Interchangeability of radiomic features between [18F]-FDG PET/CT and [18F]-FDG PET/MR

Vuong, Diem ; Tanadini-Lang, Stephanie ; Huellner, Martin W ; Veit-Haibach, Patrick ; Unkelbach, Jan ; Andratschke, Nicolaus ; Kraft, Johannes ; Guckenberger, Matthias ; Bogowicz, Marta

Abstract: **PURPOSE** Radiomics is a promising tool for identification of new prognostic biomarkers. However, image reconstruction settings and test-retest variability may influence the absolute values of radiomic features. Unstable radiomic features cannot be used as reliable biomarkers. PET/MR is becoming increasingly available and often replaces PET/CT for different indications. The aim of this study was to quantify to what extent [18F]-FDG PET/CT radiomics models can be transferred to [18F]-FDG PET/MR and thereby to investigate the feasibility of combined PET/CT-PET/MR models. For this purpose, we compared PET radiomic features calculated on PET/MR and PET/CT and on a 4D gated PET/MR dataset to select radiomic features that are robust to attenuation correction differences and test-retest variability, respectively. **METHODS** Two cohorts of patients with lung lesions were studied. In the first cohort (n=10), inhale and exhale phases of a 4D [18F]-FDG PET/MR (4DPETMR) scan were used as a surrogate for a test-retest dataset. In the second cohort (n=9), patients underwent first an [18F]-FDG PET/MR scan (SIGNA PET/MR, GE Healthcare, Waukesha) followed by an [18F]-FDG PET/CT scan (Discovery 690, GE Healthcare) with a delay of 33 min $\pm$ 5 min (PETCT-PETMR). Lesions were segmented on inhale and exhale 4D-PET phases and on the individual PET scans from PET/CT and PET/MR with two semi-automated methods (gradient-based and threshold-based). The scan resolution was 2.73x2.73x3.27 mm and 2.34x2.34x2.78 mm for the PET/CT and PET/MR, respectively. In total, 1355 radiomic features were calculated, i.e. shape (n=18), intensity (n=17), texture (n=136) and wavelet (n=1184). The intra-class correlation coefficient (ICC) was calculated to compare the radiomic features of the 4DPETMR (ICC(1,1)) and PETCT-PETMR (ICC(3,1)) datasets. An ICC>0.9 was considered stable among both types of PET scans. **RESULTS AND CONCLUSION** 4DPETMR showed highest stability for shape, intensity and texture (>80%) and lower stability for wavelet features (40%). Gradient-based method showed higher stability compared to threshold-based method except from shape features. In PETCT-PETMR, more than 61% of shape and intensity features were stable for both segmentation methods. However, a reduced stability was observed for texture (50%) and wavelet (<30%) features. More wavelet features were robust in the smoothed images (low-pass filtering) compared to images with emphasized heterogeneity (high-pass filtering). Comparing stable features of both investigations, highest agreement was found for intensity and lower agreement for shape, texture and wavelet features. Only 53.6% of stable texture features in 4DPETMR were also stable in PETCT-PETMR, and even less in case of wavelet features (40.4%). Approximately 16.9% (texture) and 43.2% (wavelet) of stable PETCT-PETMR features are unstable in 4DPETMR. To conclude, shape and intensity features were robust when comparing two types of [18F]-FDG PET scans (PET/CT and PET/MR). Reduced stability was observed for texture and wavelet features. We identified multiple origins of instability of radiomic features, such as attenuation correction differences, different uptake times and spatial resolution. This needs to be considered when models based on PET/CT are transferred PET/MR models or when combined models are used. This article is protected by copyright. All rights reserved.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: <https://doi.org/10.5167/uzh-167490>
Journal Article
Accepted Version

Originally published at:

Vuong, Diem; Tanadini-Lang, Stephanie; Huellner, Martin W; Veit-Haibach, Patrick; Unkelbach, Jan; Andratschke, Nicolaus; Kraft, Johannes; Guckenberger, Matthias; Bogowicz, Marta (2019). Interchangeability of radiomic features between [18F]-FDG PET/CT and [18F]-FDG PET/MR. *Medical Physics*, 46(4):1677-1685.

DOI: <https://doi.org/10.1002/mp.13422>

Article Type: Research Article

Title:

Interchangeability of radiomic features between [18F]-FDG PET/CT and [18F]-FDG PET/MR

Authors:

Diem Vuong¹, Stephanie Tanadini-Lang¹, Martin W. Huellner², Patrick Veit-Haibach², Jan Unkelbach¹, Nicolaus Andratschke¹, Johannes Kraft¹, Matthias Guckenberger¹, Marta Bogowicz¹

¹Department of Radiation Oncology, University Hospital Zurich and University of Zurich, Zurich, 8091, Switzerland

²Department of Nuclear Medicine, University Hospital Zurich, Zurich, 8091, Switzerland

Corresponding Author:

Name: Diem Vuong,
Department: University Hospital Zurich,
Department of Radiation Oncology,
Address: Rämistrasse 100
8091 Zurich
Switzerland
Email: diem.vuong@usz.ch

Abstract

Purpose: Radiomics is a promising tool for identification of new prognostic biomarkers. However, image reconstruction settings and test-retest variability may influence the absolute values of radiomic features. Unstable radiomic features cannot be used as reliable biomarkers. PET/MR is becoming increasingly available and often replaces PET/CT for different indications. The aim of this study was to quantify to what extent [18F]-FDG PET/CT radiomics models can be transferred to [18F]-FDG PET/MR and thereby to investigate the feasibility of combined PET/CT-PET/MR models. For this purpose, we compared PET radiomic features calculated on PET/MR and PET/CT and on a 4D gated PET/MR dataset to select radiomic features that are robust to attenuation correction differences and test-retest variability, respectively.

Methods: Two cohorts of patients with lung lesions were studied. In the first cohort (n=10), inhale and exhale phases of a 4D [18F]-FDG PET/MR (4DPETMR) scan were used as a surrogate for a test-retest dataset. In the second cohort (n=9), patients underwent first an

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.13422

This article is protected by copyright. All rights reserved.

[18F]-FDG PET/MR scan (SIGNA PET/MR, GE Healthcare, Waukesha) followed by an [18F]-FDG PET/CT scan (Discovery 690, GE Healthcare) with a delay of $33 \text{ min} \pm 5 \text{ min}$ (PETCT-PETMR). Lesions were segmented on inhale and exhale 4D-PET phases and on the individual PET scans from PET/CT and PET/MR with two semi-automated methods (gradient-based and threshold-based). The scan resolution was $2.73 \times 2.73 \times 3.27 \text{ mm}$ and $2.34 \times 2.34 \times 2.78 \text{ mm}$ for the PET/CT and PET/MR, respectively. In total, 1355 radiomic features were calculated, i.e. shape ($n=18$), intensity ($n=17$), texture ($n=136$) and wavelet ($n=1184$). The intra-class correlation coefficient (ICC) was calculated to compare the radiomic features of the 4DPETMR (ICC(1,1)) and PETCT-PETMR (ICC(3,1)) datasets. An $\text{ICC} > 0.9$ was considered stable among both types of PET scans.

Results and Conclusion: 4DPETMR showed highest stability for shape, intensity and texture ($>80\%$) and lower stability for wavelet features (40%). Gradient-based method showed higher stability compared to threshold-based method except from shape features. In PETCT-PETMR, more than 61% of shape and intensity features were stable for both segmentation methods. However, a reduced stability was observed for texture (50%) and wavelet ($<30\%$) features. More wavelet features were robust in the smoothed images (low-pass filtering) compared to images with emphasized heterogeneity (high-pass filtering).

Comparing stable features of both investigations, highest agreement was found for intensity and lower agreement for shape, texture and wavelet features. Only 53.6% of stable texture features in 4DPETMR were also stable in PETCT-PETMR, and even less in case of wavelet features (40.4%). Approximately 16.9% (texture) and 43.2% (wavelet) of stable PETCT-PETMR features are unstable in 4DPETMR.

To conclude, shape and intensity features were robust when comparing two types of [18F]-FDG PET scans (PET/CT and PET/MR). Reduced stability was observed for texture and wavelet features. We identified multiple origins of instability of radiomic features, such as attenuation correction differences, different uptake times and spatial resolution. This needs to be considered when models based on PET/CT are transferred PET/MR models or when combined models are used.

Keywords: robustness, radiomics, PET/CT, PET/MR

Introduction

Due to its non-invasive nature, imaging has gained an increasing role in identification of new prognostic biomarkers to predict treatment outcome ¹. Pre-treatment [18F]-FDG positron emission tomography (PET) scans have been investigated intensively and have shown good prognostic value in multiple oncological diseases ^{2,3}. Numerous studies have proposed and investigated so-called radiomic features as new prognostic biomarkers ⁴⁻⁶. Radiomics refers to the process of extracting a large number of quantitative imaging features that describe the intra-tumoral heterogeneity non-invasively using medical images and were partially shown superior to conventional voxel-based standard-uptake value (SUV) metrics in research settings ⁷. Radiomic features can be categorized into four types: shape, intensity, texture and wavelet. Shape and intensity features have shown good predictive power^{8,9}. However, it is known that tumors are spatial and temporal heterogeneous ¹⁰. Therefore, particularly interesting are texture and wavelet features, which focus not only on the individual voxels, but also the relation between voxels. Texture features describe the relationship of a single voxel with one or more neighboring voxels and distribution of areas with homogenous intensities, which i.e. may serve as a surrogate for intra-tumoral heterogeneity. Wavelet features are filter-based features, which enhance certain characteristics of the image based on its frequency domain information. Fast changing grey level values (associated with edges) have high spatial frequencies and low changing grey level values (smoothed regions) contain low spatial frequencies ⁵.

Biomarkers used for prognostic modelling need to be stable in terms of any variability, excluding the intra-tumoral heterogeneity ¹¹. However, the absolute values of radiomic features can be strongly influenced by different effects, which have been extensively investigated in recent years. For example, it has been shown that some features are affected by respiratory and cardiac motion during the imaging process ^{12,13} or by image reconstruction (e.g., reconstruction algorithms, noise, partial volume effect) ^{12,14}. Further, there is a strong intra-observer variability, which is related tumor segmentation in order to define the region-of-interest (ROI) ¹⁵. And lastly, random effects originated from the acquisition itself may influence the features.

Test-retest studies are established methods to assess the stability of radiomic features from those random effects, where repeated images within an elapsed time period (typically in the range of days) are acquired from the same patient. On those images, features are calculated and their variability among the scans is assessed. Test-retest studies showed high percentage of [18F]-FDG PET radiomic feature to be stable ¹⁴⁻¹⁶ and that their reproducibility

is similar or better compared to the standard SUV metric ^{6,16}. However, test-retest scans are rarely performed in clinical practice since they are cost-sensitive and there is no clinical benefit for the patients. To overcome the lack of test-retest datasets, Larue et al.¹⁷ have investigated the feasibility of different respiratory phases of gated (4D) CT imaging as an alternative to test-retest scans. They showed that a high percentage of radiomic features were stable in both 4D CT and test-retest scans for 20 non-small cell lung cancer (NSCLC) patients (85% unfiltered, 90% wavelet) ¹⁷.

Many stability studies have focused on PET, CT and PET/CT radiomics. However, nowadays, hybrid systems combining PET and magnetic resonance (MR) imaging are becoming increasingly more available and partly replaces PET/CT ¹⁸. MR imaging is an attractive and established imaging modality because it provides superior soft-tissue contrast and does not expose the patient to radiation. One of the largest challenges of PET in PET/MR imaging is the attenuation correction (AC) of photon emission ¹⁹. In PET/CT, this signal correction accounting for the attenuation of the photons in the patient's body can be calculated from the Hounsfield units (HU) derived from CT (CT-AC). Due to the setup of the imaging system, in PET/MR, this signal correction has to be derived from the MR scans (MR-AC) ^{19,20}. However, in contrast to CT, MR images do not represent the attenuation of photons, but proton density and relaxation properties. Hence, nowadays, MR-AC methods aim to relate MR information to CT HU values ^{21,22}. Several AC techniques are available and most of them involve a segmentation step in which this mapping is performed for different tissues, such as air, lung, soft tissue and fat ²³. For example in the PET/MR GE SIGNA system, prior to the PET/MR acquisition an atlas-AC map is generated from the LAVA-Flex T1 weighted images. Pre-defined linear attenuation coefficients of PET acquisition at energy of 511 keV are assigned to tissue categories to obtain a pseudo-CT which serves as an attenuation map for correction of the PET data ²⁴⁻²⁶. Hence, regional tissue changes in the lung as well as inter-patient dependency of the lungs are not accounted leading to high SUV differences ^{19,23,27,28}.

A recent study from Tsujikawa et al.²⁹ investigated the correlation of radiomic features calculated on PET scans from PET/CT and PET/MR for gynaecological as well as oral cavity/oropharyngeal cancer. They concluded, that textural features are less affected by scanner and scan protocols than shape and intensity features. However, their study is limited in number of investigated features (n=8, shape, intensity and texture) and focused on different body entities ²⁹. Since PET/CT based prognostic models have been widely investigated in the past, we want to investigate and quantify to which extend [18F]-FDG PET/CT radiomics models can be transferred to PET/MR imaging, and in particular to which

degree those two image modalities can be combined to a common model for lesion in the lung region. For this purpose, we first investigated two phases of a 4D gated PET/MR as a surrogate for a test-retest dataset according to the abovementioned methodology of Larue et al.¹⁷ to select features robust to intrinsic random effects from the image acquisition. Then, we want to compare PET radiomic features calculated on PET/MR and PET/CT respectively.

Materials and Methods

1. Study population

Similarly to the study of Larue et al.¹⁷, ten patients with lung lesions, who underwent a respiratory-gated PET/MR scan (hereon called 4DPETMR), were selected retrospectively as a surrogate for a test-retest dataset. Seven patients had suspected non-small cell lung cancer (NSCLC) and three patients had primary head and neck cancer, thyroid cancer or lymphoma. An additional dataset of nine patients with proven or suspected NSCLC from a diagnostic comparison trial (2014-2016) between PET/CT and PET/MR was included in our investigation (PETCT-PETMR). Three out of those ten patients were identical to the participants of the 4DPETMR comparison. Both observational studies were approved by the local ethic commission (Kantonale Ethikkommission Zürich) and written consent was obtained from all individual participants. Patients with a lung lesion volume of > 1 cc were enrolled.

2. Image acquisition

In 4DPETMR, a 4D [18F]-FDG-PET/MR scan was acquired with six phases of the respiratory cycle, phase-gated triggered, where first and fourth phase are maximum inhale and exhale (standard non peak inspiration), respectively. Retrospective gating was performed using bellows. [18F]-FDG (203 ± 38 MBq) was injected 66 ± 24 min prior to the acquisition. Maximum inhale and exhale phases were chosen, since the exhale phase is the most stable phase in the respiratory cycle in contrast to the inhale phase where we expect more motion artefacts.

In PETCT-PETMR, first a [18F]-FDG 3D PET/MR scan was acquired followed by a [18F]-FDG 3D PET/CT with a delay time of 33 ± 5 min. [18F]-FDG (220 ± 23 MBq) was injected once for both scans 60 ± 24 min before the first scan. 40 min after the injection, patients were transferred to the PET/MR scanner and the MR scan was acquired within 20 minutes, followed by a PET scan of 10 min. After a transfer time of 15 min, the PET/CT scan was acquired within 12-16 min. Image acquisition was performed with arms up for PET/CT and arms down for PET/MR. SUV were corrected for the injection time by normalizing voxel activity concentration to the injected dose and the patient's body weight.

[Table 1]

3. Segmentation

In 4DPETMR, the largest thoracic lesion was segmented independently in the maximum inhale and exhale phase using MIM VISTA (Version 6.7.4.). In PETCT-PETMR, only the primary tumor was included into the ROI. The segmentation was performed separately on the PET scans from both PET/CT and PET/MR. For both investigations, the segmentation was performed using two semi-automated methods: gradient-based and threshold-based segmentation. Threshold level was set patient-specific and the same level was set for both PET scans, ranging from 16% to 44% and 27% to 41% for 4DPETMR and PETCT-PETMR, respectively. Segmented ROIs were approved by an experienced physician. The respiratory motion of the lesion was determined using the Euclidean distance of the ROI center of mass of inhale and exhale phases.

4. Radiomic features and statistical analysis

Resampling of the images was performed to 3.27 mm using linear interpolation, which is the largest resolution among the studied datasets. Radiomics calculation was performed with an in-house developed radiomics software Z-Rad implemented in Python programming language (Version 2.7.10). The Z-Rad software has been validated according to the image biomarker standardization initiative (IBSI) Zwanenburg et al.³⁰. A fixed bin size of 0.25 SUV was chosen according to IBSI (**Supplement B**). In total, 1355 radiomic features were calculated, i.e. shape ($n = 18$), intensity ($n = 17$), texture ($n = 137$) and wavelets ($n = 1184$) (**Supplement A**). Statistical analysis was performed using R (Version 3.3.2). To compare the radiomic features, the intra-class correlation coefficient (ICC) was calculated. An one-way random (ICC(1,1)) and two-way mixed single model (ICC(3,1)) for the variance estimates were used for 4DPETMR and PETCT-PETMR, respectively³¹. In formulas:

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (\kappa - 1)WMS}$$

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (\kappa - 1)EMS}$$

where BMS and WMS are the between and within subject mean squares, respectively. EMS the residual

mean squares from Friedman's two-way ANOVA. The variable κ denotes the number of repeated measurements (here, PET scans) and the number of observers (here, segmentation methods)³¹. For both investigations, an ICC larger than 0.9 was considered stable among both types of PET scans. Despite the small patient cohort, the high

acceptance level of the ICC keeps type I and type II errors small (0.05 and 0.20, respectively) ³².

Results

1. Segmentation

[TABLE 2]

The SUVmax were on average higher for PET scans from PET/CT compared to those from PET/MR (15.6 and 10.37, respectively) (**Table 1**). The median relative volume difference for both investigations was relatively small (less than 11.3%). The median volume differences were overall larger in threshold-based segmented ROI than in gradient-based segmentation for both investigations. Moreover, there was a broader range in volume difference. Overall, the PETCT-PETMR dataset showed higher variability in segmented volumes than the 4DPETMR dataset (**Table 2**).

2. Stable features of 4DPETMR and PETCT-PETMR

[Figure 1]

4DPETMR: The respiratory motion of the tumors was found to be moderate. The median Euclidean distance was 3.3 mm (ranged from 0.8 to 4.5 mm) for gradient-based segmentation. The percentages of stable features showed that more than 80% of intensity and texture features yielded an ICC > 0.9 between the scans for both segmentation methods. However, only 55.6% of shape and 39.3% of the wavelet features reached this criterion (for the gradient-based method and even less for the threshold-based method) (**Figure 1**).

PETCT-PETMR: In PETCT-PETMR, a larger number of radiomic features was stable when segmentation was performed using the gradient-based compared to threshold-based method, which is in agreement with the improved reproducibility of volume using gradient-based method. More than 61% of shape and intensity features reached an ICC > 0.9 between the scans for both segmentation methods. However, only approximately half of the texture features and 28.0% of the wavelet features reached this criterion (for gradient-based and even less for threshold-based method) (**Figure 1**).

[Figure 2]

Texture and wavelet feature analysis among the respective subtypes showed that the texture features based on GLSZM has the largest ICC range (**Figure 2a**). Further, radiomic features based on LLL wavelet showed higher median stability compared to HHH wavelet features (**Figure 2b**). Additional analysis using fixed number of bins ($n = 16, 32, 64$) has been performed and can be viewed in the supplement B. It was observed that the discretization method affects the stability of the features in the PETCT-PETMR comparison, however no clear trend can be observed.

3. Shared stable radiomic features between 4DPETMR and PETCT-PETMR

[Figure 3]

Comparing stable features of both investigations, a clear difference can be observed between different feature types (**Figure 3**). Highest agreement was observed for intensity features. Only two intensity features were solely stable in 4DPETMR. Likewise, shape features showed a good agreement, with half of the features being stable in 4DPETMR and PETCT-PETMR. Texture and wavelet features showed lower agreement among stable features. Only 53.6% of stable texture features in 4DPETMR were also stable in PETCT-PETMR and even less in case of wavelet features (40.4%). Approximately 16.9% (texture) and 43.6% (wavelet) of stable PETCT-PETMR features were unstable in 4DPETMR. A majority ($> 95\%$) of those features were calculated on low-pass filtered images (**Supplement A**).

Discussion

PET/MR is gaining ground in clinical medicine, owing to high soft-tissue contrast and lack of additional radiation dose from MR. Established PET/CT radiomic models have proven to be prognostic for diseases in the lung region and different outcome parameters. For example, Apostolova et al.⁹ have shown that [18F]-FDG PET based radiomic feature asphericity is predictive for progression-free survival in NSCLC patients and Cook et al.³³ have shown [18F]-FDG PET based texture features (coarseness, contrast, and busyness) in patients with NSCLC is associated with non-response to chemoradiotherapy and with poorer prognosis. Hence, with increased availability of PET/MR scanners, question arises if PET radiomic

features from PET/CT can be transferred to PET/MR models for lesions in the thoracic region.

4DPETMR and PETCT-PETMR

First, we studied the influence of random acquisition effects on the radiomic features. Overall, approximately 40% of the features were stable in the 4DPETMR comparison. Intensity and texture features had the highest stability. Shape features were highly stable in threshold-based segmentation and showed moderate stability (50%) for gradient-based segmentation. Overall, our results are in agreement with the PET test-retest study of Leijenaar et al.¹⁴, which proves the usefulness of 4D PET imaging as surrogate of test-retest scans in case of lung lesions. Additionally, we investigated wavelet features, where a low stability was observed (approximately 40%).

Second, we compared the PET scans of PET/CT and PET/MR, and observed that more than two thirds of the investigated features were unstable. The stability within the feature types showed a reduced but still high rate of shape and intensity features (> 70%) compared to the 4DPETMR dataset (for threshold-based segmentation). However, texture and wavelet features showed lower stability, with less than half of the investigated features being stable. The lowest stability has been observed for features calculated on GLSZM in agreement with others studies which showed that features based on this matrix are more sensitive to acquisition parameters^{8,14,34}. Furthermore, HHH based wavelet features appear to be less stable compared to LLL based wavelet features, where low-pass filters suppress random noise contribution resulting in smoothed ROI.

Feature stability influences in PETCT-PETMR

Several factors could have influenced the absolute value of radiomic features in this comparison. One potential reason for this difference can be the AC difference. However, this comparison can be influenced also by scanner factors affecting PET scan acquired on different machines: e.g. the time lapse between both acquisitions, the resolution of PET scans, the intrinsic system sensitivity of both scanners and even potential differences in patient positioning or lesions location among both scans³⁵.

AC difference: Reliable and robust photon attenuation derived from MR datasets is a field of ongoing investigation in the nuclear medicine and medical physics community. We have observed a difference in SUVmax values between PET/CT and PET/MR datasets. This difference will affect the semi-automated segmentation methods used in our study which leads to a variation of segmented volumes. In fact, threshold-based segmented volumes showed higher variability in volumes compared to gradient-based, since in our study it is based on the SUVmax. This reduces the stability of the shape and intensity features. It has

been reported that SUVmax and / or SUVmean difference was less than 10% in PET corrected by MR-AC compared to the identical PET corrected by CT-AC^{19,23,27} however, considerable variation in SUV in lung tissue was observed^{19,27}. This effect may be compensated for by adjusting the attenuation coefficients.

Time lapse: In our study, both scans were acquired with 33 min time difference. Studies have reported that the [18F]-FDG concentration in tumorous tissue does not reach a plateau within 90 min post-administration³⁶. Hence, the observed difference in SUV max values was also partially caused by the PET/MR acquisition prior to the PET/CT. Ideally PET/CT and PET/MR scans were acquired with no time lapse, however these kind of careful administration of repeated acquisition are performed to optimize repeatability, but are not typical and realistic for clinical conditions^{37,38}. Studies have shown an average time difference of 33 ± 20 min between two uptake periods to better reflect the clinical setting^{38,39}. Lovat et al.⁴⁰ have investigated the effect of PET post-injection scanning time on texture features for neurofibromatosis-1 patients and have observed high agreement in texture features between the scans and no significant change in segmented tumor volume. Notably, this study investigated a different tumor entity and had different uptake times, and hence their results cannot be fully transferred to our study.

Spatial image resolution: Furthermore, variation in SUV can also originate from the difference in spatial image resolution of the compared PET scans. Shafiq et al.⁴¹ have shown that 63 of 213 investigated shape and first-order radiomics were unstable with varying voxel size. Stability of those could be improved using resampling to the same voxel size. To address this uncertainty, the images were resampled to the same spatial resolution. However, resampling alters the SUV⁴² and introduces additional variation among PET image datasets derived from PET/CT and PET/MR.

Intrinsic scanner sensitivity: Moreover, also both scanner types used in our study have a different sensitivity, as was shown by Zeimpekis et al.⁴³

Comparison of 4DPETMR and PETCT-PETMR

Nevertheless, the intersection of shape and intensity features of 4DPETMR and PETCT-PETMR remains high in our study. We have observed 100% agreement for intensity features, indicating that time lapse and AC affected the absolute SUV value but not its distribution within the lesions. Few of the texture were found to be stable in PETCT-PETMR comparison but not stable in 4DPETMR study. In terms of texture feature types, these features are associated to low grey level enhanced features, which are likely to be less sensitive to blurring sourced from the respiratory motion during the long acquisition time in conventional 3D PET images⁴⁴. In case of wavelet features, a poor stability was observed for both 4DPETMR and PETCT-PETMR. For both investigations, more features were stable

in the smoothed images (low-pass filtering) in comparison to images with emphasized heterogeneity (high-pass filtering) (**Supplement A**). This may be explained by the impact of high-pass filters applied on the images, enhancing fast varying grey-levels. Similar to texture features, two thirds of the stable wavelet features from the 4DPETMR study were affected by the differences in AC, time lapse or resolution differences. One third of wavelet features stable in 4DPETMR dataset were also stable in PETCT-PETMR comparison. Approximately 16% of wavelet features stable in PETCT-PETMR were not stable in the 4DPETMR dataset. Most of them were wavelet features calculated on low-pass filtered images, which have a reduced noise impact.

Limitations

In our study, we have shown that a high extent of shape and intensity features are stable among PET/CT and PET/MR and can be thus used interchangeably between PET/CT and PET/MR radiomic models for patients with lung lesions. In wavelet and texture features, only a few selected features are stable among both modalities. However, some limitations of our investigation need to be considered.

First, in our study, features calculated on PET scans of (3D) PET/CT and PET/MR were compared and stable features were set in agreement with a 4D gated PET/MR phase comparison to account for modality-intrinsic random effects from the image acquisition. Oliver et al.⁴⁴ have investigated the impact of respiratory motion on the stability of the 3D and 4D imaging radiomic features. They have compared 3D and 4D PET shape, intensity and first-order radiomics and they showed that features calculated on 3D and 4D PET scans differ. Differences were attributed to both ROI motion and count statistics (noise)⁴⁴. Adams et al.⁴⁵ investigated the factors affecting SUV measurements. They showed SUV varies up to 30% due to an averaging effect of motion during acquisition in 3D PET which causes blurring in the images. This results in higher susceptibility of 4D acquisition to noise. Additionally Oliver et al.⁴⁴ observed that PET features were more stable compared to CT features when comparing 3D and 4D acquisition. Conventional PET scans usually have longer acquisition times that can impact the radiomic feature stability particularly in the lung region, where image acquisition is influenced considerably by respiratory motion. One main conclusion refers to the reduced stability of CT compared to PET due to the higher spatial image resolution and therefore higher impact from respiratory motion. This is a risk of using 4D PET acquisition as a surrogate of test-retest study.

Second, uptake time correction was shown not capable to fully account for the SUV changes over time. Additionally, we assumed the FDG kinetics to be irreversible. To our knowledge, the impact of FDG kinetics on SUV heterogeneity pattern within the tumour has only been investigated by Lovat et al.⁴⁰. However, since they investigated a different tumour entity and

different uptake times, conclusion drawn from their observation are limited. Therefore, further investigations are needed to fully account for the time lapse observed in our study.

Third, partial volume effects are affecting segmenting volumes. Further investigations are needed to explore the influence of image resampling on the partial volume effect.

Fourth, limitation of this study is the small sample size of 10 and 9 patients for 4DPETMR and PETCT-PETMR dataset respectively as well as the small investigated volumes that show to influence the intra-tumoral uptake heterogeneity^{46,47}.

Finally, respiratory motion was assumed to be regular, but irregularities can introduce uncertainties in feature stability.

Recommendation

In this work we have shown a majority of radiomic features calculated on PET from PET/CT and PET/MR to be non-stable and hence, these features should not be used for outcome modelling of combined PET/CT – PET/MR models. Several approaches were investigated to overcome this inherent problem. One approach, used in a majority of radiomics-based models, is to discard all unstable features prior to the development of outcome models. However, as seen in this work, this results in a major number of features to be discarded solely due to their scanner-variability, despite their potential prognostic power as biomarker. Another approach is to post-process the PET scans to harmonize the variability. For example, effects from different pixel sizes can be addressed using resampling combined with additional low-pass filters⁴⁸. However, it has to be ensured that the harmonization addresses the scanner and image acquisition variability only and preserve the patient-variability. In a recent publication, a PET reconstruction harmonization method ComBat, first introduced in genomic studies, was suggested to correct for center-variability⁴⁹. They reported for nine investigated features (three SUV based and six texture based) that tissue dependent variability could be preserved and that they could correct for imaging protocol differences⁴⁹. Nevertheless, as we have shown, uptake time differences in PET acquisition affect the PET scans strongly, which has to be considered for any type (single- or multi-modality) of PET models.

Conclusion

In this study we have investigated the interchangeability of [18F]-FDG PET/CT based radiomic features to [18F]-FDG PET/MR for lung lesions. Despite differences in imaging techniques and influences on the SUV (time lapse, spatial resolution differences and motion), there are radiomic features which showed high stability. Shape and intensity features were stable between both image modalities. However, only a few selected texture

and wavelet features were stable (51.5% and 28%, respectively). Only a reduced number of radiomics features can be used when models from PET/CT are transferred to PET/MR or when combined image modality models for lung lesion are determined. More detailed study to investigate impact of time lapse on stability of the features is needed. Additional investigations should be performed to draw conclusion for other entities.

Acknowledgement

This project was partly supported by Swiss National Science Fonds (SNF) 310030_173303.

Disclosure of Conflicts of Interest

The authors have no relevant conflicts of interest to disclose.

References

1. O'Connor JP, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*. 2017;14(3):169-186.
2. Higashi K, Ueda Y, Arisaka Y, et al. 18F-FDG uptake as a biologic prognostic factor for recurrence in patients with surgically resected non-small cell lung cancer. *J Nucl Med*. 2002;43(1):39-45.
3. Tixier F, Hatt M, Le Rest CC, et al. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*. 2012;53(5):693-700.
4. Chicklore S, Goh V, Siddique M, et al. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. *Eur J Nucl Med Mol Imaging*. 2013;40(1):133-140.
5. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
6. Lee G, Lee HY, Park H, et al. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. *Eur J Radiol*. 2017;86:297-307.
7. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*. 2013;501(7467):338-345.
8. Sollini M, Cozzi L, Antunovic L, Chiti A, Kirienko M. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. *Scientific reports*. 2017;7(1):358.
9. Apostolova I, Rogasch J, Buchert R, et al. Quantitative assessment of the asphericity of pretherapeutic FDG uptake as an independent predictor of outcome in NSCLC. *BMC Cancer*. 2014;14(1):896.
10. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer*. 2012;12(5):323-334.
11. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-73.
12. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234-1248.
13. Grootjans W, Tixier F, van der Vos CS, et al. The Impact of Optimal Respiratory Gating and Image Noise on Evaluation of Intratumor Heterogeneity on 18F-FDG PET Imaging of Lung Cancer. *Journal of Nuclear Medicine*. 2016;57(11):1692-1698.
14. Leijenaar RT, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol*. 2013;52(7):1391-1397.

15. van Velden FH, Kramer GM, Frings V, 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):788-795.
16. Weber WA, Gatsonis CA, Mozley PD, et al. Repeatability of 18F-FDG PET/CT in Advanced Non-Small Cell Lung Cancer: Prospective Assessment in 2 Multicenter Trials. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine.* 2015;56(8):1137-1143.
17. Larue R, Van De Voorde L, van Timmeren JE, et al. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. *Radiother Oncol.* 2017;125(1):147-153.
18. Ehman EC, Johnson GB, Villanueva-Meyer JE, et al. PET/MRI: Where might it replace PET/CT? *J Magn Reson Imaging.* 2017;46(5):1247-1262.
19. Keereman V, Mollet P, Berker Y, Schulz V, Vandenberghe S. Challenges and current methods for attenuation correction in PET/MR. *MAGMA.* 2013;26(1):81-98.
20. Stefaan V, Paul KM. PET-MRI: a review of challenges and solutions in the development of integrated multimodality imaging. *Physics in Medicine & Biology.* 2015;60(4):R115.
21. Hofmann M, Pichler B, Scholkopf B, Beyer T. Towards quantitative PET/MRI: a review of MR-based attenuation correction techniques. *Eur J Nucl Med Mol Imaging.* 2009;36 Suppl 1:S93-104.
22. Wagenknecht G, Kaiser HJ, Mottaghy FM, Herzog H. MRI for attenuation correction in PET: methods and challenges. *MAGMA.* 2013;26(1):99-113.
23. Izquierdo-Garcia D, Sawiak SJ, Knesaurek K, et al. Comparison of MR-based attenuation correction and CT-based attenuation correction of whole-body PET/MR imaging. *Eur J Nucl Med Mol Imaging.* 2014;41(8):1574-1584.
24. Spick C, Herrmann K, Czernin J. 18F-FDG PET/CT and PET/MRI Perform Equally Well in Cancer: Evidence from Studies on More Than 2,300 Patients. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine.* 2016;57(3):420-430.
25. Boellaard R, Quick HH. Current image acquisition options in PET/MR. *Semin Nucl Med.* 2015;45(3):192-200.
26. Sekine T, Ter Voert EE, Warnock G, et al. Clinical Evaluation of Zero-Echo-Time Attenuation Correction for Brain 18F-FDG PET/MRI: Comparison with Atlas Attenuation Correction. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine.* 2016;57(12):1927-1932.
27. Seith F, Schmidt H, Gatidis S, et al. SUV-quantification of physiological lung tissue in an integrated PET/MR-system: Impact of lung density and bone tissue. *PLoS One.* 2017;12(5):e0177856.
28. Keereman V, Holen RV, Mollet P, Vandenberghe S. The effect of errors in segmented attenuation maps on PET quantification. *Med Phys.* 2011;38(11):6010-6019.
29. Tsujikawa T, Tsuyoshi H, Kanno M, et al. Selected PET radiomic features remain the same. *Oncotarget.* 2018;9(29):20734-20746.
30. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative-feature definitions. *arXiv preprint arXiv:161207003.* 2016.
31. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin.* 1979;86(2):420-428.
32. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17(1):101-110.
33. Cook GJ, Yip C, Siddique M, et al. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? *Journal of nuclear medicine : official publication, Society of Nuclear Medicine.* 2013;54(1):19-26.
34. Orlhac F, Soussan M, Maisonneuve JA, et al. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine.* 2014;55(3):414-422.

35. Lee SM, Goo JM, Park CM, et al. Preoperative staging of non-small cell lung cancer: prospective comparison of PET/MR and PET/CT. *European radiology*. 2016;26(11):3850-3857.
36. Hamberg LM, Hunter GJ, Alpert NM, et al. The Dose Uptake Ratio as an Index of Glucose-Metabolism - Useful Parameter or Oversimplification. *Journal of Nuclear Medicine*. 1994;35(8):1308-1312.
37. Kurland BF, Muzi M, Peterson LM, et al. Multicenter Clinical Trials Using 18F-FDG PET to Measure Early Response to Oncologic Therapy: Effects of Injection-to-Acquisition Time Variability on Required Sample Size. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*. 2016;57(2):226-230.
38. Kumar V, Nath K, Berman CG, et al. Variance of SUVs for FDG-PET/CT is greater in clinical practice than under ideal study settings. *Clin Nucl Med*. 2013;38(3):175-182.
39. Hristova I, Boellaard R, Vogel W, et al. Retrospective quality control review of FDG scans in the imaging sub-study of PALETTE EORTC 62072/VEG110727: a randomized, double-blind, placebo-controlled phase III trial. *Eur J Nucl Med Mol Imaging*. 2015;42(6):848-857.
40. Lovat E, Siddique M, Goh V, et al. The effect of post-injection (18)F-FDG PET scanning time on texture analysis of peripheral nerve sheath tumours in neurofibromatosis-1. *EJNMMI Res*. 2017;7(1):35.
41. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44(3):1050-1062.
42. Gonzalez R, Wintz P. Digital image processing. 1977.
43. Zeimpekis KG, Barbosa F, Hullner M, et al. Clinical Evaluation of PET Image Quality as a Function of Acquisition Time in a New TOF-PET/MRI Compared to TOF-PET/CT--Initial Results. *Mol Imaging Biol*. 2015;17(5):735-744.
44. Oliver JA, Budzevich M, Zhang GG, et al. Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. *Transl Oncol*. 2015;8(6):524-534.
45. Adams MC, Turkington TG, Wilson JM, Wong TZ. A Systematic Review of the Factors Affecting Accuracy of SUV Measurements. *American Journal of Roentgenology*. 2010;195(2):310-320.
46. Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*. 2014;55(1):37-42.
47. Hatt M, Majdoub M, Vallieres M, 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. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*. 2015;56(1):38-44.
48. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLOS ONE*. 2017;12(9):e0178524.
49. Orlhac F, Boughdad S, Philippe C, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*. 2018;59(8):1321-1328.

Table 1: Overview of image acquisition characteristics.

Type	PET/CT	PET/MR
Manufacturer	GE Healthcare, Waukesha	
Model	Discovery 690	SIGNA PET/MR
Reconstruction Method	VPFXS	
Attenuation Correction	CT-based	MR-based: Lava Flex pulse sequence
Resolution [mm]	2.73 x 2.73 x 3.27	2.34 x 2.34 x 2.78
SUVmax Median	15.6	10.37
SUVmax Range	3.92 to 44.67	2.35 to 39.87

Table 2: Median relative volume difference (in % and in cc) is shown for both investigations 4DPETMR and PETCT-PETMR for each segmentation method gradient-based and threshold-based. Median relative volume difference for both investigations are small. The PETCT-PETMR dataset shows higher variability in segmented volumes than 4DPETMR dataset.

		4DPETMR		PETCT-PETMR	
		[%]	[cc]	[%]	[cc]
Gradient	Median	7.2	1.5	3.8	1.42
	min	1.5	0.3	0.4	0.13
	max	18.3	8.2	39.9	7.27
Threshold	Median	9.5	1.4	11.3	4.27
	min	0.2	0.01	0.7	0.18
	max	58.8	11.0	71.2	15.98

Figure 1: Bar plots of 4DPETMR and PETCT-PETMR comparison shown for all four feature types, i.e. shape, intensity, texture and wavelet. Results are shown for both segmentation methods gradient-based and threshold-based. For the 4DPETMR dataset, highest stability was observed for shape and intensity features. Worse stability was observed for texture and wavelet features. In case of PETCT-PETMR, most stable features were of type shape and intensity. Worse stability was observed for texture and wavelet features. Gradient-based segmentation shows overall comparable or higher stability compared to the threshold-based method except from shape features.

Figure 2: Distribution intra-class correlation coefficient (ICC) of the PETCT-PETMR comparison among the investigated texture matrices **(a)** and among the wavelet types (combination of high-pass filter H and low-pass filter L) **(b)** for the two different semi-automatic segmentation methods gradient- and threshold-based segmentation. Stability of the features is observed to vary dependent on the segmentation method. Largest ICC range are observed for radiomic feature calculated from GLSZM independent on the segmentation method. In case of the wavelet features, stability of the features is shown to vary dependent on segmentation method, however there is no general trend. Overall, wavelet features calculated on HHH matrix show lowest stability, while highest stability is shown for LLL wavelet features, independent on the segmentation method.

Figure 3: Distribution of shared stable features between PETCT-PETMR comparison and the 4DPETMR comparison for gradient-based method. Good agreement between stable features of PETCT-PETMR and 4DPETMR was observed for shape and intensity features. Texture and wavelet showed only little agreement between the two investigations. Low-pass filtered features tended to have higher stability than high-pass filtered features.