter Study on Preclinical FDG-PET/CT Highlights the Impact of Standardization

Optimizing SUV Analysis: A Multicenter Study on Preclinical FDG-PET/CT Highlights the Impact of Standardization

Claudia Kuntner[1, 2], Carlos Alcaide[3], Dimitris Anestis[4], Jens P. Bankstahl[5], Herve Boutin[6,7], David Brasse[8], Filipe Elvas[9], Duncan Forster[10], Maritina G. Rouchota[4], Adriana Tavares[3], Mari Teuter[5], Thomas Wanek[1], Lena Zachhuber[1], Julia G. Mannheim[11, 12]

[1] Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria

[2] Medical Imaging Cluster (MIC), Medical University of Vienna, Vienna, Austria

[3] University of Edinburgh, Edinburgh, United Kingdom

[4] BIOEMTECH, Athens, Greece

[5] Hannover Medical School, Hannover, Germany

[6] Division of Neuroscience & Experimental Psychology, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

[7] INSERM, UMR 1253, iBrain, Université de Tours, Tours, France

[8] Institut Pluridisciplinaire Hubert Curien, Université de Strasbourg, CNRS, UMR7178, Strasbourg, France

[9] Molecular Imaging Center Antwerp, University of Antwerpen, Antwerpen, Belgium

[10] Division of Informatics, Imaging and Data Sciences, Manchester Molecular Imaging Centre, The University of Manchester, Manchester, UK.

[11] Werner Siemens Imaging Center, Department of Preclinical Imaging and Radiopharmacy, Eberhard-Karls University Tuebingen, Tuebingen, Germany

[12] Cluster of Excellence iFIT (EXC 2180) "Image Guided and Functionally Instructed Tumor Therapies", Tuebingen, Germany

Corresponding & first author: Claudia Kuntner

Address: Waehringer Guertel 18-20, 1090 Vienna, Austria

E-mail address: claudia.kuntner@meduniwien.ac.at

Telephone number: +43 (0)1 40160-64022

Short running title: Multicenter preclinical image analysis

Manuscript category: Original Article

Abstract

Purpose: Preclinical imaging, with translational potential, lacks a standardized method for defining volumes of interest (VOIs), impacting data reproducibility. The aim of this study was to determine the interobserver variability of VOI sizes and standard uptake values (SUV_{mean} and SUV_{max}) of different organs using the same [^{18}F]FDG-PET and PET/CT datasets analyzed by multiple observers. In addition, the effect of a standardized analysis approach was evaluated.

Procedures: In total, 12 observers (4 beginners and 8 experts) analyzed identical preclinical [^{18}F]FDG-PET-only and PET/CT datasets according to their local default image analysis protocols for multiple organs. Furthermore, a standardized protocol was defined, including detailed information on the respective VOI size and position for multiple organs, and all observers reanalyzed the PET/CT datasets following this protocol.

Results: Without standardization, significant differences in the SUV_{mean} and SUV_{max} were found among the observers. Coregistering CT images with PET images improved the comparability to a limited extent. The introduction of a standardized protocol that details the VOI size and position for multiple organs reduced interobserver variability and enhanced comparability.

Conclusions: The protocol offered clear guidelines and was particularly beneficial for beginners, resulting in improved comparability of SUV_{mean} and SUV_{max} values for various organs. The study suggested that incorporating an additional VOI template could further enhance the comparability of the findings in preclinical imaging analyses.

Key words (3-5): multicenter, image analysis, reproducibility, PET/CT, preclinical imaging

Introduction

Over the past few decades, preclinical molecular imaging, notably positron emission tomography (PET) combined with computed tomography (CT), has become indispensable in scientific medical research [1-2]. This approach offers multimodal imaging in preclinical models that are highly translatable to clinical settings [3-4]. PET enables quantification of biological processes in living subjects, achieved by defining regions or volumes of interest (ROIs or VOIs) on the images to extract activity concentrations (typically given in kBq/cc). Mathematical operations transform these activity concentrations into percent injected activity or dose per volume of tissue (%IA/cc or %ID/cc) by normalizing them to the administered activity or standardized uptake values (SUVs) by additionally normalizing to the body weight. The SUV is used as a semiquantitative measurement of glucose uptake in tissue from a 2-deoxy-2-[^{18}F]fluoro-D-glucose ([^{18}F]FDG) PET scan, especially in clinical practice [5]. The SUV_{mean}, reflecting the mean voxel value within a VOI, is strongly influenced by the VOI definition method and is susceptible to partial volume effects, resulting in greater variability. Conversely, the SUV_{max}, which represents the voxel with the highest radioactivity concentration, is less affected by observer variability but more affected by technical variations [6].

A major limitation in preclinical imaging is the lack of standardized or fully automated methods for defining VOIs. While some data-driven or semiautomatic segmentation methods exist, they still require observer input to define or choose the proposed cluster. Anatomy-based automatic segmentation methods rely heavily on annotated training images (magnetic resonance (MR) and/or CT), but their effectiveness hinges on the quality and quantity of the database. Currently, there is no widely accepted automated preclinical VOI delineation method. Consequently, most preclinical image analysis is manual, with observers selecting regions for analysis. Additionally, the availability of multiple software tools for preclinical PET/CT image analysis, each with different features and pipelines, further complicates the issue.

For clinical PET/CT imaging, several studies have assessed inter- and intraobserver variability and proposed methods to standardize image analysis [7-10]. Until now, there hasn't been any study conducted on preclinical PET/CT imaging that includes a standardized image analysis. Therefore, the present study assessed the variability in VOI size, $SUV_{mean,}$ and SUV_{max} measurements of multiple organs and tumors between different observers (grouped into beginners and experts) when analyzing the same preclinical [^{18}F]FDG-PET-only and [^{18}F]FDG-PET/CT datasets with free or commercially available image analysis software. Furthermore, a standardized protocol was used, and all observers reanalyzed the PET/CT datasets following this protocol; potential improvements in interobserver variability were evaluated accordingly.

Materials and Methods

Imaging data

Twelve observers analyzed dynamic [^{18}F]FDG-PET-only (dynamic images 0-75 min, 25 frames; n=6) and [^{18}F]FDG-PET/CT (dynamic images 0-60 min, 19 frames; n=7) scans of tumor-bearing mice. Two laboratories provided the datasets, which were acquired according to local regulations. The images were provided in Bq/cc together with the injected activities and weights of the mice in the scanner-specific and DICOM formats. Information regarding the animal experiments and imaging protocols can be found in the Electronic Supplementary Material (ESM). Co-registration of PET/CT data for part 2 and 3 was performed by one observer to eliminate potential co-registration-induced influences.

Of the twelve observers, eight were experts in the analysis of preclinical images (> 4 years of experience), whereas four were classified as beginners (< 1 year of experience). With the exception of the dataset providers, all observers analyzed the images independently and blinded to each other's assessments, utilizing their expertise and judgment.

Part 1: [^{18}F]FDG-PET-only image analysis and reporting

The observers were asked to analyze the images according to their standard institutional procedures, including the choice of image analysis software, the procedures for preparing the images (e.g., adjustment of the animal's position), the radiation scale and time frames, and the method of delineating VOIs. The observers were requested to delineate the following VOIs: tumor, whole brain, muscle, heart (either whole heart or left ventricle), kidneys (left and right), liver, and urinary bladder (short name bladder). An additional region covering the whole FOV was delineated on the last time frame with a predefined size (128 × 128 × 95 voxels/51.2 × 51.2 × 75.62 mm³) to assess any software-related biases in image quantitation.

After analyzing the images, the observers completed a detailed report, including SUV_{mean} and SUV_{max} (normalized to the body weight of the animals, respectively), VOI delineation method (manual, thresholding, fixed objects, etc.), and volume (in mm³). They also specified how they displayed the images (radiation scale, minimum and maximum values, kBq/cc, %IA/cc, or SUV). As the datasets were dynamic, observers indicated the time frame (individual frame or summed image) for VOI delineation. Time-activity curves (TACs) for all animals and organs were plotted. Group differences (SUV_{mean} and SUV_{max}) were determined across observers and animals based on the 10 min time frame from 55-65 min.

Part 2: [^{18}F]FDG-PET/CT image analysis and reporting

The image analysis procedure for the PET/CT datasets was identical to that for the [^{18}F]FDG-PET-only datasets. Only the whole FOV region was adjusted (256 × 256 × 159 voxels/99.377 × 99.377 × 126.564 mm³) as a different PET scanner was used for these experiments. In addition, the observers were asked to report on which dataset (PET or CT) each organ and the tumor were delineated. Group differences (SUV_{mean} and SUV_{max}) were determined across observers and animals based on the 5 min time frame from 55-60 min.

Part 3: Standardized [^{18}F]FDG-PET/CT image analysis and reporting

The authors established a standardized tumor and organ VOI definition method based on [^{18}F]FDG-PET-only and [^{18}F]FDG-PET/CT data analysis results. The protocol required to be universally applicable across image analysis software tools. Consequently, data-driven segmentation methods, such as multiclustering, were excluded from part 3, resulting in the exclusion of observer E8. Observer B3's analysis was also omitted due to inability to meet the standardized consensus specifications for VOI definition.

Observers unanimously opted to delineate organs and tumors using specific objects (ellipsoids and boxes), with predefined VOI drawing on either PET or CT images. PET-related VOIs adhered to a fixed radiation scale specified in SUV. VOIs for the brain, heart and tumor were delineated on the CT images (and verified on the respective PET images), as the CT image provided sufficient anatomical delineation to surrounding tissues. The VOIs for both muscle regions, kidneys, liver and both bladder regions were delineated on the PET images (and verified on the respective CT images) due to the fact that for most of these organs the [^{18}F]FDG uptake is very distinct and the low soft-tissue contrast of the CT doesn't enable a clear delineation to surrounding tissues.

Table 1 summarizes the objects and predefined VOI sizes and ranges. To explore VOI position influence on quantitative analysis, two muscle regions (gluteus maximus and biceps/triceps) and two urinary bladder regions (bottom and maximum fill) were included.

Table 1 Details on the standardized VOI analysis. The PET-related VOIs were delineated at the last time frame using the specified SUV radiation scale.

VOI	image used for VOI delineation	radiation scale (SUV)	shape	size	notes
tumor	CT	n.a.	ellipsoid	entire tumor	
brain	CT	n.a.	ellipsoid	7 x 5 x 10 mm^3	inside skull, control on PET that olfactory bulb and harderian glands are excluded
heart	CT	n.a.	ellipsoid	>100 and <200 mm^3	
muscle	PET	0 - 2	box	2 x 2 x 3 mm^3	gluteus maximus, avoid spill in from bladder, control on CT that no bone is included
muscle	PET	0 - 2	box	2 x 2 x 3 mm^3	biceps/triceps, control on CT that no bone is included
kidney	PET	0 - 2	ellipsoid	~ 60 mm^3	definition of right and left side
liver	PET	0 - 2	box	4 x 4 x 4 mm^3	opposite to the stomach

bladder bottom	PET	0 - 10	box	2 x 2 x 2 mm³	bottom of bladder
bladder maximum fill	PET	0 - 10	ellipsoid	entire bladder	draw on time frame with largest bladder fill

157

158 Statistical analysis

159 The mean or maximum radioactivity concentrations given as SUV_{mean} or SUV_{max} per animal
160 and organ over the 12 (part 1 and 2) and 10 (part 3) observers were used.

161 The coefficient of variation (CV, %) was calculated as the ratio of the standard deviation to the
162 mean to assess the extent of variability. Moreover, to account for the variability between
163 animals, the normalized difference was calculated for each animal and organ based on the 60
164 min values using the following equation:

165 $$normalized\ difference = \frac{individual\ value - mean\ value}{mean\ value}$$

166 The data are expressed as the mean ± standard deviation. Statistical analysis was performed
167 with Prism 9.5.0 Software (GraphPad, La Jolla, CA, USA) and SPSS Statistics (version 29.0,
168 IBM SPSS, IBM Corp., Armonk, NY, USA). Differences between the beginner and expert
169 groups were assessed by applying two-way ANOVA followed by a Bonferroni multiple
170 comparisons test, with an alpha level of 0.05 for each organ. Brown-Forsythe and Welch
171 ANOVA tests were performed to assess interobserver variability, followed by Dunnett's
172 multiple comparisons test, with individual variances computed for each comparison and organ.
173 The threshold of statistical significance was set to an adjusted p value ≤ 0.05.

Intraclass correlation coefficients (ICCs; single-measure, two-way random, absolute agreement) were calculated based on the SUV_{mean} and SUV_{max} values to determine interobserver reliability for the beginners, the experts, and all observers [11-12]. According to Koo et al. [12], ICCs less than 0.5 can be classified as poor reliability, ICCs in the range of 0.5 to 0.75 as moderate reliability, ICCs between 0.75 and 0.8 as good reliability, and ICCs greater than 0.9 as excellent reliability.

Results

Selection of image analysis software programs and VOI definition methods

Five different image analysis software programs were utilized in the present study. The selected software and the typically used output units, radiation scales, and time frames are summarized in the **Suppl. Tab. s1** (see ESM). One observer employed a data-driven segmentation method (observer E8, BrainVISA/Anatomist) that used the local means analysis method based exclusively on the dynamics (i.e., time-activity and level of uptake) of each voxel in the PET images [13-14]. The VOIs of six of the remaining eleven observers were defined in the last time frame. Some observers (3 out of 11) selected the time frame where the respective organ was clearly visible for analysis. Seven out of the eleven observers applied different radiation scales for specific organs (e.g., 0-2 SUV for muscle, 0-20 SUV for the heart), whereas the rest used a fixed radiation scale for all organs. The whole FOV region evaluated in parts 1 and 2 revealed no systematic software biases in image-based quantitation of the mean and maximum activity values (**Suppl. Fig. s1,** see ESM). These small differences were attributed to the VOI position in the whole FOV region.

Parts 1 and 2: Individual [^{18}F]FDG-PET-only and [^{18}F]FDG-PET/CT image analysis

VOI sizes

The VOI delineation methods vary from fixed objects (e.g., spheres for the whole brain and heart) to manual drawings of VOIs on consecutive slices to those using thresholds (see **Fig. 1** for examples of VOI positions and shape for each software tool). Some observers applied post-processing to re-orient the images according to the “standard” configuration in preclinical imaging (head first, prone), whereas others analyzed the images in the orientation provided by the scanner. The delineation methods used for each organ are summarized in the supplementary methods (**Suppl. Fig. s2 and s3,** see ESM) for the PET-only and PET/CT studies, respectively.

For the [^{18}F]FDG-PET-only study, the tumor VOI was excluded from the analysis because delineation was rather challenging due to the low uptake and small size of the tumors (most of the observers could not identify the tumors).

The different delineation methods resulted in considerable variability in the VOI sizes, as illustrated in **Fig. 2** ((**a**) [^{18}F]FDG-PET-only; (**b**) [^{18}F]FDG-PET/CT). The beginners delineated significantly larger liver and heart VOIs than did the experts on the PET images (part 1). The smallest variability in the VOI sizes in the beginner group was obtained for the heart (71% CV), whereas in the expert group, the smallest variability was obtained for the kidneys (52% CV). In contrast, the greatest variability was found in the muscle VOI (149% CV) for the beginner group and in the liver VOI (210% CV) for the expert group.

On the [^{18}F]FDG-PET-CT images (part 2), the beginners delineated significantly larger VOIs than did the experts in the liver, heart, and brain. The smallest variability in VOI sizes was obtained in the bladder for the beginners (37% CV) and in the tumor VOIs for the experts (40% CV). The highest variability in VOI sizes was found in the muscle for the beginners (159% CV) and in the liver for the experts (164% CV). In particular, the VOI drawn for the liver ranged from 16 to 3619 mm³, which spans two orders of magnitude. Furthermore, the VOI position for the muscle differed among the observers (e.g., for part 2, the lower left limb was delineated by seven observers, the upper left limb was delineated by four observers, and the upper right limb was delineated by one observer).

Organ-time activity curves

The organ TACs for part 1 [^{18}F]FDG-PET-only images for a representative animal, subdivided into beginner and expert groups, are shown in **Suppl. Fig. s4** (SUV_{mean}) and **Fig. s5** (SUV_{max}) in the ESM. The heart and kidney SUV_{mean} TACs exhibited greater interobserver variation in the beginner group than in the expert group. The remaining organs revealed a similar pattern between beginners and experts.

For the SUV_{max} of the TACs, the beginner group revealed greater interobserver variation for the brain and muscle; interestingly, the experts showed greater variability than the beginners for the liver and heart.

The inclusion of CT data (part 2) reduced the variability in the liver, brain, and muscle SUV_{mean} TACs, as depicted in **Suppl. Fig. s6** and **Fig. s7** (see ESM). For the SUV_{max} of the TACs (beginners: **Suppl. Fig. s8;** experts: **Suppl. Fig. s9**), reduced variability was detected mainly for the muscle. The two groups of observers determined identical SUV_{max} TACs for the tumor, kidney, and bladder.

Last time frame analysis

The SUV_{mean} and SUV_{max} values from the time frame covering 60 min were used to compare the variability between groups (beginners and experts) and individual observers. For the PET-only study, the calculated normalized difference based on the SUV_{mean} showed the greatest deviation from 0 for the heart region (-0.25 ± 0.27 for beginners and 0.13 ± 0.18 for experts) and the smallest deviation for the brain (0.01 ± 0.14 for beginners and -0.01 ± 0.14 for experts), as displayed in the upper row of **Fig. 3 (a)**. In addition, statistically significant differences were observed between the beginner and expert groups for the heart, muscle and bladder. The ICCs revealed greater reliability within the expert groups for all organs except the brain, although poor reliability was observed for the muscle and liver (ICCs<0.5).

The calculated normalized difference based on the SUV_{max} (**Fig. 3(b)**) yielded the greatest deviation from 0 for the muscle region among the beginners (0.24 ± 0.81) and for the bladder among the experts (0.14 ± 0.95). The smallest deviation was found for the kidney region (beginners: 0.01 ± 0.02; experts: -0.01 ± 0.07). Overall, no statistically significant differences between the observer groups were observed. An overview of all the ICCs, including confidence intervals (CIs), for each organ can be found in the supplementary materials (**Suppl. Tab. s2,** see ESM).

Multiple statistically significant differences in the SUV_{mean} were detected between the individual observers, especially for the heart and muscle VOIs, as shown in **Fig. 4(a)**. For the SUV_{max}, the liver and muscle indices revealed multiple significant differences among the 12 observers (**Fig. 4(b)**). The individual p values are given in Suppl. Fig. s10 (see ESM).

For the PET/CT study, the normalized difference of the muscle for beginners and experts was reduced (compare the middle row of **Fig. 3(a)**). However, statistically significant differences between the observer groups were obtained for the heart, kidneys, bladder, and tumor. The ICCs for the liver, muscle, and bladder showed improved reliability compared to those of part 1. Analyzing the normalized difference based on the SUV_{max} (**Fig. 3(b)**) yielded the largest overall spread in the liver region (0.60 ± 1.67 for the beginners and -0.25 ± 0.73 for the experts, $p<0.0001$). No improvement in reliability was detected for the ICCs based on the SUV_{max} for part 2 compared to part 1.

The interobserver SUV_{mean} and SUV_{max} variability are displayed in **Fig. 5(a)** and **6(a),** revealing multiple statistically significant differences in the heart and tumor regions (both SUV_{mean}) as well as the liver and brain regions (both SUV_{max}). The individual p values between the observers are given in Suppl. Fig. s11 and Fig. s12 (see ESM).

Part 3: standardized [^{18}F]FDG-PET/CT image analysis

The predefined VOI sizes reduced the variations, as shown in **Fig. 2(c)**. However, for the two regions for which the entire structure was to be delineated, namely, the tumor and the bladder at the maximum-fill level, significantly larger VOIs were determined by experts with great variability (tumor: beginners: 41% CV; experts: 38% CV; bladder: beginners: 56% CV; experts: 45% CV).

<u>Organ-time activity curves after standardization</u>

The standardized image analysis method reduced the variation in the SUV_{mean} TACs of the tumor, brain, liver, and kidney, as shown in panel B in the **Suppl. Fig. s6 and s7** (see ESM). The muscle and bladderTACs exhibited different patterns depending on the VOI position. The expert group obtained mostly congruent SUV_{max} TACs for the liver, heart, tumor, brain, kidneys, and bladder maximum-fill VOIs (**Suppl. Fig. s9**), whereas the beginner group obtained slightly greater variations (**Suppl. Fig. s8,** see ESM).

Last time frame analysis after standardization

The standardized analysis approach notably enhanced the normalized difference based on SUV_{mean} for most organs, depicted in the lower row of **Fig. 3(a)**, correlating with higher ICCs across most organs. Liver and brain index reliability significantly improved, achieving excellent levels post-standardization. Initially poor heart and tumor reliability transformed into good and moderate levels, respectively. Standardization notably elevated kidney index reliability from moderate to excellent levels. However, statistically significant differences persisted between observer groups for muscle gluteus maximus and urinary bladder maximum-fill regions. Improvement in normalized difference based on SUV_{max} was inconsistent post-standardization, with no improvement observed for tumor or urinary bladder (**Fig. 3(b)**). Significant differences between observer groups were found for liver and gluteus maximus region (SUV_{max}). Notably, liver and brain ICCs substantially improved in standardized analysis (liver: part 2=0.08, part 3=0.43; brain: part 2=0.00, part 3=0.65).

The interobserver variability based on the SUV_{mean} values was markedly reduced using the standardized image analysis approach. However, some statistically significant differences between observers persisted in the tumor, biceps/triceps muscle, or maximum-fill urinary bladder region (**Fig. 5(b)**).). The individual p values between the observers are given in Suppl. Fig. s13 (see ESM). For the SUV_{max}, no significant differences were found between the observers for any of the organs (**Fig. 6(b)**).

Discussion

Quantifying radioactivity concentrations in small animal organs or tumors is standard in preclinical imaging and relies on parameters such as the SUV_{mean} or SUV_{max}. However, the variability and reproducibility of these parameters among different observers within a single institution or across multiple centers remain poorly understood. Currently, each imaging lab and often each observer within the same institution applies different workflows, experiences, and judgments to analyze and segment PET images. These variations encompass factors such as the position, size, and shape of VOIs; PET image display settings; and postprocessing methods, potentially compromising comparability across observers and centers. Despite the prevalence of preclinical [^{18}F]FDG-PET/CT studies, no multicenter consensus exists on a reproducible image analysis method. This study represents the first comprehensive multicenter [^{18}F]FDG-PET/(CT) investigation into the impact of image analysis methods on results and the comparability of a standardized analysis approach. Our findings underscore the significant influence of image analysis methods on [^{18}F]FDG-PET/(CT) study outcomes, particularly regarding SUV_{mean} discrepancies attributed to regional position and size, corroborating similar observations from prior studies [15].

Our first observation was that not all observers performed post-processing to re-orient the images according to the “standard” configuration in preclinical imaging (head first, prone). Some analyzed the images in the orientation provided by the scanner, which was for the PET/CT study in feed first, prone. Thus, an agreement on the orientation of images to be used (also with regard to future automatic segmentation applications) is therefore the first step towards standardized image analysis. Without standardization, variations in VOI sizes were observed between beginners and experts for multiple organs. These differences influenced SUV_{mean} (e.g., heart) and SUV_{max} (e.g., liver in PET/CT) analyses, suggesting that VOI size impacts uptake. However, for certain organs (e.g., the liver in PET-only and the brain in PET/CT), despite significant differences in VOI size, SUV analysis was unaffected by homogeneous [^{18}F]FDG uptake.

Introducing anatomical references in part 2 reduced variability in heart and muscle regions but had no effect on liver or brain regions. However, overall reliability and comparability did not improve universally. Comparing parts 1 and 2 is challenging due to the different image sets analyzed. However, this design showcases variability between studies (e.g., small vs. large tumors with necrotic areas), mitigating potential biases from part 1 to part 2.

Based on the results from these two studies, the participants in this study reached a consensus on the standardized VOI delineation method utilized in part 3.

Standardization improved the consistency and shape of SUV_{mean} TACs in the liver, brain, and kidney, while nearly identical SUV_{max} TACs were obtained in the liver, heart, tumor, brain, kidneys, and urinary bladder. Reduced interobserver variability poststandardization was evidenced by reduced deviation and improved ICCs across organs, except for muscle and urinary bladder regions. Muscle VOIs are small and prone to spill over from adjacent bone regions, making muscle-fat differentiation challenging despite the use of anatomical information from CT scans. Intensive training and visual aids are recommended for comparability improvement. For maximum-fill bladder VOIs, inconsistent time frame choices hindered comparisons between parts 2 and 3. Nevertheless, considering its importance in dosimetric studies, assessing bladder necessity and employing frame-by-frame analysis for volumetric changes are advised.

Furthermore, the significant differences between beginners and experts found by the normalized difference analysis in the heart, kidneys, and tumor diminished after standardization (**Fig. 3(b)** and **3(c)**). We concluded that the use of a standardized approach reduced the interobserver variability in the SUV analysis. In addition, we propose to create a VOI template for each preclinical PET/CT and PET/MR study that includes a standardized VOI positioning and size as well as detailed information on the segmentation method. For multicenter studies, we recommend reaching a consensus on the use of single analysis software for evaluating and providing VOI template files. For single-center studies, a VOI

template from the first animal analyzed will ensure reproducibility for the remaining animals and help train new personnel.

In general, the SUV_{max} revealed a lower interobserver variability than the SUV_{mean} in our study. However, as the SUV_{max} represents only a single voxel within a region, the SUV_{mean} might be a more stable marker for underlying tissue uptake. Therefore, both measures can be valuable in multicenter studies.

Despite its strengths, our study has several limitations. First, mid-level observers were not included, potentially biasing the results, as experience levels were subjectively categorized as beginners or experts. Additionally, the varied backgrounds of the participating observers (e.g., physics, chemistry, biology, etc.) may have influenced interpretation. Secondly, validation using gamma-counter data was not available. Third, the use of different image analysis software led to the use of various segmentation tools, hindering detailed discrepancy identification within segmented VOIs. Finally, the standardized protocol lacked optimization, notably omitting a VOI template for precise location visualization. Addressing these limitations in future studies could enhance the accuracy and reproducibility of the findings.

It has to be noted that depending on the specific tracer used, standardized image analysis protocols need to be re-defined to address tracer-specific factors that might impact the reproducibility of image analysis. This also applies for the acquisition of the imaging data, for which standardized protocols – depending on the used tracer – can also significantly enhance reproducibility [16].

The 12 observers in this study represent 8 different preclinical imaging facilities in Europe and all observers were asked to use their default image analysis method and software tool to analyze the provided PET(/CT) data. Only 1 observer analyzed the data using an automated segmentation tool. Automatic organ segmentation has been an active field of research for decades [17-22], and current research in this field includes the development of artificial intelligence (AI)-assisted solutions [23]. Nevertheless, manual delineation will still be the

standard method for image analysis until these tools are applicable to a broader community with sufficient training databases and a variety of VOI templates. The variety of chosen software tools and methods utilized in this study represents in our opinion the current standard in preclinical imaging. However, the transition to AI-guided automatic segmentation will certainly be a strong focus within the next decade and thus will potentially improve the comparability and reliability of preclinical multicenter image analysis.

Conclusion

For the first time, the present study demonstrated the significant influence of image analysis on the obtained quantitative data; this work is intended as the basis for a discussion of further standardization approaches in preclinical imaging. Moreover, the authors aim to raise awareness of potential pitfalls when preclinical data are analyzed by multiple observers with different levels of experience. Our study verified that the comparability of image analysis significantly improves when detailed standardized image analysis protocols are used. This approach will be of particular interest not only for preclinical multicenter studies but also for studies performed over a long period within the same institution, where the observers might vary.

Acknowledgment

For this work, the methodological advice of the Institute of Clinical Epidemiology and Applied Biometry of the University of Tübingen was applied. We would like to express our sincere thanks to Mr. Blumenstock for his support.

Author contributions

CK and JGM designed the study. CK, HB and DF provided the data. CK, CA, DA, JB, HB, BD, FE, DF, MT, TW, LZ and JGM analyzed the image data. AT and MGR interpreted the data. CK and JGM performed the comparability analysis of all observer analyses. All the authors were involved in critically revising the manuscript. All the authors have read and approved the final version of the manuscript.

Funding

This work was supported by the COST Action "Correlated Multimodal Imaging in Life Sciences" (COMULIS, CA17121) and by the Chan Zuckerberg Initiative, Advancing Imaging through Collaborative Projects (COMULISglobe, 2023-321161).

Conflict of interest

Author DA and MGR are employees of the company BIOEMTECH.

Ethical approval

All applicable institutional and/or national guidelines for the care and use of animals were followed.

References

1. Lewis JS, Achilefu S, Garbow JR, Laforest R, Welch MJ (2002) Small animal imaging. current technology and perspectives for oncological imaging. Eur J Cancer 38:2173-2188.
2. Kiessling F, Pichler BJ, Hauff P (2017) Small Animal Imaging.
3. Cherry SR, Gambhir SS (2001) Use of positron emission tomography in animal research. ILAR J 42:219-232.

4. Phelps ME (2004) PET: Molecular Imaging and Its Biological Applications. 1st ed. Softcover of orig. ed. 2004 ed., Berlin: Springer.
5. Kinahan PE, Fletcher JW (2010) Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. Semin Ultrasound CT MR 31:496-505.
6. De Luca GMR, Habraken JBA (2022) Method to determine the statistical technical variability of SUV metrics. EJNMMI Phys 9:40.
7. Suzuki A, Nakamoto Y, Terauchi T, et al. (2007) Inter-observer Variations in FDG-PET Interpretation for Cancer Screening. Japanese Journal of Clinical Oncology 37:615-622.
8. Büyükdereli G, Güler M, Şeydaoğlu G (2016) Interobserver and Intraobserver Variability among Measurements of FDG PET/CT Parameters in Pulmonary Tumors. Balkan Med J 33:308-315.
9. Gerke O, Vilstrup MH, Segtnan EA, Halekoh U, Høilund-Carlsen PF (2016) How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardisation. BMC Med Imaging 16:54.
10. Guezennec C, Bourhis D, Orlhac F, et al. (2019) Inter-observer and segmentation method variability of textural analysis in pre-therapeutic FDG PET/CT in head and neck cancer. PLoS One 14:e0214299.
11. Fisher RA (1992) Statistical Methods for Research Workers. In Breakthroughs in Statistics: Methodology and Distribution, Eds. Kotz S, Johnson NL. New York, NY: Springer New York, pp 66-70.
12. Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15:155-163.
13. Maroy R, Boisgard R, Comtat C, et al. (2008) Segmentation of rodent whole-body dynamic PET images: an unsupervised method based on voxel dynamics. IEEE Trans Med Imaging 27:342-354.
14. Maroy R, Boisgard R, Comtat C, et al. (2010) Quantitative organ time activity curve extraction from rodent PET images without anatomical prior. Med Phys 37:1507-1517.
15. Habte F, Budhiraja S, Keren S, Doyle TC, Levin CS, Paik DS (2013) In situ study of the impact of inter- and intra-reader variability on region of interest (ROI) analysis in preclinical molecular imaging. Am J Nucl Med Mol Imaging 3:175-181.
16. Mannheim JG, Mamach M, Reder S, et al. (2019) Reproducibility and Comparability of Preclinical PET Imaging Data: A Multicenter Small-Animal PET Study. J Nucl Med 60:1483-1491.
17. Baiker M, Milles J, Dijkstra J, et al. (2010) Atlas-based whole-body segmentation of mice from low-contrast Micro-CT data. Med Image Anal 14:723-737.
18. Khmelinskii A, Baiker M, Kaijzel EL, Chen J, Reiber JH, Lelieveldt BP (2011) Articulated whole-body atlases for small animal image analysis: construction and applications. Mol Imaging Biol 13:898-910.
19. Wang H, Stout DB, Chatziioannou AF (2012) Estimation of mouse organ locations through registration of a statistical mouse atlas with micro-CT images. IEEE Trans Med Imaging 31:88-102.
20. Akselrod-Ballin A, Dafni H, Addadi Y, et al. (2016) Multimodal Correlative Preclinical Whole Body Imaging and Segmentation. Sci Rep 6:27940.
21. Yan D, Zhang Z, Luo Q, Yang X (2017) A Novel Mouse Segmentation Method Based on Dynamic Contrast Enhanced Micro-CT Images. PLoS One 12:e0169424.
22. Wang H, Han Y, Chen Z, Hu R, Chatziioannou AF, Zhang B (2019) Prediction of major torso organs in low-contrast micro-CT images of mice using a two-stage deeply supervised fully convolutional network. Phys Med Biol 64:245014.
23. Schoppe O, Pan C, Coronel J, et al. (2020) Deep learning-enabled multi-organ segmentation in whole-body mouse scans. Nature Communications 11:5626.

Figures

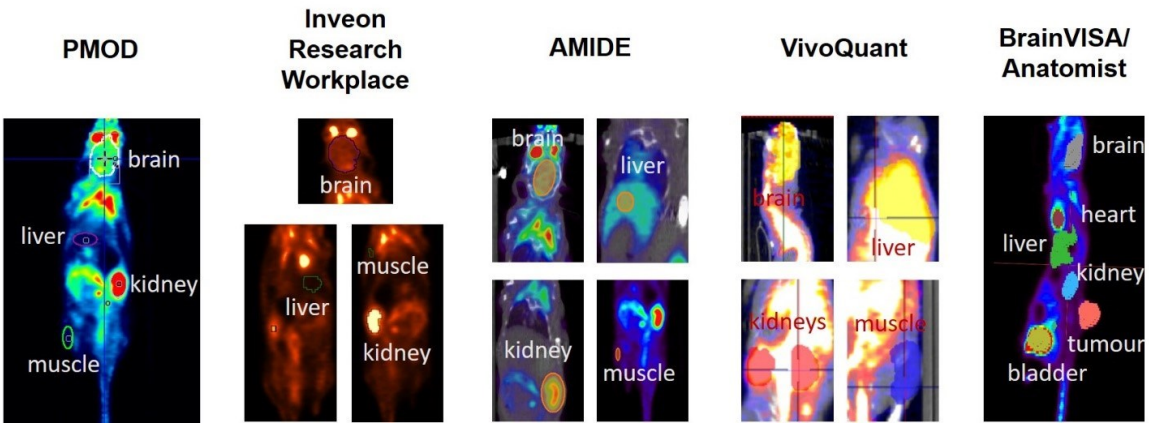


Fig. 1 Representative images of multiple VOI positions for the individual software tools utilized for analysis. With the BrainVISA software, a 3D rendering of the VOIs is displayed.

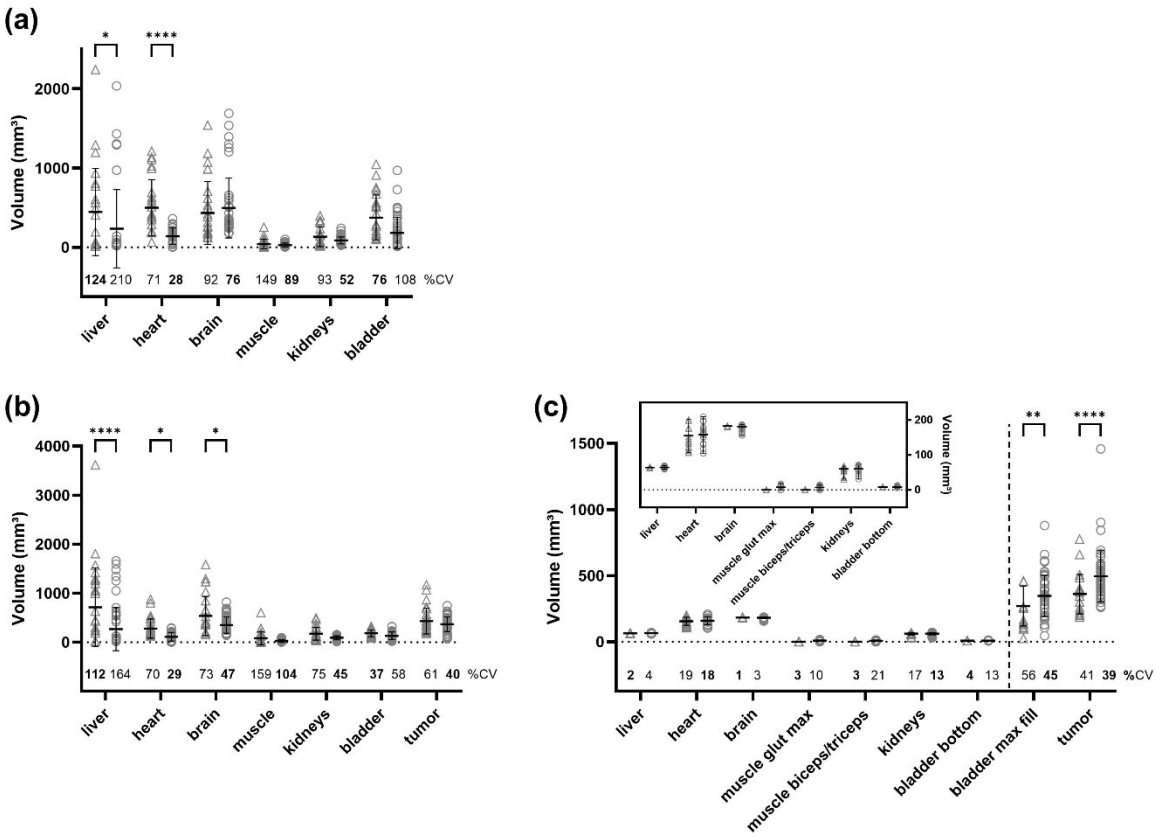
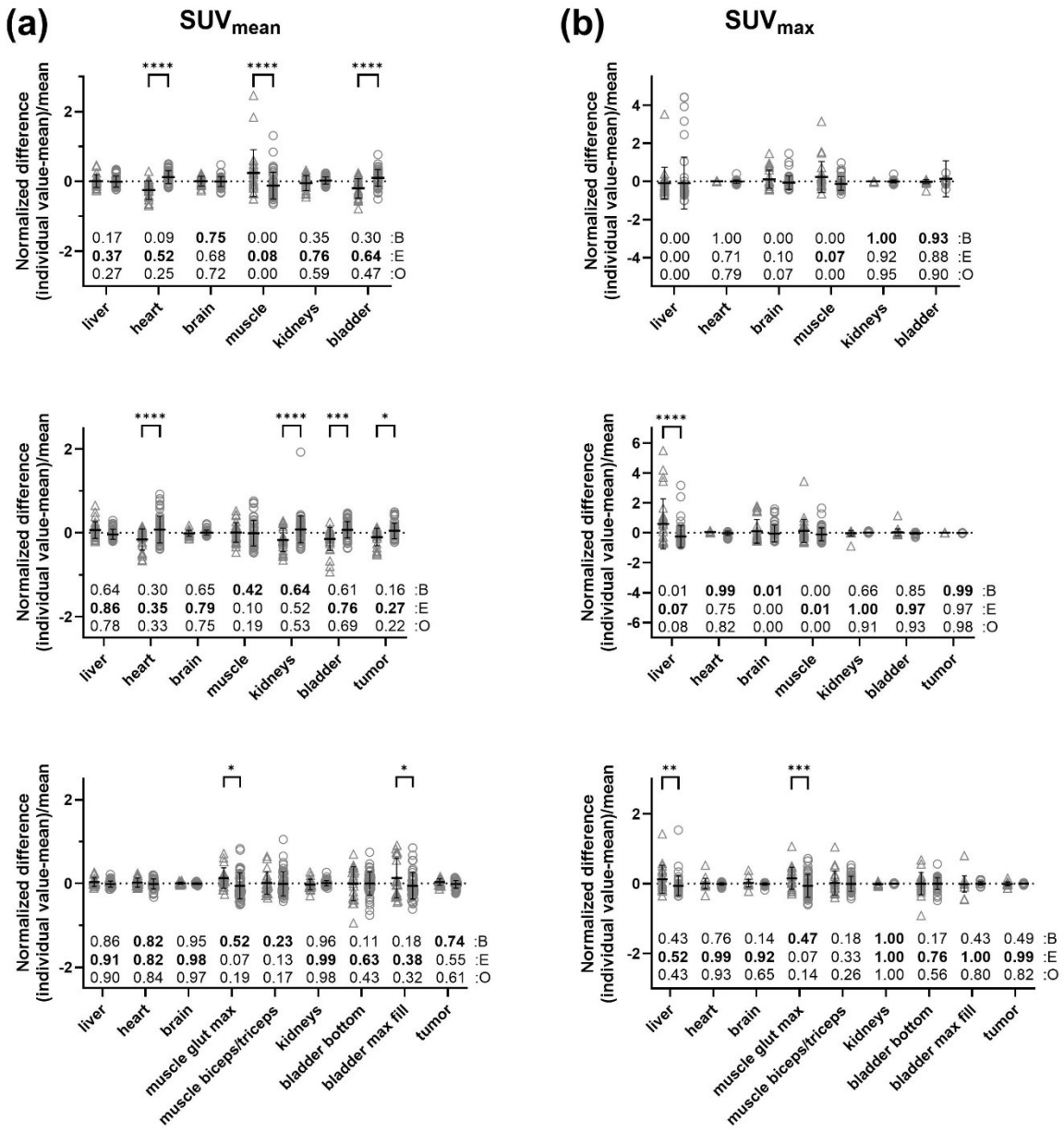


Fig. 2 VOI sizes delineated by the beginner (n=4, open triangle) or expert (n=8, open circle) group on the (**a**) [^{18}F]FDG-PET-only (n=6) and (**b**) [^{18}F]FDG-PET/CT (n=7) images. In (**c**), the

VOI sizes after the standardization procedure are shown. The mean values ± standard deviations are displayed. (*p<0.05; **p<0.01; ***p<0.001; ****p<0.0001; two-way ANOVA followed by Bonferroni multiple comparisons test). The coefficient of variation (%CV) values for each organ are provided separately for beginners and experts. The bold text marks lower %CV values for beginners or experts. (Abbreviations used: bladder – urinary bladder, muscle glut max – muscle gluteus maximus, bladder bottom – bottom of the urinary bladder, bladder max fill – urinary bladder at maximum fill).



505

506 **Fig. 3** (**a**) SUV_{mean} and (**b**) SUV_{max} analysis for the different organs for [^{18}F]FDG-PET-only
507 (upper row), [^{18}F]FDG-PET/CT (middle row) and standardized [^{18}F]FDG-PET/CT (lower row)
508 analysis by beginners (n=4/3, open triangle) and experts (n=8/7, open circle). The normalized
509 difference for each animal is plotted. The mean values ± standard deviations are displayed.
510 (*p<0.05; **p<0.01; ***p<0.001; ****p<0.0001; two-way ANOVA followed by Bonferroni
511 multiple comparisons test). The ICCs for each organ are provided separately for beginners
512 (B), experts (E), and all observers (O). A bold text indicates greater reliability for beginners or
513 experts. (Abbreviations used: bladder – urinary bladder, muscle glut max – muscle gluteus

maximus, bladder bottom – bottom of the urinary bladder, bladder max fill – urinary bladder at maximum fill).

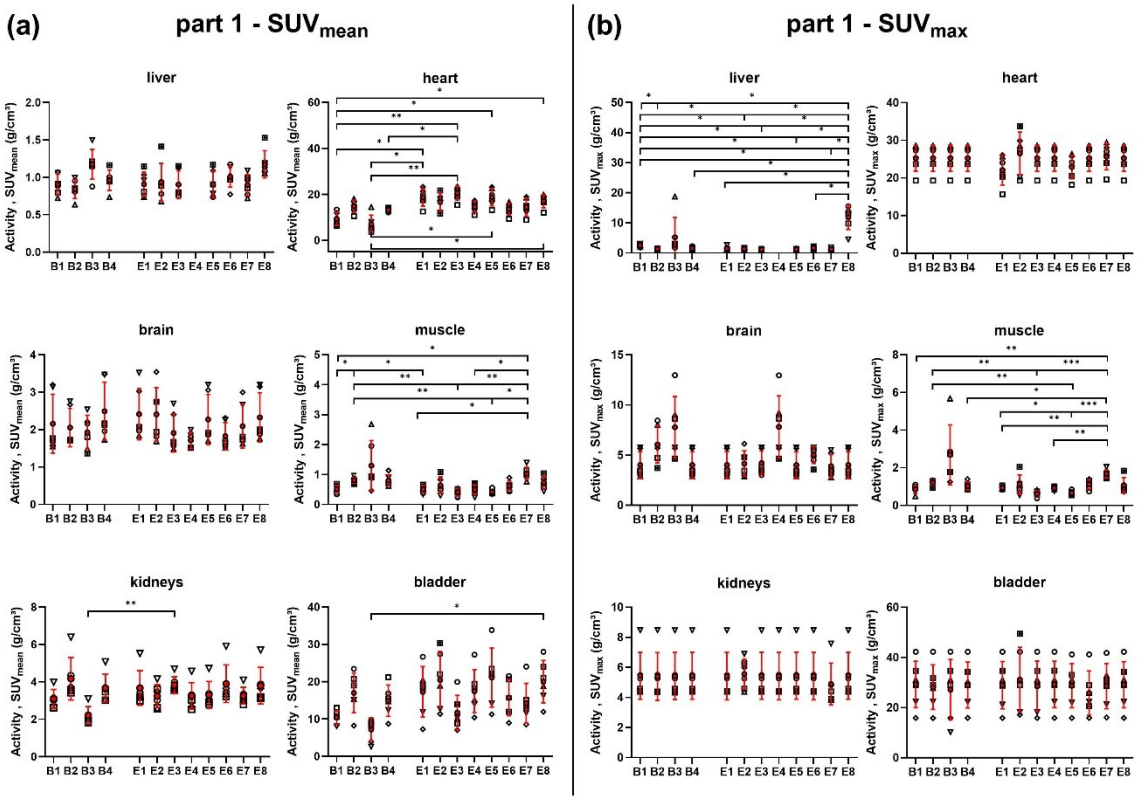
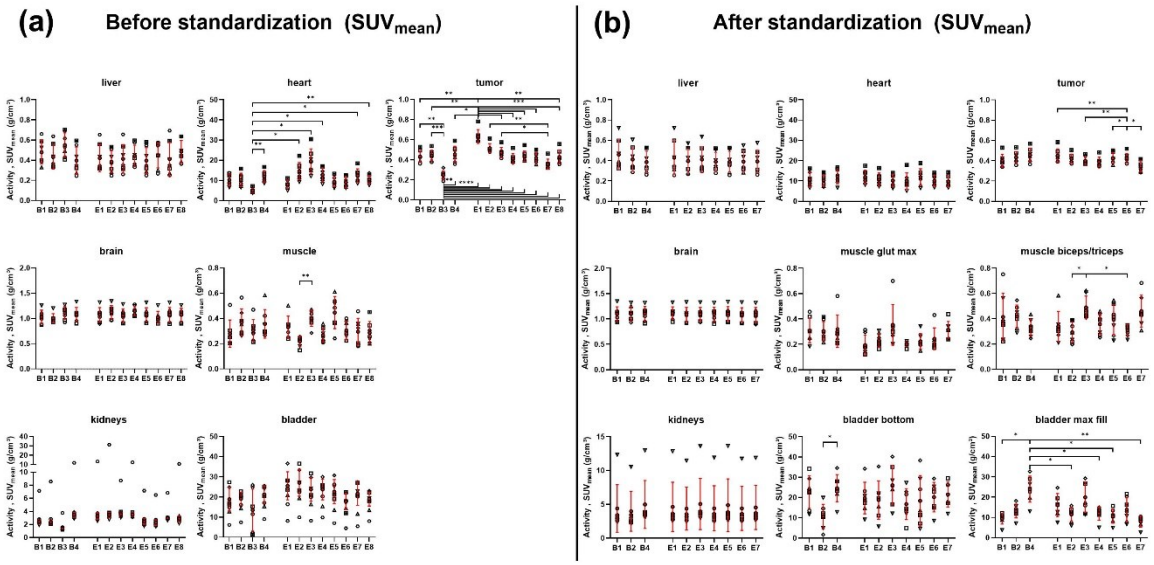


Fig. 4 (**a**) SUV_{mean} and (**b**) SUV_{max} analysis as a function of beginner or expert observers for [^{18}F]FDG-PET-only data from the liver, heart, brain, muscle, mean kidney, and urinary bladder. Individual values, as well as the mean ± standard deviation, are displayed. B1-4: beginners 1 to 4; E1-8: experts 1 to 8. Differences between individual observers were assessed by Brown-Forsythe and Welch ANOVA followed by Dunnett's T3 multiple comparisons test (*p<0.05; **p<0.01; ***p<0.001; ****p<0.0001). Expert 4 did not analyze the liver. (Abbreviations used: bladder – urinary bladder).



528

529 **Fig. 5** SUV_{mean} analysis as a function of beginner or expert observers from [^{18}F]FDG-PET/CT
530 data for the selected regions (**a**) before and (**b**) after standardization. Individual values, as well
531 as the mean ± standard deviation, are displayed. B1-4: beginners 1 to 4; E1-8: experts 1 to 8.
532 Differences between individual observers were assessed by Brown-Forsythe and Welch
533 ANOVA followed by Dunnett's T3 multiple comparisons test (*p<0.05; **p<0.01; ***p<0.001;
534 ****p<0.0001). The analyses of observers B3 and E8 were not included in the standardized
535 [^{18}F]FDG-PET/CT analysis because they were not applicable for the standardized protocol.
536 (Abbreviations used: bladder – urinary bladder, muscle glut max – muscle gluteus maximus,
537 bladder bottom – bottom of the urinary bladder, bladder max fill – urinary bladder at maximum
538 fill).

539

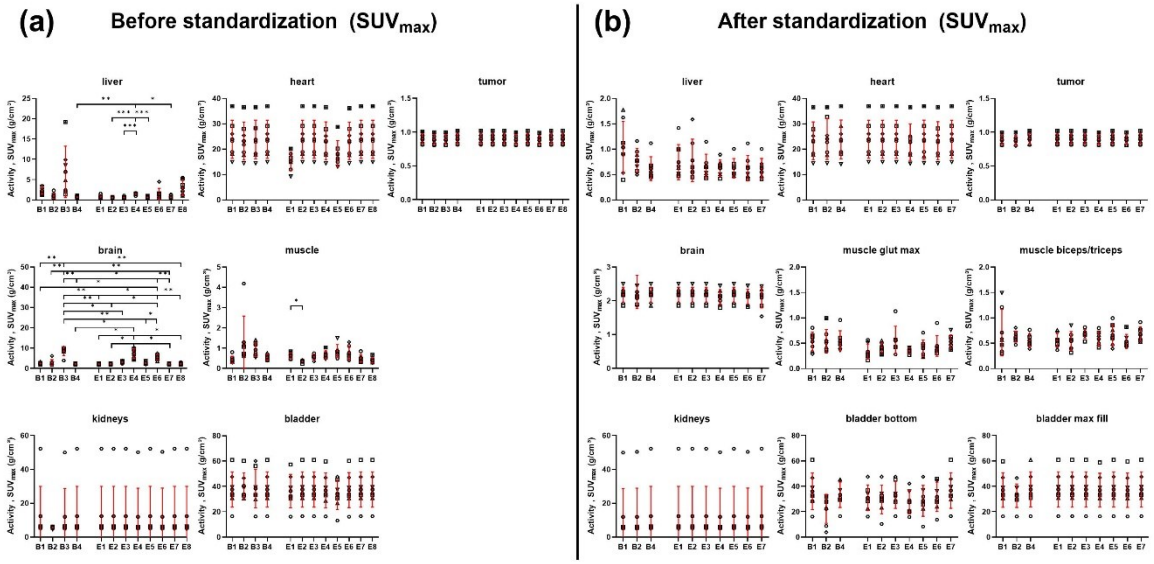


Fig. 6 SUV_{max} analysis as a function of beginner or expert observers from [^{18}F]FDG-PET/CT data for the selected regions (**a**) before and (**b**) after standardization. Individual values, as well as the mean ± standard deviation, are displayed. B1-4: beginners 1 to 4; E1-8: experts 1 to 8. Differences between individual observers were assessed by Brown-Forsythe and Welch ANOVA followed by Dunnett's T3 multiple comparisons test (*p<0.05; **p<0.01; ***p<0.001; ****p<0.0001). The analyses of observers B3 and E8 were not included in the standardized [^{18}F]FDG-PET/CT analysis because they were not applicable for the standardized protocol. (Abbreviations used: bladder – urinary bladder, muscle glut max – muscle gluteus maximus, bladder bottom – bottom of the urinary bladder, bladder max fill – urinary bladder at maximum fill).