Feasibility of radiomic feature harmonization for pooling of [^{18}F]FET or [^{18}F]GE-180 PET images of gliomas

Adrian Jun Zounek [a], Nathalie Lisa Albert [a,b,c], Adrien Holzgreve [a], Marcus Unterrainer [a,d], Julia Brosch-Lenz [a], Simon Lindner [a], Andreas Bollenbacher [a], Guido Boening [a], Rainer Rupprecht [e], Matthias Brendel [a,f,g], Louisa von Baumgarten [c,h], Joerg-Christian Tonn [b,h], Peter Bartenstein [a,b], Sibylle Ziegler [a], Lena Kaiser [a,*]

[a] Department of Nuclear Medicine, University Hospital, LMU Munich, 81377 Munich, Germany
[b] German Cancer Consortium (DKTK), Partner Site Munich, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[c] Bavarian Cancer Research Center (BZKF), 91054 Erlangen, Germany
[d] Department of Radiology, University Hospital, LMU Munich, 81377 Munich, Germany
[e] Department of Psychiatry and Psychotherapy, University of Regensburg, 93053 Regensburg, Germany
[f] German Center for Neurodegenerative Diseases (DZNE), 81377 Munich, Germany
[g] Munich Cluster for Systems Neurology (SyNergy), 81377 Munich, Germany
[h] Department of Neurosurgery, University Hospital, LMU Munich, 81377 Munich, Germany

Received 20 September 2022; accepted 22 December 2022

Abstract

Introduction: Large datasets are required to ensure reliable non-invasive glioma assessment with radiomics-based machine learning methods. This can often only be achieved by pooling images from different centers. Moreover, trained models should perform with high accuracy when applied to data from different centers. In this study, the impact of reconstruction settings and segmentation methods on radiomic features derived from amino acid and TSPO PET images of glioma patients was examined. Additionally, the ability to model and thus reduce feature differences was investigated.
Methods: [^{18}F]FET and [^{18}F]GE-180 PET data were acquired from 19 glioma patients. For each acquisition, 10 reconstruction settings and 9 segmentation methods were included to emulate multicentric data. Statistical robustness measures were calculated before and after ComBat harmonization. Differences between features due to setting variations were assessed using Friedman test, coefficient of variation (CV) and inter-rater reliability measures, including intraclass and Spearman's rank correlation coefficients and Fleiss' Kappa.
Results: According to Friedman analyses, most features (>60%) showed significant differences. Yet, CV and inter-rater reliability measures indicated higher robustness. ComBat resulted in almost complete harmonization (>87%) according to Friedman test and little to no improvement according to CV and inter-rater reliability measures. [^{18}F]GE-180 features were more sensitive to reconstruction settings than [^{18}F]FET features.
Conclusions: According to Friedman test, feature distributions could be successfully aligned using ComBat. However, depending on settings, changes in patient ranks were observed for some features and could not be eliminated by harmonization. Thus, for clinical utilization it is recommended to exclude affected features.

Keywords: Radiomics; Robustness; Data Pooling; FET PET; TSPO PET; Glioma

Abbreviations: CT, computed tomography; CV, coefficient of variation; FBP, filtered back-projection; FWHM, full width half maximum; ICC, intraclass correlation coefficient; IDH, isocitrate dehydrogenase; MRI, magnetic resonance imaging; OSEM, ordered subsets expectation maximization; p.i., post injection; PET, positron emission tomography; TBR, tumor-to-background ratio

* Corresponding author: Lena Kaiser, Dr. rer. nat., Department of Nuclear Medicine, University Hospital, LMU Munich, Marchioninistr. 15, 81377 Munich, Germany.

E-mail addresses: adrian.zounek@med.uni-muenchen.de (A. J. Zounek), nathalie.albert@med.uni-muenchen.de (N. L. Albert), adrien.holzgreve@med.uni-muenchen.de (A. Holzgreve), marcus.unterrainer@med.uni-muenchen.de (M. Unterrainer), julia.brosch-lenz@med.uni-muenchen.de (J. Brosch-Lenz), simon.lindner@med.uni-muenchen.de (S. Lindner), andreas.bollenbacher@med.uni-muenchen.de (A. Bollenbacher), guido.boening@med.uni-muenchen.de (G. Boening), rainer.rupprecht@medbo.de (R. Rupprecht), matthias.brendel@med.uni-muenchen.de (M. Brendel), louisa.vonbaumgarten@med.uni-muenchen.de (L. von Baumgarten), Joerg.Christian.Tonn@med.uni-muenchen.de (J. -C. Tonn), peter.bartenstein@med.uni-muenchen.de (P. Bartenstein), sibylle.ziegler@med.uni-muenchen.de (S. Ziegler), Lena.Kaiser@med.uni-muenchen.de (L. Kaiser).

Z Med Phys **33** (2023) 91–102
https://doi.org/10.1016/j.zemedi.2022.12.005
www.elsevier.com/locate/zemedi

1 Introduction

The most common type of primary malignant brain tumor is glioma with an overall incidence of approx. 6 per 100,000 persons. Glioblastoma, the most aggressive subtype of glioma, has a 5-year relative survival of only 7% and represents 49% of all malignant central nervous system tumors [1]. This dismal outcome underlines the urgent need for improved diagnosis, patient stratification and, consequently, improved treatment planning. Hence, many studies aim to improve the clinical performance of simple image statistics combined with clinical parameters by further including multi-modal and texture information and using machine or deep learning methods [2–4].

Several radiomic studies were performed using magnetic resonance imaging (MRI) data, which offer excellent spatial resolution and soft tissue contrast but lack specificity for tumor tissue. Therefore, positron emission tomography (PET) using amino acid radiotracers, which show increased uptake in neoplastic tissue, is now widely used [5,6] and several related studies have shown the added value of radiomic analyses for patient survival [7], tumor classification [8–10], and identification of tumor recurrence and early tumor progression [11,12]. Recently, the overexpression of the 18-kDa translocator protein (TSPO) in neoplastic tissue in addition to activated glial cells has also attracted attention as a novel imaging marker for assessing glioma microenvironment [13,14].

To properly translate radiomic models into clinical routine, they should be validated on large datasets that preferably include data from multiple centers and thus improve reproducibility and generalizability of radiomics analyses. However, reports have shown that features are sensitive to variations of several factors, including image acquisition, image reconstruction, tumor segmentation, as well as test-retest imaging [15–22]. Thus, it is essential to ensure the reproducibility and robustness of features in this regard. Several methods for removing unwanted variations have been introduced and tested. These so-called harmonization techniques aim to integrate data originating from different centers while preserving clinically relevant information [23]. The ComBat method outperformed other data adjustment methods [24] and was previously validated on radiomic features extracted from PET, MRI, and computed tomography (CT) images of cancer patients and phantoms [21,25–27]. Several statistical measures have been used in previous publications to assess the robustness of features. Orlhac et al. [25–27] used Friedman test and the equivalent Wilcoxon test to validate the ComBat harmonization method. Differences between scanners or reconstruction algorithms and test-retest variability have been assessed using either coefficient of variation (CV) [15,17], intra-class correlation coefficient (ICC) [16,19–22], or Spearman's correlation coefficient [16,22]. Since each of these measures reflects different properties of the data, their relevance may depend on the specific application. Thus, in this work, the robustness of radiomic features was analyzed by including all statistical measures applied in either of the aforementioned publications.

The main goal of this study was to assess whether radiomic feature harmonization is feasible for pooling amino acid or TSPO PET images of glioma patients. To achieve this, radiomic features were evaluated with respect to variations in image processing as encountered in multicentric studies, where data pooling is required for improved generalizability of clinical models. Furthermore, the effectiveness of ComBat feature harmonization was assessed for this specific application. Variations arising from multicentric data were emulated by reconstructing each patient dataset with different settings and applying multiple segmentation methods. To the best of our knowledge, these analyses have not been performed so far.

2 Methods

2.1 Patient data and imaging

PET images from a cohort of 19 patients diagnosed with glioma were included in this study. 10 patients were scanned at initial diagnosis before any treatment and 9 patients at tumor recurrence. Histological and molecular genetic classification according to the 2021 WHO guideline for brain tumors [28] revealed 13 glioblastomas, *IDH* wildtype; 4 astrocytoma *IDH* mutant without 1p/19q codeletion; 2 oligodendroglioma, *IDH* mutant, 1p/19q codeleted. All patients have given written informed consent to the data analysis. The study was approved by the local ethics committee (approval number 18-783).

The images were acquired on a Biograph 64 PET/CT scanner (Siemens Healthineers, Erlangen, Germany) at the Department of Nuclear Medicine of the University Hospital, LMU Munich. Immediately before each PET scan, low-dose CT was performed for attenuation correction. Each patient underwent one PET scan after administration of the radiolabeled TSPO ligand (*S*)-*N,N*-diethyl-9-(2-[^{18}F]-fluoroethyl)-5-methoxy-2,3,4,9-tetrahydro-1*H*-carbazole-4-carboxamide ([^{18}F]GE-180) and one PET scan after administration of the amino acid tracer *O*-(2-[^{18}F]-fluoroethyl)-L-tyrosine ([^{18}F]FET) on consecutive days.

[^{18}F]FET was synthesized in a 2-step process by [^{18}F]-fluoroethylation of L- and D-tyrosine as described by Wester et al. [29] and [^{18}F]GE-180 was synthesized using a FASTlab synthesizer with single-use cassettes (GE Healthcare, Chicago, Illinois, USA) [30]. Dynamic acquisitions were obtained in list mode and corrected for scattered and random coincidences, photon attenuation, radionuclide decay and detector dead time during image reconstruction. For both

radiotracers, late tracer uptake was used for radiomics analysis. The respective late static images were derived by averaging motion corrected 10-minute time frames of the dynamic studies. Frame-wise motion correction to an early 0-3 min post-injection (p.i.) image was performed using the PVIEW tool of the PMOD software (version 3.502, PMOD Technologies, Zürich, Switzerland).

For each patient, a 90-minute scan was performed after intravenous bolus injection of 172 ± 11 MBq of [^{18}F]GE-180. The aforementioned 10-minute frames were generated from 60-90 min p.i. acquisition data according to previous research [31–33]. On the following day, a 40-minute scan was carried out after bolus injection of 177 ± 9 MBq of [^{18}F]FET. In this case, 20-40 min p.i. acquisition data were used for the static images following international practice guidelines for glioma imaging with amino acid tracers [6].

2.2 Image reconstruction

Each image was reconstructed 10 times with different settings. One setting was defined as the default and comprised reconstruction parameters that were optimized for clinical quantification of brain PET images at our department [32,34]. For the remaining settings, the respective parameters were fixed to the default setting, while either the reconstruction algorithm, the matrix size, the number of subsets or the filter size were varied individually. The default reconstruction setting was OSEM3D algorithm with 4 iterations, 21 subsets, and 5 mm Gaussian post-reconstruction filter. The default matrix size of 336 × 336 × 109 with a zoom factor of 2 resulted in a voxel size of 1.018 × 1.018 × 2.027 mm^3. An overview of all included settings is given in Table 1.

2.3 Tumor segmentation

The background intensity I_{BG} was defined on the PET images as the mean intensity in a crescent shaped volume manually delineated in a non-affected brain region encompassing both white and grey matter, as recommended in the EANM/EANO/RANO/SNMMI joint practice guidelines for amino acid PET imaging [6] and described by Unterrainer et al. [35]. For comparison of reconstruction settings, volumes-of-interest (VOI) were segmented using the background intensity multiplied by a factor of 1.6 as a threshold for [^{18}F]FET and 1.8 for [^{18}F]GE-180 [34,36]. Semiautomatic segmentation was performed inside of a manually defined confining volume, using initial seeds and the region growing algorithm provided by the simpleITK library (version 2.1.1, [37]) in Python 3.9.

For comparison of feature values derived using different segmentation methods, three different threshold-based segmentation methods were employed each with three different threshold values, resulting in nine segmentation methods (Table 2). These analyses were performed on patient images reconstructed with the default setting. The intensity threshold was either derived using background intensity (I_{BG}), maximal intensity (I_{max}), or contrast ($I_{max} - I_{BG}$). The values of the threshold factors F_{BG}, F_{max}, and F_{cont} defined in Table 2 were chosen using previous literature [34,36,38]. VOIs with less than 18 voxels were considered too small [39]. Thus, the data of 2/19 patients were excluded.

Table 1
Image reconstruction settings for PET data of 19 glioma patients acquired on a Biograph 64 PET/CT system. The bold entries were selected as the default. FBP: filtered back-projection; OSEM: ordered subsets expectation maximization; FWHM: full width half maximum.

Parameter	Values
Algorithm	***OSEM3D***; OSEM2D; TrueX; FBP with 4.9 mm Hann
Matrix	128; 168; *336*
Subsets	8; 16; *21*
Filter	2; 4; *5* mm Gaussian

2.4 Radiomic feature extraction

Initially, PET images were normalized to the background signal by dividing all voxel intensities by I_{BG} to improve inter-patient comparability yielding tumor-to-background ratio (TBR) images. Feature extraction was performed with the Python package PyRadiomics (version 3.0.1, [40]). Voxels were resampled to 2 × 2 × 2 mm^3 with a b-spline interpolator and TBR values were discretized using a fixed bin width as recommended by Leijenaar et al. [22] to preserve quantitative characteristics and improve inter- and intra-patient comparability of radiomic features. In accordance with previous publications, the bin width was set to the interquartile range of TBR values devided by 4, which yields 0.13 [41,42]. Overall, 107 features from the following categories were extracted from each image: first order statistics (n = 18), 3D shape features (n = 14), and texture features (n = 75). Detailed feature definitions, most of which are compliant with the definitions published by the Image Biomarker Standardization Initiative (IBSI, [43]), can be found in the PyRadiomics documentation [40].

2.5 Feature harmonization

The ComBat method is an empirical Bayes framework that was proposed to harmonize data originating from different sites [44]. It assumes that the data are affected by site-specific additive and multiplicative effects. The neuroCombat package (version 0.2.12, [45]) was implemented in Python to apply ComBat for each feature separately with no adjustments for biological covariates assuming non-

Table 2
Image segmentation methods for PET data of 19 glioma patients acquired on a Biograph 64 PET/CT system.

Method	Threshold	Empirical factors F
Background intensity	$F_{BG} \cdot I_{BG}$	1.4; 1.6; 1.8
Maximum intensity	$F_{max} \cdot I_{max}$	0.4; 0.45; 0.5
Contrast	$F_{cont} \cdot (I_{max} - I_{BG}) + I_{BG}$	0.3; 0.35; 0.4

parametric variables. In this study, different radiomic feature distributions resulting from a variation of reconstruction settings and segmentation methods were aligned using ComBat by removing batch effects. ComBat was fitted and applied independently to assimilate the features derived from each of the following subgroups with the respective number of feature distributions given in brackets: all reconstructions (10), algorithm (4), matrix (3), subsets (3), filter (3), all segmentations (9), background (3), maximum (3), and contrast (3) (see Tables 1 and 2).

2.6 Statistical analysis

Each of the statistical measures described below was calculated using the radiomic features of the entire patient cohort before and after ComBat harmonization. The evaluation was performed separately for [^{18}F]FET and [^{18}F]GE-180 data. The different statistical measures allow for a separate quantification of differences between feature distributions, variability of feature values, and changes in patient ranks.

Friedman test was employed using the Python package SciPy (version 1.7.3, [46]) to compare the distributions of feature values with respect to patients, whereby each distribution originated from a different setting. Statistically significant differences between distributions were indicated by p-values less than 0.05, therefore percentages of robust features exhibiting p-values greater than 0.05 were reported.

Mean coefficients of variation (CV) over all patients were calculated in a feature-wise manner with SciPy to characterize the within-patient variance relative to the mean value from different settings. A threshold of 0.1 was used to identify robust features to provide a reference value for comparison with results from previous robustness studies [15,17].

Intraclass correlation coefficients (ICC) were estimated in Python with the Pingouin package (version 0.5.0, [47]) to quantify the within-patient variance relative to the between-patient variance [48]. According to the guideline published by Koo and Li [49], the model for two-way mixed effects, consistency, single rater, and single measurement was selected. Following their recommendation for evaluating reliability without considering the 95% confidence interval of the ICC estimate, features were categorized as robust when their ICC exceeded a value of 0.9.

Differences between patient ranks were quantified by computing pair-wise Spearman's rank correlation coefficients and Fleiss' Kappa using the Python packages SciPy and statsmodels (version 0.14.0, [50]). Whereas Spearman's rank correlation coefficient was calculated directly from the feature values, Fleiss' Kappa was derived from patient ranks. Since the Spearman's correlation coefficient can only be determined for two rankings at a time, the calculation was performed by averaging over all pair-wise coefficients. Thus, Fleiss' Kappa was also computed directly on the patient ranks to include a measure that eliminates the need for averaging. Thresholds for defining robust features were set to 0.9 for Spearman's rank correlation and to 0.4 for Fleiss' Kappa based on previous studies and recommendations [16,22,51].

Furthermore, spaghetti plots were generated for an exemplary feature to visually inspect the influence of setting variations.

3 Results

Fig. 1 shows boxplots of ICC and Spearman's correlation coefficient, and supplementary Fig. S1 shows boxplots of CV and Fleiss' Kappa. The percentages of features with $p > 0.05$ are listed in Table 3 for reconstruction settings and Table 4 for segmentation methods. Percentages for $CV > 0.1$, $ICC < 0.9$, Spearman's correlation coefficient > 0.9, and Fleiss' Kappa > 0.4 are listed in supplementary Tables S2a and S2b (Supplementary Material 2). All percentages reported in this section are for a variation of all settings and/or all feature classes. A complete list of individual values for every feature is provided in Tables S3a-S3d (Supplementary Material 3) for the variation of all settings. All tables contain results with and without ComBat harmonization.

3.1 Robustness of radiomic features

The percentages of robust features according to Friedman test were low for the variation of reconstruction settings ([^{18}F]FET: 2%; [^{18}F]GE-180: 1%) with highest robustness for shape features (14%; 7%) and for changes in matrix size (27%; 33%) and number of subsets (26%; 29%). Less than 1% of first order and texture features were robust to a variation of all reconstruction settings. In contrast, CV and inter-rater reliability measures indicated a moderate to high robustness for variable reconstruction settings (Fig. 1 and S1), whereby texture features presented with lower percent-

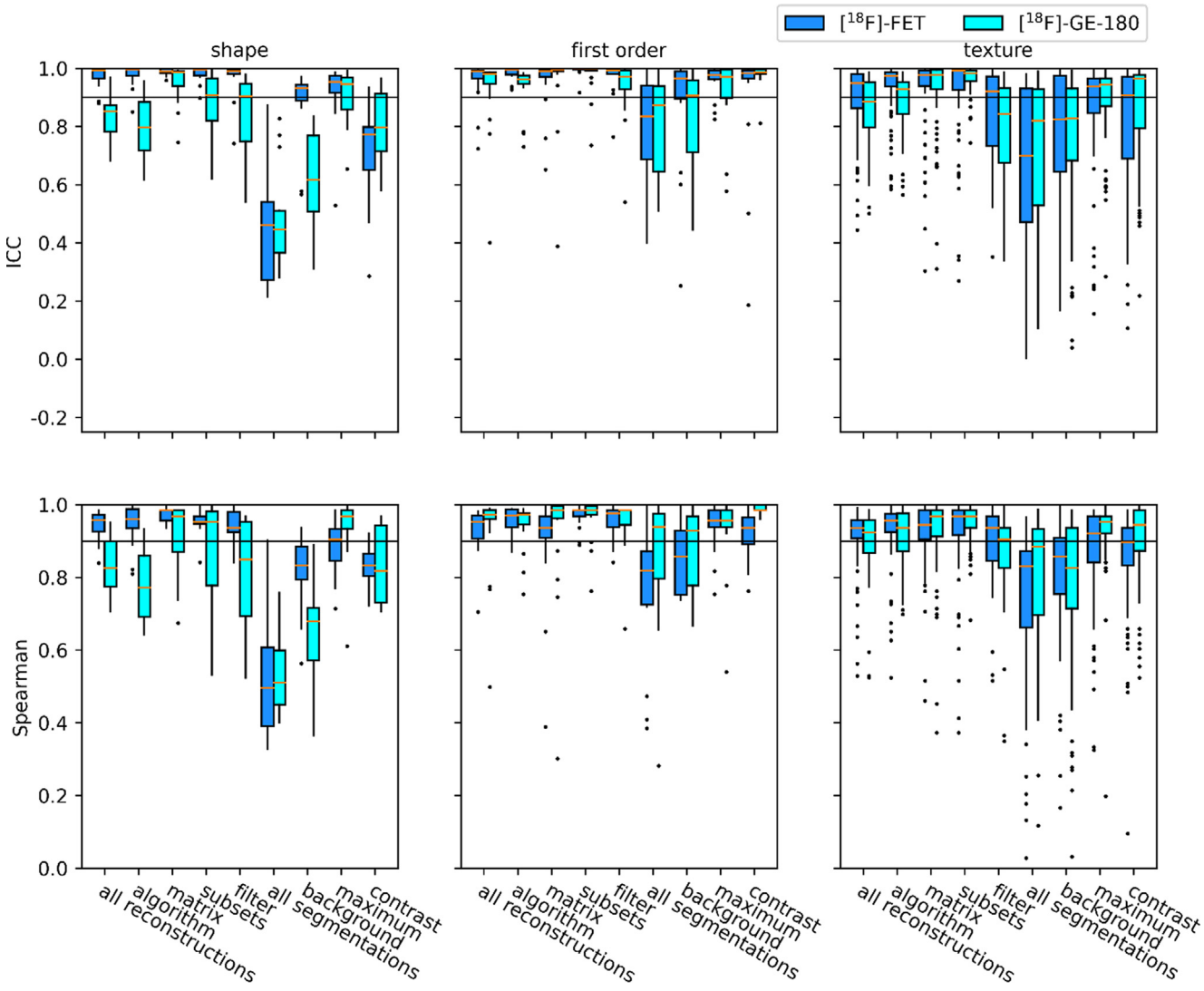

Figure 1. Boxplots showing intraclass correlation coefficients (ICC) and Spearman's rank correlation coefficients of shape, first order and texture features. Statistics were calculated by comparing feature values between all reconstruction settings, all segmentation methods, and subgroups thereof. The horizontal lines indicate the robustness thresholds for the respective measure.

ages compared to shape and first order features. Furthermore, lower robustness was observed for [^{18}F]GE-180 compared to [^{18}F]FET. All statistical measures implied a high sensitivity to the choice of post-reconstruction filter, especially for texture features.

For variation of segmentation methods, the fraction of robust features according to Friedman test was slightly increased but still low ([^{18}F]FET: 22%; [^{18}F]GE-180: 17%), whereby shape features were least robust (7%; 7%). ICC and Spearman's correlation indicated a moderate to low robustness (Fig. 1), with first order features being the most robust and shape features being the least robust. In this case, both measures indicated a lower feature robustness for [^{18}F]GE-180.

3.2 Effect of ComBat harmonization

ComBat feature harmonization enabled an almost perfect assimilation of features as assessed using Friedman test for the variation of both reconstruction settings and segmentation methods ([^{18}F]FET: >89%; [^{18}F]GE-180: >90%). The only exception was a residual sensitivity of several shape features to the variation of reconstruction settings. According to CV and ICC, ComBat caused an overall improvement in robustness, whereas the settings which already presented with very high percentages of features with CV < 0.1 and ICC > 0.9 showed little to no increase. For a variation of reconstruction settings, the percentage of robust features after ComBat harmonization according to CV and ICC

Table 3
Percentages of features without significant differences between reconstruction settings before and after ComBat harmonization. Percentages are color-coded with shades of green ranging from white for low values to dark green for high values. CV: coefficient of variation; ICC: intraclass correlation coefficient.

			All		Shape		First order		Texture	
			Pre	Post	Pre	Post	Pre	Post	Pre	Post
Friedman p > 0.05	[^{18}F]FET	All reconstructions	2%	89%	14%	57%	0%	100%	0%	92%
		Algorithm	10%	89%	14%	64%	11%	100%	9%	91%
		Matrix	27%	88%	36%	43%	17%	100%	28%	93%
		Subsets	26%	92%	14%	64%	17%	100%	31%	95%
		Filter	6%	94%	14%	93%	6%	100%	4%	93%
	[^{18}F]GE-180	All reconstructions	1%	90%	7%	64%	0%	94%	0%	93%
		Algorithm	10%	87%	21%	64%	0%	94%	11%	89%
		Matrix	33%	95%	71%	86%	17%	94%	29%	97%
		Subsets	29%	89%	21%	57%	17%	100%	33%	92%
		Filter	5%	94%	14%	71%	0%	100%	4%	97%

Table 4
Percentages of features without significant differences between segmentation methods before and after ComBat harmonization. Percentages are color-coded with shades of green ranging from white for low values to dark green for high values. CV: coefficient of variation; ICC: intraclass correlation coefficient.

			All		Shape		First order		Texture	
			Pre	Post	Pre	Post	Pre	Post	Pre	Post
Friedman p > 0.05	[^{18}F]FET	All segmentations	22%	94%	7%	100%	28%	100%	24%	92%
		Background	35%	99%	7%	100%	33%	100%	40%	99%
		Maximum	27%	91%	7%	100%	28%	89%	31%	89%
		Contrast	27%	94%	21%	79%	11%	100%	32%	96%
	[^{18}F]GE-180	All segmentations	17%	93%	7%	100%	11%	94%	20%	91%
		Background	22%	91%	7%	79%	22%	89%	25%	93%
		Maximum	35%	98%	7%	100%	39%	100%	39%	97%
		Contrast	31%	85%	14%	86%	17%	94%	37%	83%

remained lowest for texture features. Similarly, the percentages of shape features with CV < 0.1 and ICC > 0.9 remained lowest for variable segmentation methods. Spearman's rank correlation and Fleiss' Kappa were not affected by the ComBat method (see Tables S2a-S2b, Supplementary Material 2).

4 Discussion

A large number of publications report the clinical relevance of radiomic features derived from PET images of glioma patients [3]. Hence, pooling data and applying trained models to data from different centers becomes essential to improve generalizability of models and ultimately enable translation into clinical routine. Therefore, in this study, sensitivity of radiomic features derived from [^{18}F]FET and [^{18}F]GE-180 PET images of glioma patients was quantified with respect to variations in image reconstruction settings and tumor segmentation methods. Since feature robustness has previously been evaluated by different statistical measures, we compared their results and critically assessed their usefulness in judging the success of harmonization.

In previous studies, Friedman test was applied for evaluation of ComBat performance [25–27], whereas CV and ICC

were frequently applied to assess feature variance and inter-rater reliability and have been complemented by Spearman's rank correlation coefficient [15–17,19–22]. Friedman test quantifies significant differences in feature distributions considering the paired nature of feature values of each patient. Since feature harmonization using ComBat allows to improve the correspondence between feature value distributions, this property can be directly evaluated using Friedman test [25–27]. CV and inter-rater reliability measures describe substantially different aspects of feature robustness. CV quantifies the within-patient variance relative to the mean feature value of the patient and ICC relative to the between-patient variance [48]. Spearman's rank correlation coefficient and Fleiss' Kappa quantify whether patients are ranked differently within each setting, which could adversely affect e.g. classification tasks. Since changes in patient ranks cannot be compensated using feature harmonization, variations leading to a low rank correlation need to be avoided.

Overall results showed that PET radiomic features were highly sensitive to the choice of image segmentation methods and, in accordance with the literature for [^{18}F]FDG PET, reconstruction settings [15,17,52]. Rank-based measures implied that a variation of segmentation methods is more likely to change patient ranks with respect to feature values than a variation of reconstruction settings. As this variability cannot be diminished, it is important to first carefully select a clinically meaningful segmentation method and then consistently apply the chosen segmentation method to all patient data.

The high impact of different post-reconstruction filters is most likely explained by the strong effect of smoothing on object boundaries, image texture, as well as voxel intensities in general. This finding contradicts results of a previous study using [^{18}F]FDG for the assessment of lung lesions [15], where the choice of matrix size had the strongest impact on radiomic features as assessed by CV. However, the impact of matrix size might be reduced in this study as the applied radiomics pipeline included resampling to the same voxel size before feature extraction.

The sensitivity difference of shape features between reconstruction and segmentation was expected, as image segmentation directly relates to the shape of a VOI, whereas reconstruction rather affects voxel intensities and their interrelations. Especially in lesions with a more spread-out tracer uptake, VOIs that were generated with different segmentation methods showed significant differences in shape features, as exemplarily seen in Fig. 2.

The lower robustness to a variation of reconstruction settings of features from the [^{18}F]GE-180 data compared to the [^{18}F]FET data might be explained by the potential contribution of low inflammation-related PET signal in [^{18}F]GE-180 images and by the lower activity concentration in healthy background resulting in an increased noise contribution especially when combined with a narrow post-reconstruction filter. This is visualized in Fig. 2 for an example glioma, where in case of a 2 mm Gaussian filter, the tumor volume is rather compact for [^{18}F]FET, while it is broad and patchy for [^{18}F]GE-180. Evidently, the feature extraction process is therefore also dependent on the interplay between reconstruction and segmentation. Hence, the robustness of radiomic features can be influenced by tracer-specific uptake patterns especially when a solely PET-based radiomics workflow including tumor segmentation is used. This implies that the distribution of suspiciously increased biological signal, which is driven by tracer characteristics, may affect the sensitivity of radiomic features to a variation of reconstruction settings or segmentation methods.

As assessed using Friedman test, ComBat harmonization successfully assimilated most radiomic features, which is in line with previous publications validating the ComBat method [25–27]. However, CV and ICC showed only little to no improvements and rank correlation measures were unchanged. Similarly, only little improvement of ICC was observed after Combat harmonization of CT based features as reported by Ligero et al. [21].

The different aspects quantified by statistical measures can be visualized using spaghetti plots as presented in Fig. 3 for the shape feature mesh volume derived from [^{18}F]GE-180 PET data. Feature values of one patient for different settings are connected by lines. For zero variance, a straight horizontal line reflects perfect agreement of feature values from different settings, whereas a jagged line reflects increased variance. Fig. 3 depicts spaghetti plots before and after ComBat harmonization for a variation of reconstruction settings (Fig. 3a, b) and for a variation of segmentation methods (Fig. 3c, d). The mesh volume presents with only few intersecting lines and is therefore less sensitive to a variation of reconstruction settings according to rank correlation measures. Yet, for instance the upper-most blue line in Fig. 3a is slightly jagged and transformed to a straight line after harmonization (Fig. 3b), leading to a slightly decreased CV and satisfactory alignment of distributions according to Friedman test. In contrast, for a variation of segmentation methods, a large number of lines are jagged and intersecting before harmonization, which is reflected by all included statistical measures. Although feature distributions of each segmentation method were well aligned after harmonization, the feature values of each individual patient showed a high variability between settings as quantified using CV and ICC. Furthermore, mesh volume ranks of different patients remained the same as visualized by persisting intersecting lines and quantified using Spearman's rank correlation and Fleiss' Kappa. These observations are in line with the above-mentioned increased sensitivity of shape features to tumor segmentation.

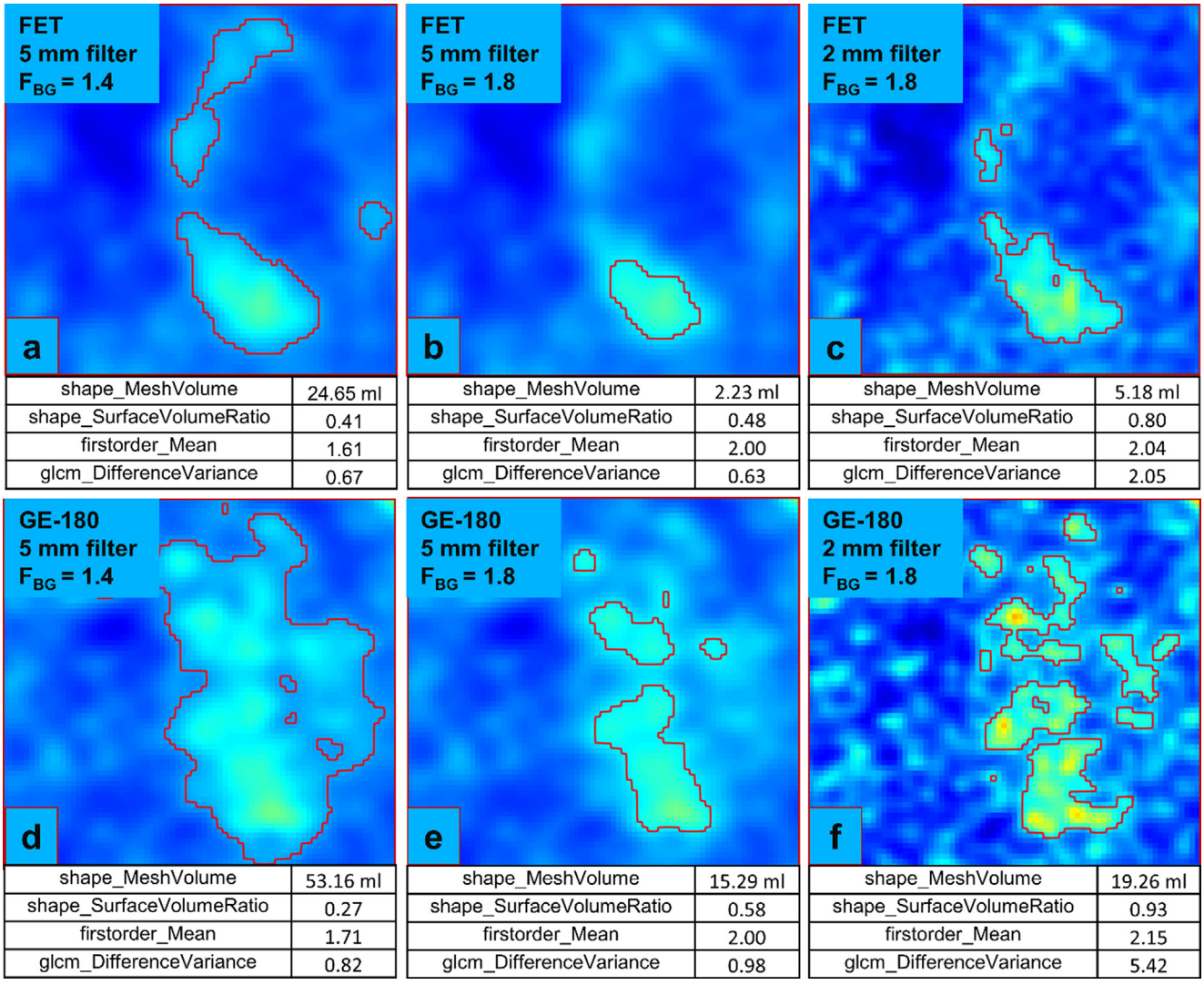

Figure 2. [^{18}F]FET (a-c) and [^{18}F]GE-180 (d-f) PET images showing the same axial slice through a lesion (TBR intensity window: 0-5). Tumor VOIs, marked by red contours, were generated after applying different smoothing filters and segmentation thresholds F_{BG}. The values of four radiomic features are shown for each image.

A limitation of this study is the small sample size, which was restricted due to the large number of included reconstruction protocols (20 per patient for both radiotracers, 380 overall). Results were derived from a mixed patient cohort, which comprises gliomas at initial diagnosis as well as recurrent tumors. To explicitly exclude features which cannot be harmonized using ComBat, the presented analyses should be reperformed with a larger sample size for each specific clinical task and patient group of interest.

One potential caveat concerning ComBat is the occurrence of negative feature values after harmonization for features that can only assume positive values per definition, which was for example observed for the feature mesh volume for one patient (Fig. 3d). Thus, the biological meaning of the harmonized values is uncertain in these cases. In general, it is not clear how well biological variations are retained by ComBat, as ground truth data are usually unavailable to correlate radiomic feature values to the underlying biology. Furthermore, it remains to be evaluated, whether ComBat model fitted to a small patient cohort can be meaningfully applied on data from of a larger or even different patient population. Yet, the striking improvement of feature similarity among different settings in terms of overlapping feature distributions will increase the performance of clinically relevant features unless they are susceptible to patient rank variability. The clinical benefit of feature harmonization for glioma classification, survival prediction, or identification of tumor recurrence will be assessed in a separate study.

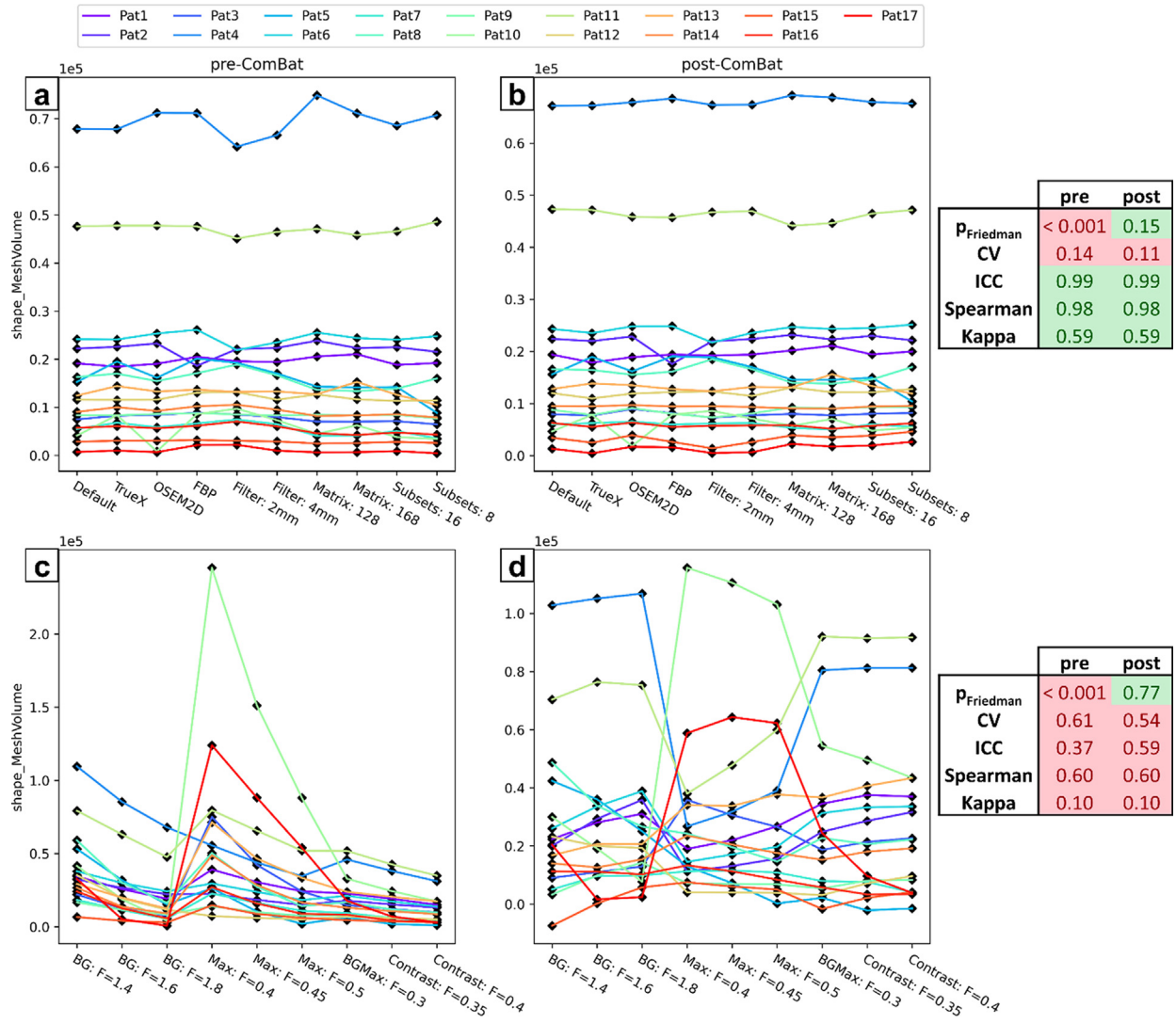

Figure 3. Spaghetti plots showing the values of the shape feature mesh volume after extraction from the [^{18}F]GE-180 data before and after ComBat harmonization. a & b: distributions over all patients for every reconstruction setting. c & d: distributions over all patients for every segmentation method. Results of the statistical analyses are highlighted in green or red depending on the respective robustness threshold ($p_{Friedman}$: 0.05; CV: 0.1; ICC: 0.9; Spearman: 0.9; Kappa: 0.4).

5 Conclusions

In this study, radiomic feature robustness and the applicability of ComBat feature harmonization was assessed with regard to variations that are typically encountered when data from multiple centers are pooled. From the findings it can be concluded that radiomic features derived from [^{18}F]FET or [^{18}F]GE-180 data of glioma patients are highly susceptible to setting variations, whereby [^{18}F]GE-180 features display higher sensitivity to a variation of reconstruction settings compared to [^{18}F]FET. This implies that feature robustness is tracer dependent. Although feature value distributions can be assimilated using ComBat, variable patient ranks cannot be compensated. However, poor agreement between patient ranks may have a significant impact on the biological interpretability and clinical applicability of radiomic features. Hence, multicentric data can be successfully pooled for selected clinically relevant features when ComBat harmonization is employed and rank variability is considered.

Ethics approval and consent to participate

The study was authorized by the local ethics committee (18-783) in accordance with the ICH Guideline for Good

Clinical Practice (GCP) and the Declaration of Helsinki. Written informed consent was obtained from all individual patients included in this study.

Availability of data and material

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to ethical restrictions.

Funding

This project was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) (FOR 2858 project number 421887978 and Research Training Group GRK 2274). M.B. was funded by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy within the framework of the Munich Cluster for Systems Neurology (EXC 2145 SyNergy – ID 390857198).

Author contributions

AJZ, NLA, SZ, LK: Conceptualization; AJZ, SZ and LK: Methodology; AJZ and LK: Formal analysis and investigation; AJZ, AH and LK: Data curation; AJZ and LK: Software; AH, MU, JBL, SL, AB, GB, MB, LvB: Resources; AJZ and LK: Writing – original draft; NLA and SZ: Writing – review and editing; NLA, RR, JCT, PB, SZ: Funding acquisition; NLA, SZ and LK: Supervision. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: N.L.A. is member of the EANM Neuroimaging Committee (NIC). M.B. received speaker honoraria from Roche, GE healthcare and Life Molecular Imaging and is an advisor of Life Molecular Imaging and a member of the EANM NIC. All other authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to thank PD Dr. rer. biol. hum. Michael Lauseker (Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich) for kindly providing his expertise in the field of statistical analysis.

Appendix A Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.zemedi.2022.12.005.

References

[1] Low JT, Ostrom QT, Cioffi G, Neff C, Waite KA, Kruchko C, et al. Primary brain and other central nervous system tumors in the United States (2014–2018): A summary of the CBTRUS statistical report for clinicians. Neuro-Oncol Pract 2022;9:165–182. https://doi.org/10.1093/nop/npac015.

[2] Galldiks N, Zadeh G, Lohmann P. Artificial Intelligence, Radiomics, and Deep Learning in Neuro-Oncology. Neurooncol Adv 2020;2:iv1-iv2. https://doi.org/10.1093/noajnl/vdaa179.

[3] Lohmann P, Meissner AK, Kocher M, Bauer EK, Werner JM, Fink GR, et al. Feature-based PET/MRI radiomics in patients with brain tumors. Neurooncol Adv 2020;2:iv15 -iv21. https://doi.org/10.1093/noajnl/vdaa118.

[4] Singh G, Manjila S, Sakla N, True A, Wardeh AH, Beig N, et al. Radiomics and radiogenomics in gliomas: a contemporary update. Br J Cancer 2021;125:641–657. https://doi.org/10.1038/s41416-021-01387-w.

[5] Albert NL, Weller M, Suchorska B, Galldiks N, Soffietti R, Kim MM, et al. Response Assessment in Neuro-Oncology working group and European Association for Neuro-Oncology recommendations for the clinical use of PET imaging in gliomas. Neuro Oncol 2016;18:1199–1208. https://doi.org/10.1093/neuonc/now058.

[6] Law I, Albert NL, Arbizu J, Boellaard R, Drzezga A, Galldiks N, et al. Joint EANM/EANO/RANO practice guidelines/SNMMI procedure standards for imaging of gliomas using PET with radiolabelled amino acids and [18F]FDG: version 1.0. Eur J Nucl Med Mol Imaging 2019;46:540–557. https://doi.org/10.1007/s00259-018-4207-9.

[7] Pyka T, Gempt J, Hiob D, Ringel F, Schlegel J, Bette S, et al. Textural analysis of pre-therapeutic [18F]-FET-PET and its correlation with tumor grade and patient survival in high-grade gliomas. Eur J Nucl Med Mol Imaging 2016;43:133–141. https://doi.org/10.1007/s00259-015-3140-4.

[8] Lohmann P, Lerche C, Bauer EK, Steger J, Stoffels G, Blau T, et al. Predicting IDH genotype in gliomas using FET PET radiomics. Sci Rep 2018;8:13328. https://doi.org/10.1038/s41598-018-31806-7.

[9] Haubold J, Demircioglu A, Gratz M, Glas M, Wrede K, Sure U, et al. Non-invasive tumor decoding and phenotyping of cerebral gliomas utilizing multiparametric 18F-FET PET-MRI and MR Fingerprinting. Eur J Nucl Med Mol Imaging 2020;47:1435–1445. https://doi.org/10.1007/s00259-019-04602-2.

[10] Li Z, Kaiser L, Holzgreve A, Ruf VC, Suchorska B, Wenter V, et al. Prediction of TERTp-mutation status in IDH-wildtype high-grade gliomas using pre-treatment dynamic [18F]FET PET radiomics. Eur J Nucl Med Mol Imaging 2021;48:4415–4425. https://doi.org/10.1007/s00259-021-05526-6.

[11] Lohmann P, Kocher M, Ceccon G, Bauer EK, Stoffels G, Viswanathan S, et al. Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis. Neuroimage Clin 2018;20:537–542. https://doi.org/10.1016/j.nicl.2018.08.024.

[12] Lohmann P, Elahmadawy MA, Gutsche R, Werner J-M, Bauer EK, Ceccon G, et al. FET PET Radiomics for Differentiating Pseudoprogression from Early Tumor Progression in Glioma Patients Post-Chemoradiation. Cancers 2020;12:3835. https://doi.org/10.3390/cancers12123835.

[13] Zinnhardt B, Roncaroli F, Foray C, Agushi E, Osrah B, Hugon G, et al. Imaging of the glioma microenvironment by TSPO PET. Eur J Nucl Med Mol Imaging 2021;49:174–185. https://doi.org/10.1007/s00259-021-05276-5.

[14] Galldiks N, Langen K-J, Albert NL, Law I, Kim MM, Villanueva-Meyer JE, et al. Investigational PET tracers in neuro-oncology—

What's on the horizon? A report of the PET/RANO group. Neuro Oncol 2022. https://doi.org/10.1093/neuonc/noac131.
[15] 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;56:1667–1673. https://doi.org/10.2967/jnumed.115.156927.
[16] Whybra P, Parkinson C, Foley K, Staffurth J, Spezi E. Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. Sci Rep 2019;9:9649. https://doi.org/10.1038/s41598-019-46030-0.
[17] 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 2019;60:864–872. https://doi.org/10.2967/jnumed.118.217612.
[18] Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J Radiol 2019;20:1124. https://doi.org/10.3348/kjr.2018.0070.
[19] Barry N, Rowshanfarzad P, Francis RJ, Nowak AK, Ebert MA. Repeatability of image features extracted from FET PET in application to post-surgical glioblastoma assessment. Phys Eng Sci Med 2021. https://doi.org/10.1007/s13246-021-01049-4.
[20] Gutsche R, Scheins J, Kocher M, Bousabarah K, Fink GR, Shah NJ, et al. Evaluation of FET PET Radiomics Feature Repeatability in Glioma Patients. Cancers 2021;13:647. https://doi.org/10.3390/cancers13040647.
[21] Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 2021;31:1460–1470. https://doi.org/10.1007/s00330-020-07174-0.
[22] 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. https://doi.org/10.1038/srep11075.
[23] Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol 2020;65:24TR02. https://doi.org/10.1088/1361-6560/aba798.
[24] Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS One 2011;6:e17238.
[25] Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med 2018;59:1321–1328. https://doi.org/10.2967/jnumed.117.199935.
[26] Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology 2019;291:53–59. https://doi.org/10.1148/radiol.2019182023.
[27] Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol 2021;31:2272–2280. https://doi.org/10.1007/s00330-020-07284-9.
[28] Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol 2021;23:1231–1251. https://doi.org/10.1093/neuonc/noab106.
[29] Wester HJ, Herz M, Weber W, Heiss P, Senekowitsch-Schmidtke R, Schwaiger M, et al. Synthesis and Radiopharmacology of O-(2-[¹⁸F]fluoroethyl)-L-Tyrosine for Tumor Imaging. J Nucl Med 1999;40:205–212.
[30] Wickstrøm T, Clarke A, Gausemel I, Horn E, Jørgensen K, Khan I, et al. The development of an automated and GMP compliant FASTlab™ Synthesis of [18F]GE-180; a radiotracer for imaging translocator protein (TSPO). J Label Compd Radiopharm 2014;57:42–48. https://doi.org/10.1002/jlcr.3112.
[31] Feeney C, Scott G, Raffel J, Roberts S, Coello C, Jolly A, et al. Kinetic analysis of the translocator protein positron emission tomography ligand [18F]GE-180 in the human brain. Eur J Nucl Med Mol Imaging 2016;43:2201–2210. https://doi.org/10.1007/s00259-016-3444-z.
[32] Unterrainer M, Fleischmann DF, Vettermann F, Ruf V, Kaiser L, Nelwan D, et al. TSPO PET, tumour grading and molecular genetics in histologically verified glioma: a correlative 18F-GE-180 PET study. Eur J Nucl Med Mol Imaging 2020;47:1368–1380. https://doi.org/10.1007/s00259-019-04491-5.
[33] Vomacka L, Albert NL, Lindner S, Unterrainer M, Mahler C, Brendel M, et al. TSPO imaging using the novel PET ligand [18F]GE-180: quantification approaches in patients with multiple sclerosis. EJNMMI Res 2017;7:89. https://doi.org/10.1186/s13550-017-0340-x.
[34] Kaiser L, Holzgreve A, Quach S, Ingrisch M, Unterrainer M, Dekorsy FJ, et al. Differential Spatial Distribution of TSPO or Amino Acid PET Signal and MRI Contrast Enhancement in Gliomas. Cancers (Basel) 2021;14:53. https://doi.org/10.3390/cancers14010053.
[35] Unterrainer M, Vettermann F, Brendel M, Holzgreve A, Lifschitz M, Zahringer M, et al. Towards standardization of (18)F-FET PET imaging: do we need a consistent method of background activity assessment? EJNMMI Res 2017;7:48. https://doi.org/10.1186/s13550-017-0295-y.
[36] Pauleit D, Floeth F, Hamacher K, Riemenschneider MJ, Reifenberger G, Muller HW, et al. O-(2-[18F]fluoroethyl)-L-tyrosine PET combined with MRI improves the diagnostic assessment of cerebral gliomas. Brain 2005;128:678–687. https://doi.org/10.1093/brain/awh399.
[37] Lowekamp B, Chen D, Ibanez L, Blezek D. The Design of SimpleITK. Frontiers. Neuroinformatics 2013;7. https://doi.org/10.3389/fninf.2013.00045.
[38] Kaiser L. Non-invasive quantification of CNS pathology with dynamic PET information: Investigation of advanced methods for the characterisation of multiple sclerosis and glioma lesions [Dissertation]. LMU Munich 2019.
[39] Vomacka L, Unterrainer M, Holzgreve A, Mille E, Gosewisch A, Brosch J, et al. Voxel-wise analysis of dynamic 18F-FET PET: a novel approach for non-invasive glioma characterisation. EJNMMI Res 2018;8:91. https://doi.org/10.1186/s13550-018-0444-y.
[40] Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104–e107. https://doi.org/10.1158/0008-5472.can-17-0339.
[41] Li Z, Kaiser L, Holzgreve A, Ruf VC, Suchorska B, Wenter V, et al. Prediction of TERTp-mutation status in IDH-wildtype high-grade gliomas using pre-treatment dynamic [(18)F]FET PET radiomics. Eur J Nucl Med Mol Imaging 2021;48:4415–4425. https://doi.org/10.1007/s00259-021-05526-6.
[42] Li Z, Holzgreve A, Unterrainer LM, Ruf VC, Quach S, Bartos LM, et al. Combination of pre-treatment dynamic [(18)F]FET PET radiomics and conventional clinical parameters for the survival stratification in patients with IDH-wildtype glioblastoma. Eur J Nucl Med Mol Imaging 2022. https://doi.org/10.1007/s00259-022-05988-2.
[43] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for

High-Throughput Image-based Phenotyping. Radiology 2020;295:328–338. https://doi.org/10.1148/radiol.2020191145.
[44] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2006;8:118–127. https://doi.org/10.1093/biostatistics/kxj037.
[45] Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 2018;167:104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024.
[46] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–272. https://doi.org/10.1038/s41592-019-0686-2.
[47] Vallat R. Pingouin: statistics in Python. J Open Source Softw 2018;3:1026. https://doi.org/10.21105/joss.01026.
[48] Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–428.
[49] Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012.
[50] Seabold S, Statsmodels PJ. Econometric and Statistical Modeling with Python. Proceedings of the Python in Science Conference: SciPy, 2010.
[51] Fleiss JL, Chilton NW. The measurement of interexaminer agreement on periodontal disease. J Periodontal Res 1983;18:601–606. https://doi.org/10.1111/j.1600-0765.1983.tb00397.x.
[52] Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49:1012–1016. https://doi.org/10.3109/0284186x.2010.498437.