Physica Medica 59 (2019) 117–126

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

# Impact of acquisition count statistics reduction and SUV discretization on PET radiomic features in pediatric 18F-FDG-PET/MRI examinations

Marco Branchini[a,*], Alessandra Zorz[a], Pietro Zucchetta[b], Andrea Bettinelli[a], Francesca De Monte[a], Diego Cecchin[b], Marta Paiusco[a]

[a] Medical Physics Department, Veneto Institute of Oncology IOV – IRCCS, Padova, Italy
[b] Nuclear Medicine Unit, Department of Medicine DIMED, University Hospital of Padua, Padova, Italy

ARTICLE INFO

*Keywords:*
Radiomics
PET features
PET/MRI
Reproducibility
Count statistics reduction
SUV discretization

ABSTRACT

*Purpose:* The evaluation of features robustness with respect to acquisition and post-processing parameter changes is fundamental for the reliability of radiomics studies. The aim of this study was to investigate the sensitivity of PET radiomic features to acquisition statistics reduction and standardized-uptake-volume (SUV) discretization in PET/MRI pediatric examinations.

*Methods:* Twenty-seven lesions were detected from the analysis of twenty-one 18F-FDG-PET/MRI pediatric examinations. By decreasing the count-statistics of the original list-mode data (3 MBq/kg), injected activity reduction was simulated. Two SUV discretization approaches were applied: 1) resampling lesion SUV range into fixed bins numbers (FBN); 2) rounding lesion SUV into fixed bin size (FBS). One hundred and six radiomic features were extracted. Intraclass Correlation Coefficient (ICC), Spearman correlation coefficient and coefficient-of-variation (COV) were calculated to assess feature reproducibility between low tracer activities and full tracer activity feature values.

*Results:* More than 70% of Shape and first order features, and around 70% and 40% of textural features, when using FBS and FBN methods respectively, resulted robust till 1.2 MBk/kg. Differences in median features reproducibility (ICC) between FBS and FBN datasets were statistically significant for every activity level independently from bin number/size, with higher values for FBS. Differences in median Spearman coefficient (i.e. patient ranking according to feature values) were not statistically significant, varying the intensity resolution (i.e. bin number/size) for either FBS and FBN methods.

*Conclusions:* For each simulated count-statistic level, robust PET radiomic features were determined for pediatric PET/MRI examinations. A larger number of robust features were detected when using FBS methods.

## 1. Introduction

Imaging has a prominent role in the diagnosis and management of cancer patients. Computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) provide complementary information about tumor anatomy and physiology. In particular, 18F-FDG-PET imaging has become essential in staging and treatment response assessment due to its sensitivity to track glucose-avid lesions [1]. In recent years, the hybrid modality PET/MRI has been introduced in clinical practice [2]. Its main advantages compared to widespread PET/CT are essentially the superior contrast of soft tissue images, the possibility of acquiring physiologic information with MR and the lack of x-ray exposition from the CT component. This last property is particularly interesting for pediatric imaging because of their higher radiosensitivity [3,4].

Nowadays, the field of Radiomics, i.e. the extraction of a large number of quantitative features from tumor images for prognostic profiling of cancer lesions [5], is capturing more and more interest in the cancer research community. Radiomic biomarkers are determined by searching for correlations of radiomic features with patient outcomes. In particular, radiomics enables a more comprehensive non-invasive description of whole tumor heterogeneity than the analysis of a limited number of biopsy samples. Moreover, radiomics uses clinically acquired images without the need for different or additional imaging examinations.

PET radiomics, first investigated by the pioneering study of El Naqa et al. [6] for cervix and head and neck tumors, has been employed to derive bio-markers for many other cancers [7–10]. The repeatability

* Corresponding author.
*E-mail address:* branchini.marco@gmail.com (M. Branchini).

https://doi.org/10.1016/j.ejmp.2019.03.005
Received 1 December 2018; Received in revised form 2 March 2019; Accepted 7 March 2019
Available online 16 March 2019

and sensitivity of radiomics features to different acquisition protocols and post-processing (i.e. reproducibility or robustness) have been carefully investigated [11]. For example, the impact of reconstruction and delineation on repeatability of radiomic features was investigated by van Velden et al., Yan et al. [12,13] while Leijeneaar et al. focused on the impact of Standardized Uptake Volume (SUV) discretization to textural features [14,15]. These efforts are of outmost importance for the interchangeable use of PET-derived features when images are acquired from different scanners or for multicenter studies and for a safe and effective translation of radiomics from research into clinical practice. Indeed, a standardization of imaging parameters is necessary to improve radiomics reliability and prognostic power [14,16,17]. In this context, published studies focusing on PET/MRI radiomics are still limited [18,19].

The feasibility of activity reduction in 18F-FDG-PET/MRI examinations has been investigated with particular focus to lesion detectability, SUV reliable measurements [20] and quantitative tumor burden information [21]. Tsujikawa et al. [22] recently presented a study focused on absolute variations and correlations of a limited number of PET features between PET/CT and PET/MR acquisitions in oropharyngeal and gynecological cancer, considering two different PET reconstructions from PET/MRI examinations based on different acquisition durations and finding high correlations between feature values extracted from the three images. To the best of our knowledge, nobody has analyzed the robustness of a large set of PET radiomics features and the impact of different methods of SUV discretization on patient data when decreasing the acquisition count statistics.

The aim of this study was indeed twofold: first, to investigate the sensitivity of a large set of PET radiomic features to acquisition statistics reduction in pediatric PET/MRI examinations and, second, to evaluate the effect of different discretization approaches on textural feature robustness while decreasing the injected activity.

## 2. Materials and methods

### 2.1. PET/MRI acquisition, activity reduction simulation and VOI delineation

Twenty-one whole-body 18F-FDG-PET/MRI pediatric examinations performed on a 3T clinical PET/MR scanner (Biograph mMR, Siemens Healthcare, Erlangen, Germany) were considered. Patients consisted of 17 children (9 girls and 8 boys, four patients imaged twice) with an average age of 12.7 ± 3.1 (6.9–16.9) years and a body mass index of 19.6 ± 4.4 (13.3–31.2) kg/m$^2$ who were diagnosed with solid cancers (11 Hodgkin's lymphoma, 2 non-Hodgkin's lymphoma, 4 rhabdomyosarcomas). 18F-FDG activity equal to 3 MBq/kg body weight (mean administered activity: 139.4 ± 48.2, range: 86–263 MBq) was administered according to the European Guidelines [23]. PET/MR acquisitions started 124.9 ± 40.4 min (range: 66–214) after injection. All patients fasted at least 6 h before the examination. Attenuation correction was performed with Siemens MLAA approach based on standard MR Dixon sequence. Corrections for random coincidences, daily normalization, dead-time losses and scatter were applied. The reconstruction algorithm employed was 3D Ordinary Poisson Ordered Subset Expectation Maximization (OSEM) with 3 iterations, 21 subset, 172 × 172 matrix, a voxel size of 4.17 × 4.17 × 2.03 mm and 4 mm Gaussian filter, as recommended by the manufacturer.

To simulate acquisition count statistics reduction, for each examination the full tracer activity (FTA) list-mode data (3 MBq/kg, 5 min/bed position) were truncated at 4, 3, 2.5, 2 and 1 min for bed position corresponding to proportional decreasing injected activities (2.4, 1.8, 1.5, 1.2 and 0.6 MBq/kg) [20,24]. In the following, we refer to the images at simulated lower activity as LTA. Six simulated images were thus reconstructed for each patient using the research software package Siemens e7-tools (Fig. 1).

Images were imported in RayStation software (ver. 5.0.2, RaySearch Laboratories, Stockholm, Sweden) for lesion delineation. Segmenting the VOIs in each image is more realistic compared to the use of a fixed VOI representing the actual situation in which only the acquisition with reduced statistical counts is available. For each examination, VOIs were thus defined applying a threshold at 40% of SUVmax in each reconstructed image using the same segmentation's bounding box. Eventually, VOIs were checked by an expert nuclear physician.

Fifty-six lesions were detected in the FTA image with average volume equal to 14.2 ml. Lesions with volume < 5 ml were discarded from the analyses to minimize partial volume effects [13,25,26]: twenty-seven lesions in total were thus considered in the analysis (average volume 26.7 ± 29.2 ml, range 5.2–135.1 ml).

### 2.2. SUV discretization and features extraction

Images and VOIs were imported into the open-source IBEX software platform (v.1.0β, M.D. Anderson, Houston, Texas, USA) [27] where images were normalized to the body weight SUV. Only voxels with more than 50% of their volume inside VOI contour were analyzed for all feature categories. SUV discretization was performed for textural features calculation to decrease noise and have a finite number of intensity values for meaningful texture analysis [14]. Two SUV discretization approaches were applied to VOI voxels for textural features computation:

– for each voxel inside the lesion, SUVs were resampled at 16 and 64 levels [14,26], i.e. a fixed number of bins (FBN) were considered, by using the Ibex preprocess modules BitDepthRescaleRange and RoundtoNearest. In particular, the resampling was performed using:

$$R(x) = round\left(\frac{I(x) - SUV_{min}}{SUV_{max} - SUV_{min}} * (N_{bin}-1)\right) + 1$$

where I(x) is the SUV of the voxel x, $SUV_{min}$ and $SUV_{max}$ are the minimum and maximum SUV in the VOI and $N_{bin}$ is equal to 16 or 64 [14,28]. These values for bin numbers were chosen to obtain similar average bin width to the second approach described next in order to get similar average level of detail in the description of tumor heterogeneity. Furthermore, 64 bins have been employed in many studies [14,26,29] and was proven to be optimal for FBN methods [30]. In this way, lesion-dependent bin sizes were obtained (these methods will be called "FBN16" and "FBN64" in the text).

– SUV were discretized using a fixed bin size (FBS), i.e. a fixed intensity resolution, independent from lesion SUV, as proposed by Orlhac et al. [31], using:

$$R(x) = round\left(\frac{I(x)}{L} * (D-1)\right) + 1 \cong round\left(\frac{I(x)}{B}\right)$$

where B is equal to 0.25 or 0.10 (these methods will be called "FBS025" and "FBS010"). This resampling was performed in Ibex fixing the textural features parameters GrayLimits equal to 0 and 20 (L = 20) to include all the tumor SUVs of our dataset and NumLevels (D) equal to 80 or 200. This operation is equivalent to rounding SUVs to the nearest multiple of 0.25 or 0.10. A bin size of 0.25, also used by other authors [12,14], was chosen to obtain SUV bins (20 on average, since the mean SUVmax in our dataset was 8.5 g/ml and a 40% threshold was applied for delineation) and to have relatively large bin widths. To investigate if using smaller bin sizes has some impact on feature robustness, a bin size equal to 0.10 was also considered [32]. In this case, the resulting average bin number was about 50.

One-hundred-six features were extracted (Table 1): 15 Shape, 6 Intensity Histogram, 27 Intensity Direct, 22 Gray Level Co-occurrence Matrix 2D (GLCM), 22 Gray Level Co-occurrence Matrix (GLCM) 3D

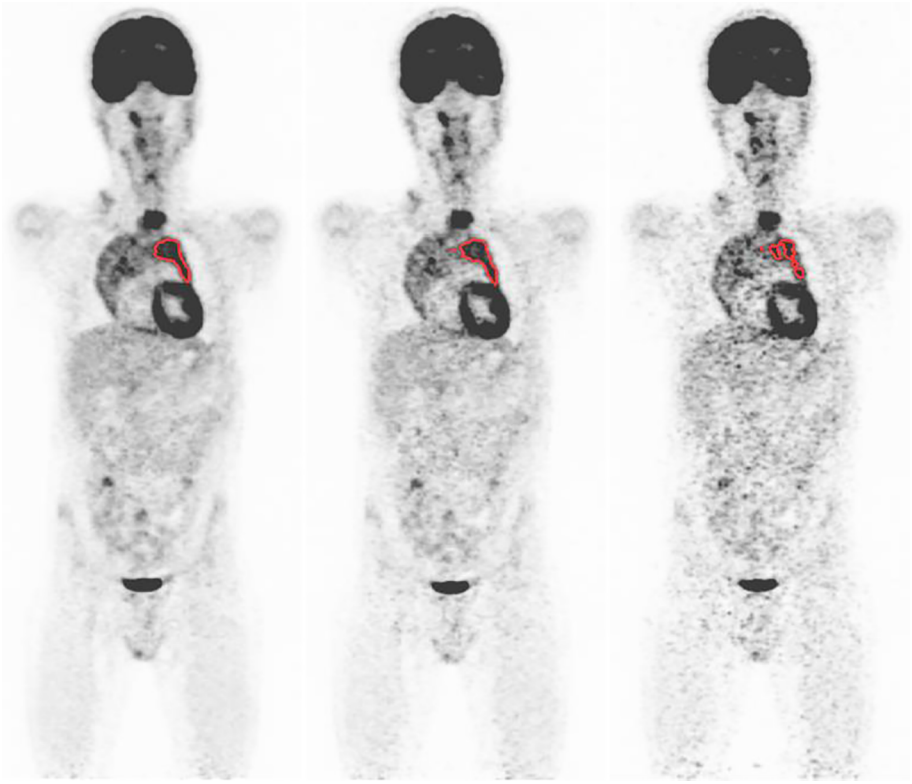

**Fig. 1.** Coronal view of the same examination reconstructed at 3 MBq/kg, 1.5 MBq/kg and 0.6 MBq/kg (from right to left) each with the same lesion segmented at 40% of SUVmax.

and 11 from Gray Level Run Length Matrix 2D (GLRLM). GLCM and GLRLM were calculated for a one-voxel neighborhood distance and averaged over all possible directions (4 directions for 2D and 13 for 3D). For Intensity Histogram feature extraction, 80 bins in the SUV range 0–20 were considered. A complete list of references describing the extracted features is reported in [27].

### 2.3. Data analysis

Features were calculated for each lesion and lower activity level (LTA), and then compared to feature values calculated in the Full Tracer Activity (FTA) image, which was considered as ground truth. For each feature, differences in feature values from those calculated in FTA images were assessed by calculating three quantitative metrics:

a) Intraclass Correlation Coefficient (ICC) using a one-way random single-measure model defined as:

$$ICC(f, LTA) = \frac{MSB(f_{LTA}, f_{FTA}) - MSW(f_{LTA}, f_{FTA})}{MSB(f_{LTA}, f_{FTA}) + MSW(f_{LTA}, f_{FTA})}$$

where $f_{LTA}$ and $f_{FTA}$ are the values of the feature f evaluated at the simulated injected activity level LTA and FTA, MSB and MSW are the mean squares between-subjects and within subjects obtained by one-way ANOVA;

b) Spearman's rank correlation coefficient (ρ);

c) Coefficient of Variation (COV) defined as:

$$COV(f, LTA) = \frac{SD(f_{LTA}/f_{FTA})}{mean(f_{LTA}/f_{FTA})}$$

**Table 1**
Groups and names of extracted features.

| Group | No. of features | Features |
|---|---|---|
| Shape[a] | 15 | Compactness 1, Compactness 2, Convex, Convex Hull Volume, Convex Hull Volume 3D, Max 3D Diameter, Mean Breadth, Number Of Voxel, Orientation, Roundness, Spherical Disproportion, Sphericity, Surface Area, Surface Area Density, Volume |
| Intensity Histogram[a] | 6 | Inter Quartile Range, Kurtosis, Mean Absolute Deviation, Median Absolute Deviation, Skewness, Range |
| Intensity Direct[a] | 30 | Energy, Global Entropy, Global Max, Global Mean, Global Median, Global Std, Global Uniformity, Inter Quartile Range, Kurtosis, Local Entropy Max, Local Entropy Mean, Local Entropy Median, Local Entropy Min, Local Entropy Std, Local Range Max, Local Range Mean, Local Range Median, Local Range Min, Local Range Std, Local Std Max, Local Std Mean, Local Std Median, Local Std Min, Local Std Std, Mean Absolute Deviation, Median Absolute Deviation, Range, Root Mean Square, Skewness, Variance |
| GLCM 2D/3D[b] | 22 + 22 | Autocorrelation, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Entropy, Dissimilarity, Energy, Entropy, Homogeneity, Homogeneity 2, Information Measure of Correlation 1, Information Measure of Correlation 2, Inverse Difference Moment Normalized, Inverse Difference Normalized, Inverse Variance, Maximum Probability, Sum Average, Sum Entropy, Sum Variance, Variance |
| GLRLM[b] | 11 | Gray Level Nonuniformity, High Gray Level Run Emphasis, Long Run Emphasis, Long Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Low Gray Level Run Emphasis, Run Length Nonuniformity, Run Percentage, Short Run Emphasis, Short Run High Gray Level Emphasis, Short Run Low Gray Level Emphasis |

[a] Not dependent from SUV discretization.
[b] Computed for all SUV discretization methods.

where, as before, $f_{LTA}$ and $f_{FTA}$ are the values of the feature f evaluated at the simulated injected activity level LTA and FTA. Mean and SD were evaluated on different subjects (i.e. lesions) for every LTA level.

ICC and ρ values larger than 0.90 were considered indicators of excellent reliability [12] and correlation of patients rankings according to features values [14], respectively. Furthermore, a COV [13,33] equal to 0.20 was arbitrary employed to identify features with a large variability of the ratio between LTA and FTA feature values indicating feature instability. If a feature respected all three of the criteria (ICC > 0.90, Spearman coefficient > 0.90 and absolute value of COV < 0.20) for each activity level down to activity level A, it was considered robust until this simulated injected activity.

At each decreased activity level, paired Wilcoxon signed-rank tests were computed for ICC, ρ and COV to assess statistically significant differences of feature robustness to activity reduction when using one of the four discretization methods compared to the others. P values less than 0.01 were considered significant. To quantify the sensitivity of each textural feature to the applied discretization method when decreasing activity, a difference of ICC values > 0.03 was considered relevant similarly to the study of van Velden et al. [12]. Between these sensitive features, those with ICC > 0.90 when using FBS methods but ICC < 0.90 when using FBN methods and vice versa were selected.

Furthermore, to facilitate the evaluation of radiomic feature reproducibility compared to that of standard SUV measurements, the average relative difference (RD) in comparison to FTA values was calculated for each feature f as:

$$RD(f, LTA) = \frac{1}{N_L}\sum_{L}\frac{f_L(LTA) - f_L(FTA)}{f_L(FTA)}$$

where L stands for the lesions and $N_L$ is the number of lesions.

Moreover, to investigate whether feature robustness is dependent on tumor type, Hodgkin lymphoma (20 lesions) was compared to other tumors (7 lesions). In addition, a specific analysis was performed on features that have been proven to hold prognostic or predicting value for Hodgkin lymphoma [34–37].

## 3. Results

### 3.1. Sensitivity of features to decreasing activity and impact of SUV discretization method

Shape and Intensity Direct features showed good performance when decreasing activity with more than 70% of robust features up to 1.2 MBq/kg even if robust Shape features dropped to about 20% at the lowest activity level (Fig. 2). Between the six considered Intensity Histogram features, four resulted robust down to 0.6 MBq/kg. (Supplemental Materials).

A consistent fraction of textural features was robust to activity reduction until 1.2 MBq/kg considering FBS, while a lower number of robust features were achieved using FBN for all activity levels (Fig. 3). In particular, considering FBS025 and FBS010, around 70% of features resulted robust to activity reduction on the basis of the employed criteria until 1.2 MBq/kg. However, considering all textural features, about 40% of features were robust at 1.2 MBq/kg for FBN16 and FBN64 (Fig. 3). The difference between FBS and FBN was particularly evident for GLRLM features, especially at 1.2 and 0.6 MBq/kg, where less than 20% of the considered GLCM features was found robust for both FBN methods (Fig. 3).

### 3.2. Statistical analysis of ICC, Spearman correlation coefficient and COV for textural features

When using FBS discretization, higher median values and, generally, a lower IQR (1st–3rd interquartiles range) were displayed for ICC and Spearman ρ in comparison to FBN datasets (Fig. 4 and Supplemental material). Considering all textural features, the difference in median ICC and Spearman ρ coefficient between FBS and FBN datasets were statistically significant for every activity level (paired Wilcoxon signed-rank tests) independently from bin size or bin number ($p \leq 0.003$ in 40/40 tests). Differences in median COV between FBS and FBN datasets (Supplemental materials) were not statistically significant for all activity levels except for 2.4 MBq/kg where each FBS dataset was statistically different to both FBN and for 1.2 MBq/kg, where FBS010 was statistically different to FBN16. However, the COVs of robust features to activity reduction were generally lower for FBS discretization method than FBN method as shown in Supplemental Fig. 3 for 1.5 MBq/kg.

When comparing FBS025 with FBS010, ICC values resulted statistically significantly larger for FBS025 only for A = 1.5 MBq/kg, even if the absolute variation between median values was very small (0.988 vs 0.987) while Spearman ρ coefficient and COV did not display any statistically significant difference except for COV at 1.2 MBq/kg (0.090 and 0.088 for FBS025 and FBS010, respectively). Comparing FBN16 and FBN64, differences in ICC values were statistically significant only at A = 1.5 and 1.2 MBq/kg (0.914 vs 0.929 and 0.899 vs 0.896, respectively) while Spearman ρ and COV did not show any statistically significant difference.

The comparison of single textural feature ICC, ρ and COV values between FBS010 and FBN64 (representative of FBS and FBN discretization methods) showed generally higher values of ICC and Spearman ρ for FBS010 in comparison to FBN64 (Fig. 5 and Supplemental material). At 1.5 MBq/Kg, 35 out of 55 features resulted sensitive to the discretization methods applied showing differences of ICC values larger than 0.03 if using FBS010 vs FBN64. In particular, between these features, twelve showed ICC > 0.90 using FBS010 and ICC < 0.90 using FBN64 (eight were robust as defined by the three criteria) and three vice versa (one was robust) (Table 2).

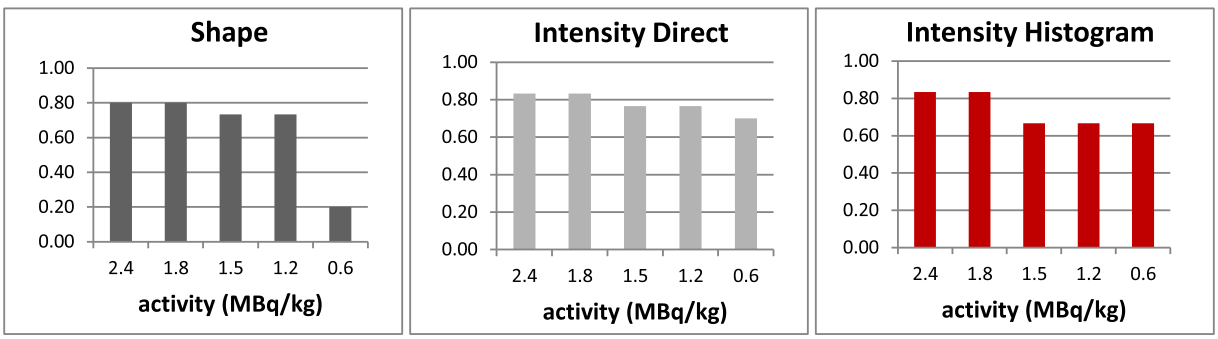


**Fig. 2.** Percentage of robust features for Shape, Intensity Direct and Intensity Histogram groups at LTA levels.

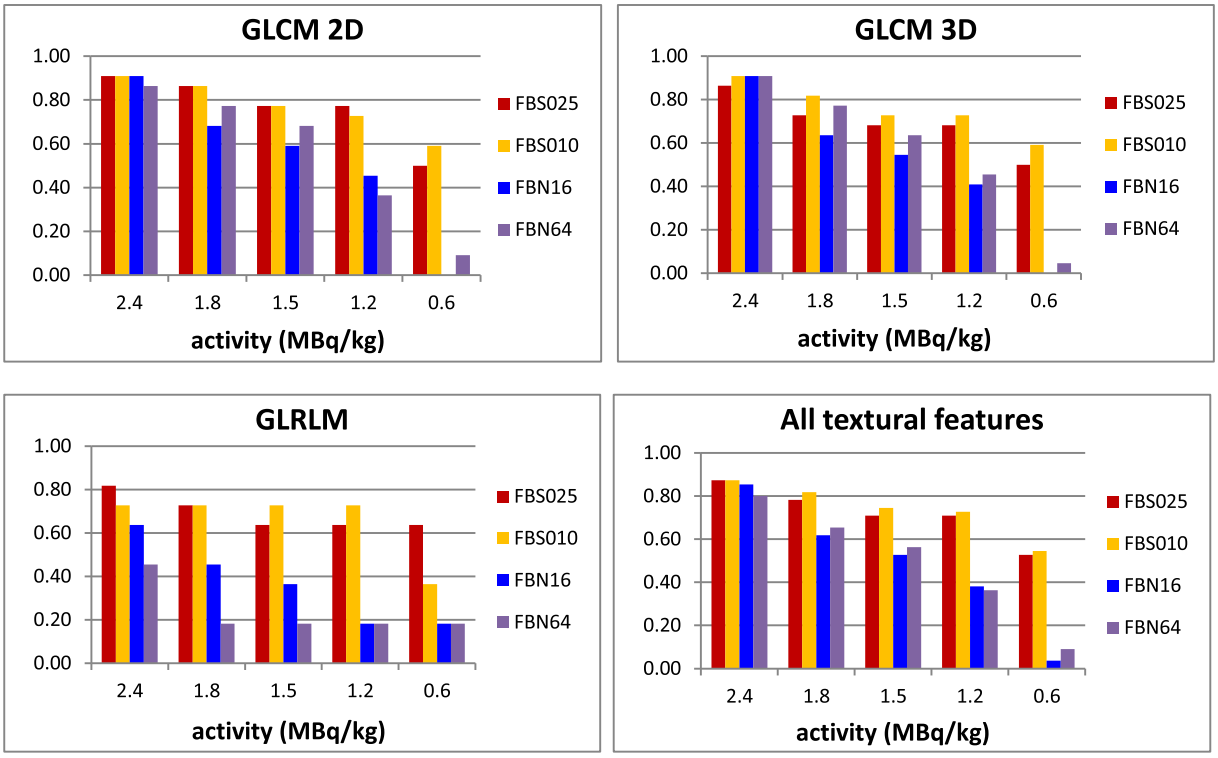


**Fig. 3.** Percentage of robust textural features at LTA levels.

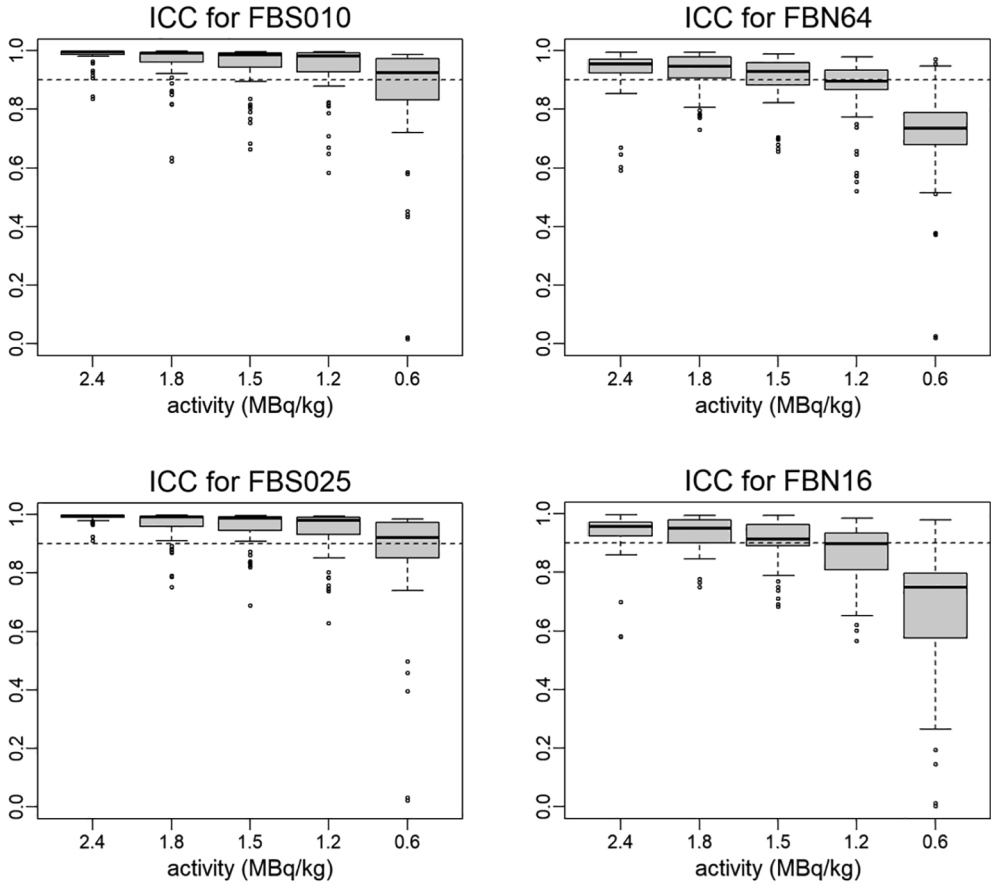


**Fig. 4.** Boxplots of ICC for textural features using FBS010, FBS025, FBN64, FBN16 (dashed line in correspondence of ICC value of 0.90).

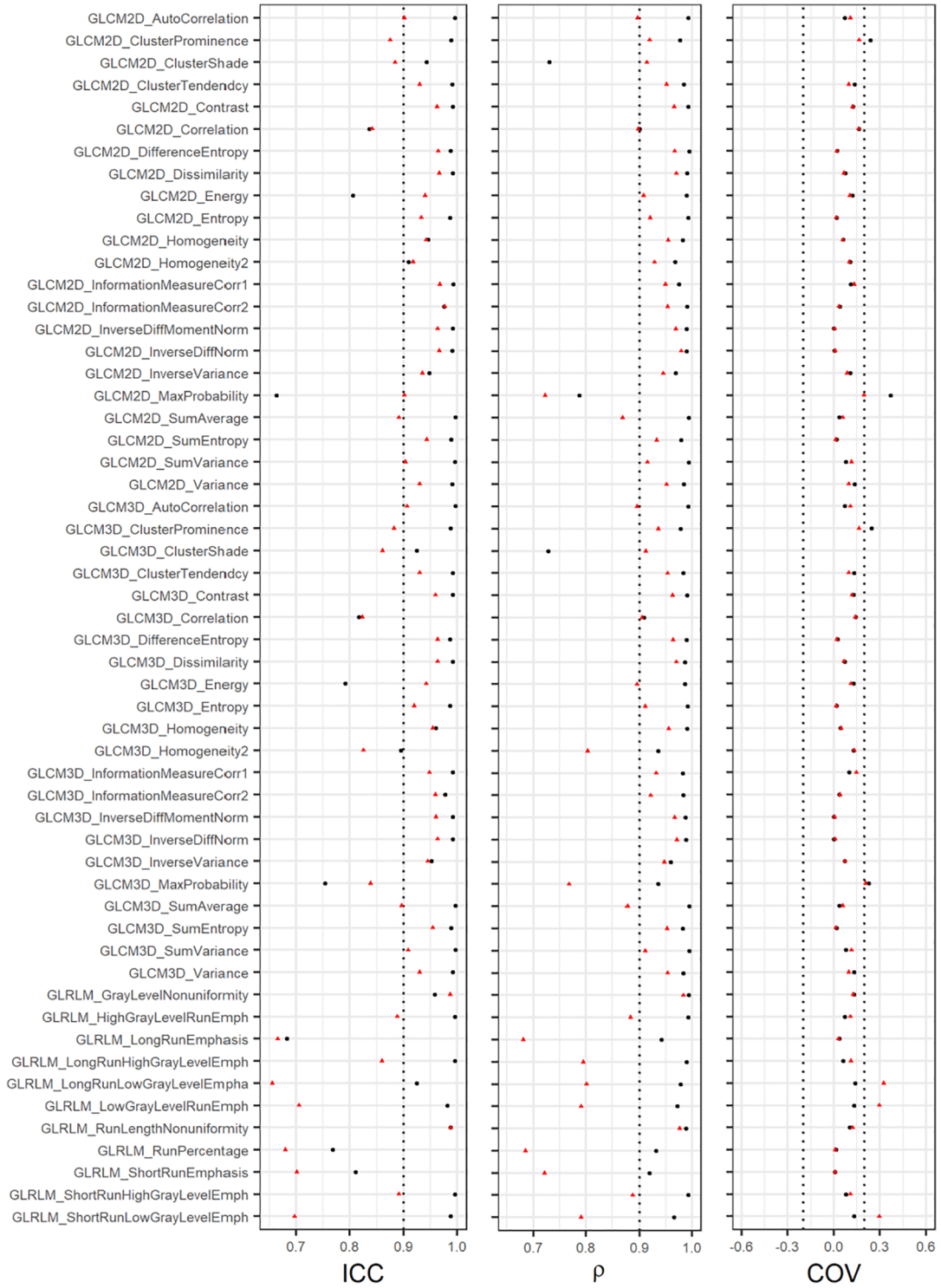


**Fig. 5.** ICC, Spearman coefficient and COV of textural features at 1.5 MBq/kg for FBS010 (black circles) and FBN64 (red triangles).

### *3.3. Analysis of feature RD and features robust for all discretization methods*

SUVmean and SUVmax were robust up to the lowest activity level (0.6 MBq/kg) respecting the employed criteria and showed RD equal to +1% and +3%, respectively, at 1.5 MBq/kg. Focusing on 1.5 MBq/kg, 21 (70%) over 30 Intensity Direct features and 32 (58%), 27 (49%) and 25 (45%) of the 55 textural features, respectively for FBS010/FBS025, FBN64 and FBN16 datasets, were robust and showed absolute RD < 5%, thus comparable to SUV measurements variations.

Twenty-four textural features were found robust for every activity level up to 1.5 MBq/kg independently from the discretization method (Table 3). These features are particularly interesting because it was shown that an activity reduction of 50% is feasible while preserving lesion detectability, SUV measurements and quantitative tumor burden information [20,21] and thus they can be reliably computed from images with clinical usefulness. Among these features, eleven showed absolute RD < 5% for all discretization methods (Table 3).

**Table 2**
Features with difference of ICC values > 0.03 between FBS010 and FBN64 at 1.5 MBq/kg with ICC > 0.90 for one of the two discretization methods. Features robust as defined by the three criteria described in the Methods are marked with *.

| Group | Features with ICC > 0.90 using FBS010 and ICC < 0.90 using FBN64 | Features with ICC > 0.90 using FBN64 and ICC < 0.90 using FBS010 |
|---|---|---|
| GLCM 2D | Cluster Prominence<br>Cluster Shade<br>Sum Average* | Energy*<br>Max Probability |
| GLCM 3D | Cluster Prominence<br>Cluster Shade<br>Sum Average* | Energy |
| GLRLM | High Gray Level Run Emphasis*<br>Low Gray Level Run Emphasis*<br>Long Run High Gray Level Emphasis*<br>Long Run Low Gray Level Emphasis*<br>Short Run High Gray Level Emphasis*<br>Short Run Low Gray Level Emphasis* | |

**Table 3**
Groups and names of robust textural features to activity reduction up to 1.5 MB/kg independently from the discretization employed. Features with absolute RD < 5% are marked with *.

| Group | No. of features | Features |
|---|---|---|
| GLCM 2D | 11 | Cluster Tendendcy, Contrast, Difference Entropy*, Dissimilarity, Homogeneity, Homogeneity 2, Inverse Difference Moment Normalized*, Inverse Difference Normalized*, Inverse Variance, Sum Entropy*, Variance |
| GLCM 3D | 11 | Cluster Tendendcy, Contrast, Difference Entropy*, Dissimilarity, Entropy*, Homogeneity*, Inverse Difference Moment Normalized*, Inverse Difference Normalized*, Inverse Variance, Sum Entropy*, Variance |
| GLRLM | 2 | Gray Level Nonuniformity, Run Length Nonuniformity* |

### 3.4. Analysis of features robustness for Hodgkin lymphoma

A larger proportion of robust shape feature was found for Hodgkin lymphoma than for the other kinds of tumors (non-Hodgkin lymphoma and rhabdomyosarcomas): at 1.5 MBq/kg about 80% and 25% of Shape features were reproducible, respectively (Fig. 6). Also for Intensity Direct and textural features, Hodgkin-lymphomas showed a larger number of robust features even if the relative difference was lower than for the Shape group. Minimum levels of activity for feature robustness of features that have been proven to hold prognostic or predicting value for Hodgkin lymphomas are reported in Table 4.

## 4. Discussion

In the present study, the robustness of radiomic features to acquisition count statistics reduction and the impact of SUV discretization were investigated quantitatively comparing simulated images with lower count statistics to original full count statistics images. To our knowledge, this is the first study that investigated with patient data the effect of SUV dicretization on PET radiomic features when decreasing the acquisition counts statistics. The awareness of the acquisition counts statistics with respect to which radiomics features are robust is necessary to characterize whether they could be reliably employed in low-dose examinations and, consequently, to select low-activity

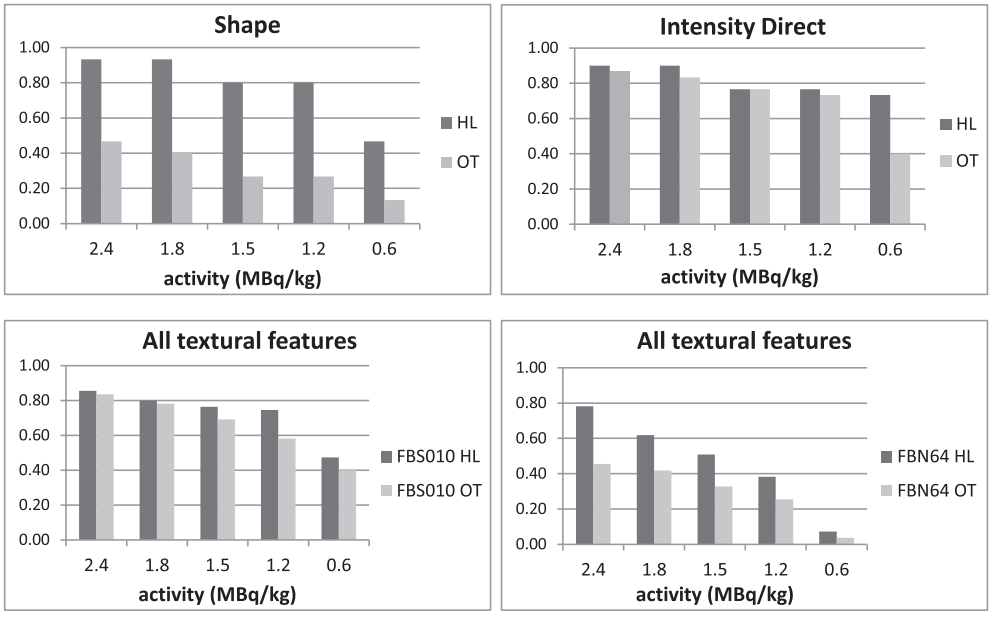


**Fig. 6.** Percentage of robust features of Shape, Intensity Direct and texture at LTA levels for Hodgkin lymphoma (HL) vs other tumors (OT).

**Table 4**
Minimum activity level to which prognostic features for Hodgkin lymphoma were found robust (considering HL lesions only).

| Group | Feature | Minimum activity level for feature robustness [MBq/kg] |
|---|---|---|
| Shape | Volume | 0.6 |
| Intensity Direct | Global Max (SUVmax) | 0.6 |
| | Kurtosis | 1.8 |
| | Global Entropy | 1.8 |
| | Skewness | 3.0 |
| GLCM 2D | Entropy | 0.6 (FBS010) – 1.8 (FBN64) |
| | Maximum Probability | 3.0 |
| | Contrast | 1.2 |
| | Dissimilarity | 0.6 (FBS010) – 1.2 (FBN64) |
| GLCM 3D | Entropy | 0.6 (FBS010) – 2.4 (FBN64) |
| | Maximum Probability | 3.0 |
| | Contrast | 1.2 |
| | Dissimilarity | 1.2 |
| GLRLM | Short Run Emphasis | 3.0 |
| | Low Gray Level Run Emphasis | 1.2 (FBS010) – 3.0 (FBN64) |
| | Short Run High Gray-level Emphasis | 0.6 (FBS010) – 2.4 (FBN64) |

protocols for radiomics studies. Furthermore, it would be of remarkable value to assess whether a sensible comparison between results of different studies using different injected activity or acquisition duration [24] protocols is possible. Activity reduction is particularly important for pediatric patients because they are more sensitive to radiation and, due to longer life expectancy, have a greater probability of developing radio-induced cancers. Recent studies [20,21] have proven that an injected activity reduction up to 50% (from 3 MBq/kg to 1.5 MBq/kg) is feasible in pediatric PET/MR while preserving lesion detectability and SUV measurements.

Feature robustness was analyzed focusing on measurement reliability (ICC), patient ranking correlations (Spearman's rank correlation coefficient) and dispersion of feature values at LTA in comparison to FTA (COV). Each of these metrics is important for robustness and reliability assessment, each highlighting different aspects. The ICC metric describes reliability between measurements, taking in consideration both intra-subjects (i.e. lesions) and inter-subject variability. Employing ICC with a one-way random single-measure model, the analysis of the difference between feature values at LTA and FTA is comparable to the feature repeatability investigation performed in other studies [12,14]. At simulated injected activity of 1.5 MBq/kg, 16 Intensity Direct features (out of 30) and 23 and 2 textural features (out of 55), using FBS and FBN methods respectively, showed ICC values larger than the ICC value reported for the repeatability of SUVmean and SUVpeak (both equal to 0.989) by van Velden et al. [12]. Considering FBS, thus, a significant fraction of textural features showed reproducibility to activity reduction comparable to SUVpeak "test-retest" repeatability. The analysis of Spearman rank correlation coefficient is particularly useful for the interpretation of radiomic features as suggested by Leijenaar et al. [14]. Indeed, if patients ranking according to feature values is only marginally affected when decreasing injected activity, similar patient stratifications should be achieved indifferently at LTA and FTA when correlating features values to patient outcome or tumors phenotypes. Intra-patient feature COVs were calculated for the ratios between values at LTA and FTA. In this way, the standardized variability of feature relative changes from FTA values was assessed. The larger the intra-patient COV of a feature (i.e. variability of the ratios), the lower the reproducibility of the feature to activity reduction.

Having assessed all three of these reliability metrics, it was proven that around 70% of shape and intensity direct features and more than 55% or 70% of textural features (using FBN or FBS discretization methods) have excellent reproducibility when reducing activity up to 1.5 MBq/kg in PET/MR pediatric examinations with their corresponding values computed on FTA images.

The impact of SUV resampling on sensitivity to activity reduction was investigated in depth in this study. Four different discretization methods were analyzed: FBS025, FBS010, FBN16 and FBN64. The selected bin sizes and bin numbers are representative of highly detailed (FBS010 and FBN64) and coarse-grained (FBS025 and FBN16) descriptions of tumor SUV distributions for the analyzed patients. It was shown that resampling SUV using a fixed-bin size method resulted in better overall robustness of textural features than using a fixed bin number method i.e. lesions dependent bin sizes. Indeed, differences in textural features ICC and Spearman correlation coefficient between FBS and FBN datasets were statistically significant at every activity level. Furthermore, the number of robust textural features was larger when using FBS in comparison to FBN approaches and the difference became more evident at the lowest activity levels (around 70% vs 40% of textural features were reproducible at 1.2 MBq/kg for FBS and FBN approaches, respectively). Moreover, considering ICC differences larger than 0.03 to select sensitive features to the discretization methods, when halving the injected activity, eight of these sensitive features resulted robust only for FBS010 while one was robust for FBN64 only. It should be noted, nonetheless, that an in-depth analysis of the clinical relevance of the robustness difference of each feature would require careful Bland-Altman analyses that were recently started by our group and will be the focus of further investigation.

This result about the discretization methods is in agreement with the test-retest study of van Velden et al. [12], where a higher repeatability of textural feature values was achieved using a fixed bin-size method of SUV resampling. However, it is in contrast to the recent published study of Presotto et al. [29] in which authors analyzed PET feature sensitivity to acquisition statistics through experimental phantom image analyses. In that study, the authors computed textural features inside spheres of 39 mm diameter for all six sectors of Jazack phantom varying the acquisition duration. They found that feature values had stronger dependence on acquisition duration if using the FBS discretization method in comparison to FBN. Differences with our results could be partly due to the different metrics employed in the analyses and in the experimental setting employed, i.e. uniform activity outside cold road of different diameters. Even if using such a phantom has the advantage of ensuring a high level of repeatability and standardization, it does not fully describe the complexity of tracer distribution inside tumour lesions. A direct comparison of specific feature stability when varying the count statistics between the present study and the recent study by Tsujikawa et al. [22] is not straightforward because of the differences in the scanners, reconstruction corrections (TOF and PSF), patient ages and cancer types, injected activity protocols (fixed 185 MBq vs 3 MBq/kg) and metrics employed in the analyses. Nonetheless, $SUV_{mean}$, kurtosis, entropy (GLCM), homogeneity (GLCM), short run emphasis (GLRLM), and long run emphasis (GLRLM) shared consistent results with the GLCM features showing the higher correlations between different counts statistics while discordant results were found for the stability of first order feature skewness.

The effects of different bin size and bin number were also examined in the present work. For GLCM 3D and GLRLM feature groups, using a thin fixed bin size (FBS010) resulted in a slightly higher number of robust features than FBS025 and FBN64 also showed higher performance than FBN16 for GLCM 2D/3D till 1.5 MBq/kg. These results suggest that textural features are able to robustly characterize high intensity resolution SUV distributions (i.e. using small average bin sizes) inside tumor lesions even at lower injected activity levels. They are also in agreement with the study of Orlhac et al. [25] in which authors consistently showed for three different tumor sets corresponding to different acquisition/reconstruction protocols that at least 32 Gy levels levels should be used for texture indices calculation using a

FBN approach to provide robust metrics with respect to different values of bin number. Furthermore, using 32–64 levels was shown to maximize the repeatability of some features calculated on intensity size-zone matrices (e.g. size-zone variability and large-area emphasis) that were found to be significant predictors of patient response [30].

An interesting methodological consideration about SUV discretization and its effects on inter-patient texture comparison and interpretation was referred by Leijenaar et al. [14]: resampling SUV with an FBS discretization methods is similar to what is routinely done when comparing SUVs between different patients. Indeed, FBS resampling reduces the possible SUVs without changing the intensity resolutions between patients. Furthermore, the FBS approaches that we employed were equivalent to simple rounding operations and, because the lower bound for resampling was set to zero for every lesion [31], the minimum value of the lesion SUV was not subtracted as instead was performed for FBN methods and proposed for the FBS method by Leijenaar et al. [14]. In this way, a completely lesion-independent SUV discretization was performed. Conversely, when using FBN, the scale of measurement (i.e. the intensity resolution) was lesion-dependent resulting in textural measurements that even if numerically identical could describe very different levels of heterogeneity. For example, using FBN, the feature Autocorrelation from GLCM is invariant by multiplication of image grey-level values by a constant. Thus, two SUV distributions that differ only for a large multiplicative constant, would result in identical values of the feature Autocorrelation even if one of the two have all values very close to 1 and the other one has a large value range. Indeed, as observed by Orhlac et al. [31], the absolute range of SUV in a lesion already provides information about its heterogeneity.

We should underline that the results of the present study, even if attractive, came from the analysis of images coming from the same scanner model and same image reconstruction parameters. Different hardware and software can indeed originate differences in image values and, consequently, in feature values. In recent papers, differences coming from the use of different scanners [38,39], reconstruction parameters and algorithm [13,29,39,40], VOI delineations [12,25,32,41] and inter-observer variability in segmentation [11,33,41] have been carefully investigated.

In particular, Yan et al. [13] presented an analysis of feature robustness to reconstruction parameters variation, i.e. voxel size, full width at half maximum of the Gaussian postprocessing filter, and showed that seven heterogeneity descriptors were as robust as SUVpeak and SUVmean. A further reduction in the number of clinical useful features using a fixed bin number discretization method (64 bins) has been proposed in [16] taking into consideration correlation existing among radiomics features and robustness in respect to segmentation methods. Between these features, six were considered in the present study and proved generally low sensitivity to activity reduction too with the exception of GLRLM features if using FBN64: Global Entropy (ID) was robust till 1.2 MBq/kg, Difference Entropy, Inverse difference normalized and Inverse difference moment normalized (GLCM) resulted robust to the lowest activity level if using FBS and to the second to last if using FBN methods; high grey level run emphasis and low grey level run emphasis (GLRLM) were robust for FBS at every activity level while for FBN only High grey level run emphasis was highly robust only at 2.4 MBq/kg (Supplemental material).

Orlhac et al. showed that the segmentation method has a substantial impact on a large number of features [25]. Using a fixed threshold of SUVmax is frequently used in the clinics and has been recently shown to be non-inferior to SUVpeak based segmentation for primary lung tumor volume quantification [42]. Even if it was demonstrated that the repeatability of the metabolically active volume is higher using a fixed threshold of highest peak SUV ($SUV_{hp}$) corrected for local background [43], there is still no consensus as to which segmentation method maximizes PET radiomics feature repeatability and reproducibility when varying different parameters. The delineation method could have an impact also on feature robustness when decreasing count statistics and this point should be investigated in a future study.

Furthermore, many studies showed that a standardization of acquisition protocols, reconstruction parameters and post-processing is essential for a reliable quantification of tumor heterogeneity biomarkers [14,16,17,28,44]. In particular, regarding SUV quantization method, Leijenaar et al. [14] showed that textural feature value depends on SUV intensity resolution and the patient rankings according to feature value obtained using FBS or FBN discretization methods are generally different. Thus, the choice of the discretization method is determinant in the selection of the best possible patient stratification. A fine-tuned selection of the pre-processing to be made for every feature was recently proposed by Fave et al. [45] for CT features in the prediction of lung cancer patient outcomes. In this way, different preprocessings are tested to maximize the prognostic power of each different feature. In principle, this idea could also be investigated for the analyses of PET features but with the disadvantage of increasing the time for the pre-processing.

This study presents some limitations. The reduction of tracer activity was simulated truncating the original list mode data at fractions of bed time, possibly introducing limited bias due to temporal and spatial changes of radionuclide distribution. The range of considered patient ages was rather wide and different types of tumors were considered. In the analysis of features robustness differences between Hodgkin lymphomas and other tumors, a small sample of other tumors was examined. The assessment of radiomics feature robustness in specific type of cancers and patient ages together with an increment of the number of patients analyzed will be the focus of a future work.

## 5. Conclusions

In this study, highly robust PET radiomic features for activity reduction in PET/MRI pediatric examinations were determined for each of the considered simulated injected activity level. These features could be reliably employed for tumor metabolic characterization in low-dose pediatric PET/MR examinations. A consistent fraction of features showed reproducibility similar or better than simply SUV measurements. About half of textural features resulted highly robust till half count statistics reduction independently from the SUV discretization method applied. A larger number of PET textural features were found to be reproducible using FBS discretization approaches compared with FBN especially at the lowest simulated activity levels in particular for GLRLM group, while a less relevant impact on reproducibility was observed varying bin size (FBS) and bin number (FBN).

## Acknowledgement

We thank Judson Jones and Bjoern Jakoby (Siemens Healthcare) for precious assistance with setting up E7-tools reconstruction software.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2019.03.005.

## References

[1] Lin P, Koh ES, Lin M, Vinod SK, Ho-Shon I, Yap J, et al. Diagnostic and staging impact of radiotherapy planning FDG-PET-CT in non-small-cell lung cancer. Radiother Oncol 2011;101(2):284–90.
[2] Schäfer JF, Gatidis S, Schmidt H, Gückel B, Bezrukov I, Pfannenberg CA, et al. Simultaneous whole-body PET/MR imaging in comparison to PET/CT in pediatric oncology: initial results. Radiology 2014;273(1):220–31.
[3] Miglioretti DL, Johnson E, Williams A, Greenlee RT, Weinmann S, Solberg LI, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA Pediatr 2013;167(8):700–7.
[4] BEIR VII PHASE II. Health risks from exposures to low levels of ionizing radiation. Washington, DC: The National Academies Press; 2006.

[5] Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Phys Med 2017;38:122–39.
[6] El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 2009;42(6):1162–71.
[7] Bundschuh RA, Dinges J, Neumann L, Seyfried M, Zsótér N, Papp L, et al. Textural parameters of tumor heterogeneity in $^{18}$F-FDG PET/CT for therapy response assessment and prognosis in patients with locally advanced rectal cancer. J Nucl Med 2014;55(6):891–7.
[8] Cheng NM, Fang YH, Chang J, Huang CG, Tsan DL, Ng SH, et al. Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma. J Nucl Med 2013;54(10):1703–9.
[9] Soussan M, Orlhac F, Boubaya M, Zelek L, Ziol M, Eder V, et al. Relationship between tumor heterogeneity measured on FDG-PET/CT and pathological prognostic factors in invasive breast cancer. PLoS One 2014;9(4):e94017.
[10] Zaidi H, Alavi A, El Naqa I. Novel quantitative PET techniques for clinical decision support in oncology. Semin Nucl Med 2018;48(6):548–64.
[11] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys 2018;102(4):1143–58.
[12] van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol 2016;18(5):788–95.
[13] Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of image reconstruction settings on texture features in 18F-FDG PET. J Nucl Med 2015;56(11):1667–73.
[14] Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep 2015;5:11075.
[15] Desseroit M-C, Tixier F, Weber WA, Siegel BA, Cheze Le Rest C, Visvikis D, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. J Nucl Med 2017;58:406–11.
[16] Buvat I, Orlhac F, Soussan M. Tumor texture analysis in PET: where do we stand? J Nucl Med 2015;56(11):1642–4.
[17] Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv preprint arXiv:1612.07003, 2016.
[18] Yin Q, Hung SC, Wang Lin W, Fielding JR, Rathmell WK, et al. Associations between tumor vascularity, vascular endothelial growth factor expression and PET/MRI radiomic signatures in primary clear-cell-renal-cell-carcinoma: proof-of-concept study. Sci Rep 2017;7:43356.
[19] Antunes J, Viswanath S, Rusu M, Valls L, Hoimes C, Avril N, et al. Radiomics analysis on FLT-PET/MRI for characterization of early treatment response in renal cell carcinoma: a proof-of-concept study. Transl Oncol 2016;9(2):155–62.
[20] Gatidis S, Schmidt H, la Fougère C, Nikolaou K, Schwenzer NF, Schäfer JF. Defining optimal tracer activities in pediatric oncologic whole-body 18F-FDG-PET/MRI. Eur J Nucl Med Mol Imaging 2016;43(13):2283–9.
[21] Zucchetta P, Branchini M, Zorz A, Bodanza V, Cecchin D, Paiusco M, et al. Quantitative analysis of image metrics for reduced and standard dose pediatric (18)F-FDG PET/MRI examinations. Br J Radiol 2019;92(1095):20180438.
[22] Tsujikawa T, Tsuyoshi H, Kanno M, Yamada S, Kobayashi M, Narita N, et al. Selected PET radiomic features remain the same. Oncotarget 2018;9(29):20734–46.
[23] Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. Association of nuclear medicine (EANM).FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0, Vols. Eur J Nucl Med Mol Imaging 2015;42(2):328–54.
[24] Alessio AM, Sammer M, Phillips GS, Manchanda V, Mohr BC, Parisi MT. Evaluation of optimal acquisition duration or injected activity for pediatric 18F-FDG PET/CT. J Nucl Med 2011;52(7):1028–34.
[25] Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med 2014;55(3):414–22.
[26] Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J Nucl Med 2015;56(1):38–44.
[27] Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys 2015;42(3):1341–53. https://doi.org/10.1118/1.4908210.
[28] Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in (18)F-FDG-PET scans of oesophageal cancer. Eur Radiol 2015;25(9):2805–12.
[29] Presotto L, Bettinardi V, De Bernardi E, Belli ML, Cattaneo GM, Broggi S, et al. PET textural features stability and pattern discrimination power for radiomics analysis: an "ad-hoc" phantoms study. Phys Med 2018;50:66–74.
[30] Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med 2012;53(5):693–700.
[31] Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. PLoS One 2015;10(12):e0145063.
[32] Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, Chen W. Robustness of radiomic features in [(11)C]choline and [(18)F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. Vols. Mol Imaging Biol 2016;18(6):935–45.
[33] Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52(7):1391–7.
[34] Bagci U, Yao J, Miller-Jaster K, Chen X, Mollura DJ. Predicting future morphological changes of lesions from radiotracer uptake in 18F-FDG-PET images. PLoS One 2013;8(2):e57105.
[35] Parvez A, Tau N, Hussey D, Maganti M, Metser U. (18)F-FDG PET/CT metabolic tumor parameters and radiomics features in aggressive non-Hodgkin's lymphoma as predictors of treatment outcome and survival. Ann Nucl Med 2018;32:410–6.
[36] Ko KY, Liu CJ, Ko CL, Yen RF. Intratumoral heterogeneity of pretreatment 18F-FDG PET images predict disease progression in patients with nasal type extranodal natural killer/T-cell lymphoma. Clin Nucl Med 2016;41(12):922–6.
[37] Ben Bouallègue F, Tabaa YA, Kafrouni M, Cartron G, Vauchot F, Mariano-Goulart D. Association between textural and morphological tumor indices on baseline PET-CT and early metabolic response on interim PET-CT in bulky malignant lymphomas. Med Phys 2017;44(9):4608–19.
[38] Beichel RR, Smith BJ, Bauer C, Ulrich EJ, Ahmadvand P, Budzevich MM, et al. Multi-site quality and variability analysis of 3D FDG PET segmentations based on phantom and clinical image data. Med Phys 2017;44(2):479–96.
[39] Papp L, Rausch I, Grahovac M, Hacker M, Beyer T. Optimized feature extraction for radiomics analysis of (18)F-FDG-PET imaging. J Nucl Med 2018. pii: jnumed.118.217612.
[40] Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. Eur Radiol 2017;27(11):4498–509.
[41] Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell'Oca I, et al. Quantifying the robustness of [(18)F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. Phys Med 2018;49:105–11.
[42] Mercieca S, Belderbos J, van Loon J, Gilhuijs K, Julyan P, van Herk M. Comparison of SUVmax and SUVpeak based segmentation to determine primary lung tumour volume on FDG PET-CT correlated with pathology data. Radiother Oncol 2018;129(2):227–33.
[43] Frings V, van Velden FH, Velasquez LM, Hayes W, van de Ven PM, et al. Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study. Radiology 2014;273(2):539–48.
[44] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present… any future? Eur J Nucl Med Mol Imaging 2017;44(1):151–65.
[45] Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer. Sci Rep 2017;7(1):588.