Journal of Nuclear Medicine, published on August 16, 2019 as doi:10.2967/jnumed.119.229724

# Experimental multicenter and multivendor evaluation of PET radiomic features performance using 3D printed phantom inserts

Elisabeth Pfaehler[1], Joyce van Sluis[1], Bram B.J. Merema[2], Peter van Ooijen[3], Ralph C.M. Berendsen[4], Floris H.P. van Velden[5], and Ronald Boellaard[1,6]

[1]Department of Nuclear Medicine and Molecular Imaging, Medical Imaging Center, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [2]Department of Oral and Maxillofacial Surgery, University Medical Center Groningen, Groningen, The Netherlands; [3]Department of Radiology, University Medical Center Groningen, University of Groningen, Groningen,The Netherlands; [4]Department of Medical Physics, Zuyderland Medical Center, Heerlen, The Netherlands; [5]Department of Radiology, Section of Nuclear Medicine, Leiden University Medical Center, Leiden, The Netherlands, and [6]Department of Radiology & Nuclear Medicine, VU University Medical Center, Amsterdam, The Netherlands

Short running title: Multivendor and multicenter study Radiomics

Keywords: $^{18}$F-FDG PET/CT radiomic features, feature harmonization, image reconstruction

Word count: 5868

Financial support: This work is part of the research program STRaTeGy with project number 14929, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This study was financed by the Dutch Cancer Society, POINTING project, grant 10034.

Correspondence to:

Elisabeth Pfaehler

Department of Nuclear Medicine and Molecular Imaging

University Medical Center Groningen, Groningen

Phone: (+31) 503613471 / Fax: (+31) 5036

E-mail: e.a.g.pfaehler@umcg.nl

# ABSTRACT

**Background:** The sensitivity of radiomic features to several confounding factors, such as reconstruction settings, makes clinical use challenging. In order to investigate the impact of harmonized image reconstructions on feature consistency, a multicenter phantom study was performed using 3D printed phantom inserts reflecting realistic tumor shapes and heterogeneity uptakes.

**Methods:** Tumors extracted from real PET/CT scans of patients with Non-Small Cell Lung Cancer served as model for three 3D printed inserts. Different heterogeneity pattern were realized by printing separate compartments that can be filled with different activity solutions. The inserts were placed in the NEMA image quality phantom and scanned various times. First, a list-mode scan was acquired and five statistical equal replicates were reconstructed. Secondly, the phantom was scanned four times on the same scanner. Thirdly, the phantom was scanned on six PET/CT systems. All images were reconstructed using EARL-compliant and locally clinically-preferred reconstructions. EARL-compliant reconstructions were performed without (EARL1) or with (EARL2) point-spread function (PSF). Images were analyzed with and without resampling to 2 mm cubic voxels. Images were discretized with a fixed bin width (FBW) of 0.25 and a fixed bin number (FBN) of 64 bins. The intraclass correlation coefficient (ICC) of each scan setup was calculated and compared across reconstruction settings. An ICC above 0.75 was regarded as high.

**Results:** The percentage of features yielding a high ICC was the largest for the statistical equal replicates (70%- 91% for FBN, 90%-96% for FBW discretization). For scans acquired on the same system, the percentage decreased, but the majority of features still resulted in a high ICC (FBN: 52%-63%, FBW: 75%-85%). The percentage of features yielding a high ICC decreased more in the multicenter setting. In this case, the percentage of features yielding a high ICC was larger for images reconstructed with EARL-compliant reconstructions: e.g. 40% for EARL1, 60% for EARL2 vs. 21% for the clinically-preferred setting for FBW discretization. When discretized with FBW and resampled to isotropic voxels, this benefit was more pronounced.

**Conclusion:** EARL-compliant reconstructions harmonize a wide range of radiomic features. FBW discretization and a sampling to isotropic voxels, pronounces the benefits of EARL-compliant reconstructions.

# INTRODUCTION

Personalized cancer treatment is one of the main promises of modern medicine. Analyzing the combinations of patient genetics and tumor phenotype in medical images can provide additional information on treatment response and diagnosis and has therefore the potential to help in clinical decision making (*1*). One part of this approach is the rapidly growing field "radiomics" that aims to extract a large number of feature values from medical images describing tumor phenotype as well as tumor inter- and intra-heterogeneity (*2*–*4*). In Positron Emission Tomography combined with Computed Tomography (PET/CT) images, radiomics has shown promising results in the assessment of treatment response and patient survival for several cancer types, such as head-and-neck or lung cancer (*5*,*6*).

Besides these positive results, many studies reported on the limitations and challenges of radiomics, including the sensitivity of feature values to differences in reconstruction algorithm, voxel size, smoothing, and discretization method (*7*–*9*). In order to make radiomic studies comparable over patients, institutions, and scanners, it is essential that radiomic features are harmonized across centers. The European Association of Nuclear Medicine (EANM) attempts to reduce this variability of measurements in multicenter clinical trials in its EANM Research Ltd (EARL) accreditation program (*10*). For this purpose, it harmonizes basic standard uptake value (SUV) features based on the $SUV_{max}$, $SUV_{mean}$ and $SUV_{peak}$ by comparing phantom scans of the NEMA NU2-2012 image quality (IQ) phantom. For this purpose, centers choose one reconstruction setting that is in line with the standards provided by EARL and uses an iterative reconstruction algorithm (EARL1). It has been shown that also reconstructions including resolution modeling (based on the Point-Spread-Function (PSF)) can be used to harmonize PET/CT systems (EARL2) (*11*). Additional to the EARL-compliant reconstructions, every center applies usually also one reconstruction with settings leading to optimal lesion detection that is used for clinical reads. As illustrated in Figure 1, the quality of a PET/CT image differs across these three reconstruction settings which have therefore a high impact on the extracted radiomic features (see Table 1).

The EARL harmonization is based on basic SUV features. To the best of our knowledge, there is no multicenter experimental study yet that investigated the effect of EARL harmonization on the variability of complex radiomics features. For this purpose, one object that reflects realistic heterogeneity uptake has to be scanned at multiple centers and the feature values across centers have to be compared. Commercial available phantoms such as e.g. the NEMA image quality phantom are not optimal, as they contain only spherical and homogeneous uptake objects. Therefore, in this study, 3D printed phantom inserts were designed and built according to tumors extracted from typical PET-scans and reflecting more realistic uptake distributions than seen with spheres. These inserts were scanned at three institutions on six different PET/CT systems. Feature values were extracted from EARL-compliant (EARL1 and

EARL2) and local clinically preferred reconstructions. The reliability, repeatability, and reproducibility of radiomic features was reported.

# MATERIALS AND METHODS

## Phantom Design and 3D Printing

Three 3D printed phantom inserts were used in this study. PET scans of patients with Non-Small-Cell-Lung-Cancer (NSCLC) served as models for the inserts. For this purpose, several NSCLC tumors showing various heterogeneity uptake pattern were visually checked. Three tumors with different shapes and uptake characteristics were selected as models for the 3D print. These tumors were segmented, slightly smoothed, scaled, and converted to a stereolithography (STL) file in order to make the printing possible. Differences in heterogeneity uptake were realized by printing two separate compartments that can be filled with different activity solutions. The heterogeneity uptake patterns include a homogeneous tumor (tumor 1), a tumor with heterogeneity uptake in the sagittal view (tumor 2) and a tumor with a necrotic core (tumor 3). The size of the inserts are displayed in Table 2.  The printing was performed by a FormLabs Form 2 printer, which relies on a stereolithography (SLA) technique to cure its photopolymeric Formlabs Clear FLGPCL02 resin (Formlabs Inc., Somerville, Ma, USA). A picture of the 3D inserts as well as the corresponding tumors are displayed in Figure 2. The inserts were placed at equal distances in the NEMA NU-2 IQ phantom. The feature values of the phantom inserts were verified to be within the range of radiomic feature values extracted from 10 FDG PET/CT studies of NSCLC patients (*12*). More than 82% of the features are well within the clinically expected range, while only 1.6% show a large variation from the clinical data. Therefore, the inserts generate feature values that are representative for clinical data.

## Phantom scans

In order to obtain features comparable across institutions and PET/CT-systems, only features that are reliable, repeatable, and reproducible should be used. Whereby reliable features are defined as features yielding only marginal differences when extracted from images obtained under exactly the same conditions. While repeatable features are features that result in small differences when extracted from various scans of the same subject. Reproducibility refers to features that remain almost the same when acquired using different PET/CT systems, image acquisition and reconstruction settings.

To measure reliability, the NEMA IQ phantom containing the inserts was scanned once on a Biograph mCT64 (Siemens Healthcare, Knoxville, TN, USA). The scan was acquired in list-mode and five statistically replicates of 60 s were reconstructed. Three different reconstruction settings were applied: An EARL-compliant reconstruction (EARL1, time-of-flight (TOF) with 5 mm full-width-at-half-maximum (FWHM) Gaussian smoothing), an EARL-compliant reconstruction including PSF (EARL2, PSF+TOF with 5 mm FWHM), and the clinical preferred setting of this institution (PSF+TOF with 7mm FWHM). The homogeneous insert, the outer part of the necrotic core as well as the lower part of the third insert were filled with an activity solution so that a tumor-to-background ratio (TBR) of around 10:1 was achieved. The upper part of the third tumor was filled with an activity solution leading to a TBR of 5:1, while the necrotic core of tumor and spheres were filled with water (see Figure 2). The five statistical equal replicates represent an ideal situation as the five images only differ in noise pattern.

To measure repeatability, the phantom was scanned four times on the same system (Siemens Biograph mCT64) independently. I.e. for every scan the phantom was filled with an activity solution and placed on a slightly different position in the scanner. For differences in phantom filling, the scan duration was adjusted so that statistically equal replicates were obtained. The exact amount of activity in tumors, spheres, and background is listed in Table 3 for each scan. Images were reconstructed using the same reconstruction settings as described above. For every scan the inserts were delineated separately what could lead to slightly different delineations. Therefore, this scenario reflects a more realistic clinical setup.

Furthermore, a multicenter study was performed in order to measure reproducibility. The inserts were scanned at three institutions on six PET/CT systems including four Siemens Healthcare (Biograph mCT40, Biograph mCT64, Horizon with extra ring of detectors (TrueV option), and Biograph Vision), one Philips Healthcare (Vereos) and one GE Healthcare (Discovery MI 4 ring) system. The data were reconstructed with a clinically relevant scan duration of 60 s. The scan duration was adjusted for differences in phantom fillings across centers. **Table 3** lists the phantom fillings for each scan. Also here, images were reconstructed using the scanner defined reconstruction settings complying with the EANM standards (EARL1 and EARL2), as well as the locally clinically preferred settings of each institution. Applied reconstruction algorithm, matrix size, and smoothing kernel of the reconstructed images are listed in **Table 4**. The inserts were segmented separately for each scan.

## PET analysis

Segmentations were performed with an in-house software developed for the analysis and segmentation of PET images.. Segmentations were done manually on the low-dose CT for each scan.

An in-house developed software for the calculation of radiomic features programmed in C++ was used for feature calculation (*13*). All calculated feature values follow the definitions of the Image Biomarker Standardization Initiative (IBSI) and are tested to be in compliance with the available benchmarks (*14*). In total, 436 radiomic features were extracted. Before feature calculation, the images were converted to SUV values so that the phantom background had a mean SUV value of 1. Features were calculated for images consisting of the original voxel size, as well as for images resampled to 2 mm cubic voxels as recommended (*15*). Image and binary segmentation mask were resampled using trilinear interpolation. Before the extraction of textural features, images were discretized using a fixed number of 64 bins (FBN), as well as a fixed bin width (FBW) of 0.25.

## Statistical analysis

Data analysis was performed with Python 3.6.3 using the packages numPy, sciPy, and matplotlib (*16*) for figure plotting. Statistical analysis was performed using R within the python environment with the Python-R interface rPy2.

### *Feature reliability, repeatability, and reproducibility*

In order to measure feature consistency (i.e. reliability, repeatability, and reproducibility) for the three different scan setups, the intraclass correlation coefficient (ICC) was calculated using the irr package (version 0.84), available from the Comprehensive R Archive Network (http://www.r-project.org). A two-way single measure model was used to evaluate the consistency of features for all scans. Every 3D printed insert was regarded as a tumor in a patient, while each scan was regarded as one observer. The ICC is defined as ratio of inter-cluster variability and sum of inter-cluster and intra-cluster variability. Therefore, ICC values vary from zero to one, with one representing perfect agreement. Furthermore, a high ICC implies that the intra-cluster variability is low when compared with the inter-cluster variability indicating that a feature with a high ICC can distinguish well between inserts. An ICC higher than 0.9 is regarded as excellent, values between 0.75 and 0.9, between 0.6 and 0.75, and below 0.6 are regarded as good, moderate, and poor, respectively (*17*).

ICC values were compared between reconstruction settings, discretization methods, and original vs. resampled data using a non-parametric permutation test. A permutation test compares two groups by checking differences in test statistics for the groups. The test swaps randomly the elements of both groups for all possible combinations. If the statistics do not change after swapping, the null hypothesis cannot be rejected. All p-values below 0.01 were considered as statistically significant. A Benjamini-Hochberg procedure with a false discovery rate of 0.25 was performed to diminish the chance of a Type I error for multiple comparisons. The permutation test was performed using the R package perm (version 1.0-0.0) for each feature group separately.

# RESULTS

All calculated radiomic features are listed in supplemental files 1, 2, and 3 (for EARL1, EARL2, and clinical reconstructions, respectively) including their ICCs for each reconstruction setting and discretization method.

Figure 3 displays the percentage of features resulting in an excellent, good, moderate, or bad ICC sorted by feature groups for the statistical equal replicates and both discretization methods. The total percentage of excellent, good, and moderate ICC values was comparable across all reconstruction settings with the highest values for FBW discretization (96.7% for EARL1, 97.4% for EARL2, and 97.9% for the clinically preferred setting vs. 83.2%, 94.2%, and 94.7% for FBN discretization, respectively) (see also supplemental Table 1). The EARL1 setting yielded the lowest percentage of features with an excellent ICC. When comparing the feature groups, the differences in ICC values were only significant for GLRLM features (p-value<0.01). A discretization with FBW resulted in more reliable features than FBN discretization, but the ICC values resulted only in significant differences for GLCM features. Resampling to cubic voxels had almost no influence on reliability, although it led to a slight increase in number of reliable features (see supplemental Figure 1) with no significant differences in ICC values.

In comparison, the percentages of features yielding excellent, good, moderate, or bad ICCs for the four scans acquired on the same system are displayed in Figure 4. The number of features yielding an excellent ICC decreased when compared with the five statistical equal replicates. However, the majority of features still resulted in a good or moderate ICC. Also here, a discretization with FBW led to the highest percentage of features with a moderate or better ICC (87.8% for EARL1, 90.3% for EARL2, and 91.8% for the clinically preferred reconstruction vs. 78.2%, 82.1%, and 77.1% for FBN discretization) with a slight increase after resampling (see supplemental Table 2) with significant differences for GLCM features (p-value<0.01). The differences between clinically preferred and EARL-compliant reconstructions were also not significant, but the clinical preferred reconstruction yielded the highest and the EARL1 setting the lowest percentage of repeatable features. The only feature group that resulted in less repeatable features after resampling were the morphological features (see supplemental Figure 2).

In the multicenter setting, the percentage of features yielding a moderate or better ICC is low when compared with the other scan settings (see Figure 5). Also here, a discretization with FBW led to the largest percentage of features with an ICC higher than 0.6 (71.7% for EARL1, 84.9% for EARL2, 32.3% for the clinically preferred setting vs. 49.3%, 49.5%, and 38% for FBN discretization). Significant differences in ICC values between the two discretization methods were found only for the EARL-compliant reconstructions and some textural feature groups (GLCM and GLRLM features for both EARL-compliant reconstructions, NGLDM and GLSZM for EARL2). When discretized with FBN, only small and non-significant discrepancies can be observed between the reconstruction settings. While for FBW discretization, the difference between EARL-compliant reconstructions and clinical preferred reconstructions led for the majority of textural feature groups to significant differences. It should be noted that in the multicenter setting, the local clinically preferred reconstructions differed substantially between sites and scanners, while this was not the case in the single scanner experiments desribed. Significant differences in ICC values between EARL1 and EARL2 were only observed for GLCM and GLRLM features when discretized with FBW. A resampling to cubic voxels was beneficial, especially for textural feature groups, although the differences were not significant (see supplemental Figure 3). Also here, the only feature group resulting in less reproducible features after resampling was the group of morphological features, where a significant difference was observed (see supplemental Table 3).

# DISCUSSION

To the best of our knowledge, this is the first multicenter and multivendor experimental study that investigates the impact of EARL-compliant reconstructions on the repeatability and reproducibility of radiomic features. Our results suggest that in a multicenter setting, the use of EARL-compliant reconstructions leads to a larger number of reproducible features. A reason for this might be that the clinical preferred reconstructions varied to a large extend in terms of spatial resolution/contrast recovery across PET/CT systems. As radiomic features are sensitive to resolution and image noise, these  variations could be the reason for a higher variation of radiomic features (*18*). This is in line with the fact that differences in feature consistency between reconstruction settings were not visible in the five statistical equal replicates and the four scans acquired on the same scanner where the same local clinically preferred reconstruction was applied.

In the multi-center setting, EARL-compliant images yield comparable image quality. This might be the reason for the low differences in reliability, repeatability and reproducibility of these two reconstruction settings. This result is in line with the findings of Kaalep et *al*. who reported that a harmonization of PET/CT systems using PSF reconstructions is feasible (*11*). Furthermore, our results support the findings of Lasnon et *al*. who showed that images reconstructed with PSF and in line with the EARL-standard can be used for the harmonization of radiomic features (*19*).

While EARL-compliant reconstructions yield similar contrast recoveries, the amount of smoothing for clinically preferred settings differed across PET/CT systems.  The lower spatial resolution with EARL-compliant reconstructions seems to be beneficial in terms of repeatability and reproducibility, but might also eliminate important heterogeneity information which is visible in some of the clinically preferred reconstructions. . This effect is lower in the updated EARL standards (EARL2), which yields higher contrast recoveries and spatial resolution, and is therefore preferred for future multicenter studies. One limitation of this study is that we do not report the accuracy of feature values. As it was demonstrated before that radiomics features are biased as function of  acquisition parameters, image reconstruction settings, and noise (*18*,*20*,*21*), there is an urgent need for standardization of feature values in order to reduce the variability (in bias) of radiomic features across centers. Therefore, we focused on feature consistency and the feasibility of using existing harmonization procedures on the reproducibility of radiomic features**.** Nonetheless, as an high ICC also indicates that features can differentiate well between inserts, our results suggest that EARL-compliant reconstructions also result in more meaningful features, especially when using the EARL2 settings. This is in line with the findings of Aide et *al.* who showed that images reconstructed with higher resolution reconstructions improved the characterization of breast tumors when compared with EARL1 (*22*)**.**

Use of physical phantoms also have limitations, as the 3D printed inserts reflect only three coarse heterogeneity patterns. However, they provide a more realistic scenario than publicly available phantoms **containing** only spheres. Furthermore, phantoms have the advantage to provide a more reproducible setting than patient scans as the activity solution filled in spheres and background can be matched closely across experiments performed in different institutions.

Moreover, our study confirms previous findings (on clinical datasets) such as the impact of image discretization on the reliability and repeatability of radiomic features. Previous studies reported a better repeatability as well as less sensitivity to differences in delineations for FBW discretization (*7*,*10*,*23*). Furthermore, Orlhac et al. demonstrated that a discretization with FBW led to more meaningful features i.e. features that can distinguish well between tumor types (*23*). Our results also confirm the benefit of a discretization with FBW as it resulted in more consistent features especially for EARL-compliant reconstructions.

The impact of voxel size on radiomic feature values has also been studied before (*24*,*25*). Hatt et *al*. recommended the use of isotropic voxels with voxel size of 2 mm (*15*). Our study supports this recommendation. Especially in the multicenter setting, a resampling to cubic voxels led to a better reproducibility of radiomic features. A possible explanation might be that a common voxel size might lead to more comparable features as a large amount of features is sensitive to differences in slice thickness and voxel size (*26*,*27*). The only feature group not benefiting from resampling were the morphological features. This effect was only observed in the scan setups where each scan was segmented separately. A possible reason might be that the resampling of the tumor segmentation might lead to different results depending on the initial position of the delineation in the image.

The impact of tumor delineation on the sensitivity of radiomic features was also reported previously (*7*,*28*,*29*). Our results confirm this finding, as the number of features yielding an excellent ICC decreased from the five statistical equal replicates to the four scans acquired on the same system (with repositioning and thus redefinition of tumor delineation). However, differences in number of features resulting in a moderate or better ICC might also be caused by differences in phantom filling and phantom positioning. Mansor et *al.* demonstrated that basic SUV features

($SUV_{max}$, $SUV_{peak}$, $SUV_{mean}$) are affected by phantom repositioning (*30*), so it is likely that repositioning also affects more complex textural features. However, as a patient repositioning and differences in tumor delineation across institutions are part of the general clinical workflow, it is questionable if features highly sensitive to these changes are feasible to be used for radiomic analysis in the clinic.

In summary, this study shows that PET/CT system performance harmonization by the use of EARL-compliant reconstructions increases the repeatability and reproducibility of a large number of radiomic features. A discretization with FBW and a sampling to 2 mm isotropic voxels have beneficial impact on feature repeatability and reproducibility.

# CONCLUSION

This study reports on the impact of EARL-compliant reconstructions on the reliability, repeatability, and reproducibility of radiomic features in comparison with clinical preferred reconstructions. Our results show that the use of EARL-compliant reconstructions is beneficial and leads to a larger number of reliable, repeatable, and reproducible features. A discretization with FBW and a resampling to cubic 2mm voxels increases the percentage of consistent features. The study suggests that EARL-compliant reconstructions should be used for radiomic analysis, especially in a multicenter setting. Use of the updated EARL2 standards is preferred because of its higher contrast recovery and spatial resolution while providing similar radiomics performance compared to EARL1 standards (*11*).

**Financial support**: This work is part of the research program STRaTeGy with project number 14929, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This study was financed by the Dutch Cancer Society, POINTING project, grant 10034.

**Acknowledgments**: We would like to thank Hinke Schokker and Johan R. de Jong for the help with the phantom scans.

**Disclosure of Conflicts of Interest**: The authors have no relevant conflicts of interest to disclose.

**Ethical approval**: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**KEY POINTS:**

QUESTION: Which reconstruction algorithm leads to the most stable radiomic features in a multi-center and multi-vendor setting?

PERTINENT FINDINGS: Harmonized image reconstructions (EARL-compliant) led to a larger number of reliable, repeatable, and reproducible radiomic features. This effect increased when images were discretized with a fixed bin width and resampled to isotropic voxels before feature extraction.

IMPLICATIONS FOR PATIENT CARE: In order to make radiomic features comparable across multiple centers, multi-center radiomic studies should be performed using harmonized (EARL-compliant) reconstructions, images should be discretized using a fixed bin width and resampled to isotropic voxels.

# References

1. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.

2. Avanzo M, Stancanello J, El I. Beyond imaging : The promise of radiomics. *Phys Medica*. 2017;38:122-139.

3. Lambin P, Rios-velazquez E, Leijenaar R. Radiomics : Extracting more information from medical images using advanced feature analysis. 2012:441-446.

4. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278:563-577.

5. Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Sci Rep*. 2017;7:46349.

6. Parmar C, Leijenaar RTH, Grossmann P, et al. Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer. *Nat Sci Reports*. 2015;5:1-10.

7. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol Imaging Biol*. 2016;18:788-795.

8. Leijenaar RTH, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability. *Acta Oncol (Madr)*. 2013;52:1391-1397.

9. Desseroit M-C, Tixier F, Weber WA, et al. Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non–Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort. *J Nucl Med*. 2017;58:406-411.

10. Leijenaar RTH, Nalbantov G, Carvalho S, 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.

11. Kaalep A, Sera T, Rijnsdorp S, et al. Feasibility of state of the art PET/CT systems performance harmonisation. *Eur J Nucl Med Mol Imaging*. 2018;45:1344-1361.

12. Kolinger GD, Vállez García D, Kramer GM, et al. Repeatability of [18F]FDG PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients. *EJNMMI Res*. 2019;9:14.

13. Pfaehler E, Zwanenburg A, de Jong JR, Boellaard R. RaCaT: An open source and easy to use radiomics calculator tool. Wang Y, ed. *PLoS One*. 2019;14:e0212223.

14. Zwanenburg A, Leger S, Vallières M, Löck S, Initiative  for the IBS. Image biomarker standardisation initiative. 2016.

15. Hatt M, Tixier F, Pierce L, et al. Characterization of PET / CT images using texture analysis : the past , the present … any future ? *Eur J Nucl Med Mol Imaging*. 2017;44:151-165.

16. Oliphant TE. Python for Scientific Computing. *Comput Sci Eng*. 2007;9:10-20.

17. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15:155-163.

18. Pfaehler E, Beukinga RJ, de Jong JR, et al. Repeatability of 18 F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med Phys*. December 2018.

19. Lasnon C, Majdoub M, Lavigne B, et al. 18F-FDG PET/CT heterogeneity quantification through textural features in the era of harmonisation programs: a focus on lung cancer. *Eur J Nucl Med Mol Imaging*.

2016;43:2324-2335.

20. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J Med Imaging*. 2015;2:041002.

21. Yan J, Chu-Shern JL, Loi HY, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. *J Nucl Med*. 2015;56:1667-1673.

22. Aide N, Salomon T, Blanc-Fournier C, Grellard J-M, Levy C, Lasnon C. Implications of reconstruction protocol for histo-biological characterisation of breast cancers using FDG-PET radiomics. *EJNMMI Res*. 2018;8:114.

23. Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. 18F-FDG PET-Derived Textural Indices Reflect Tissue-Specific Uptake Pattern in Non-Small Cell Lung Cancer. Adusumilli PS, ed. *PLoS One*. 2015;10:e0145063.

24. Orlhac F, Nioche C, Soussan M, Buvat I. Understanding Changes in Tumor Texture Indices in PET: A Comparison Between Visual Assessment and Index Values in Simulated and Patient Data. *J Nucl Med*. 2017;58:387-392.

25. Orlhac F, Theze B, Soussan M, Boisgard R, Buvat I. Multiscale Texture Analysis: From 18F-FDG PET Images to Histologic Images. *J Nucl Med*. 2016;57:1823-1828.

26. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol*. 2015;60:5471-5496.

27. 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*. November 2018:jnumed.118.217612.

28. Bashir U, Azad G, Siddique MM, et al. The effects of segmentation algorithms on the measurement of 18F-FDG PET texture parameters in non-small cell lung cancer. *EJNMMI Res*. 2017;7:60.

29. Altazi BA, Zhang GG, Fernandez DC, et al. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. *J Appl Clin Med Phys*. 2017;18:32-48.

30. Mansor S, Pfaehler E, Heijtel D, Lodge MA, Boellaard R, Yaqub M. Impact of PET/CT system, reconstruction protocol, data analysis method, and repositioning on PET/CT precision: An experimental evaluation using an oncology and brain phantom. *Med Phys*. 2017;44:6413-6424.

31. 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. *Sci Rep*. 2017;7:1-15.

# Figures

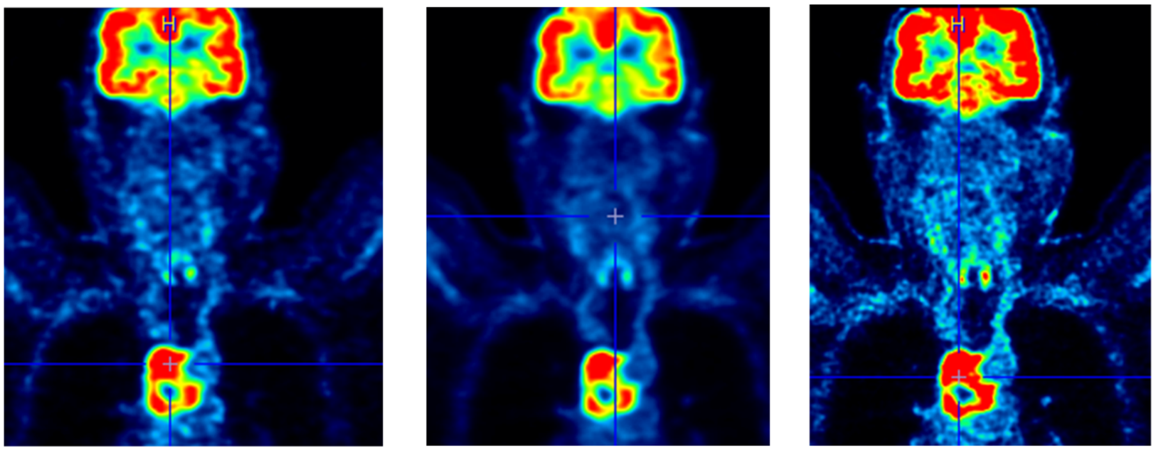

**Figure 1: Different reconstructions of the same patient scan of a patient with Non-Small-Cell-Lung-Cancer (NSCLC) acquired at the Siemens Vision: From left to right: Reconstruction with EARL1, EARL2, and the clinical preferred reconstruction**

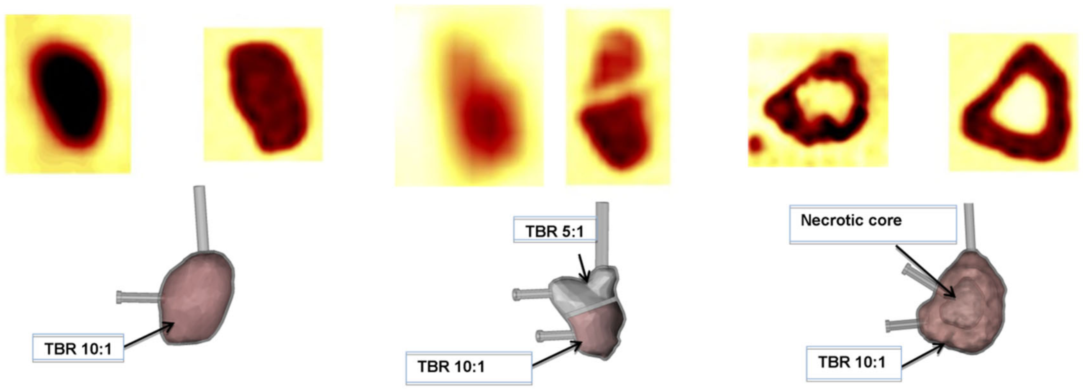

**Figure 2: Upper row: Original tumor in PET/CT image (left) and PET/CT scan of phantom insert (from left to right tumor 1, tumor 2, tumor 3);**

**Lower row: Corresponding STL-images with tumor-to-background ratios**

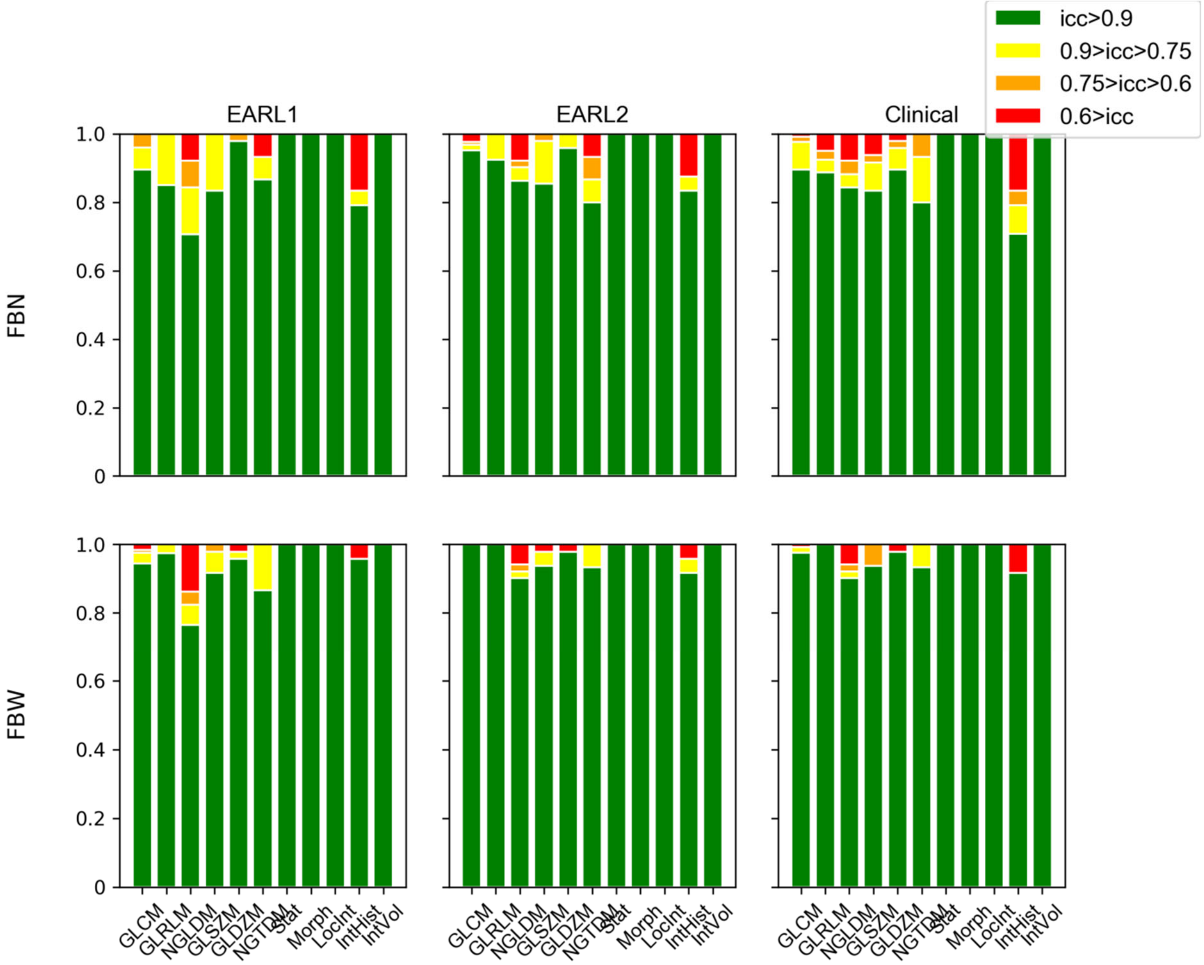

**Figure 3: Percentage of features extracted from the five statistical equal replicates yielding an excellent, good, moderate, or bad ICC for Fixed bin number (FBN) (first row) and Fixed bin width (FBW) discretization (second row) for the different feature groups (GLCM: Grey-Level-Co-occurrence-Matrix, GLRLM: Grey-Level-Run-Length-Matrix, NGLDM: Neighboring-Grey-Level-Dependence-Matrix GLSZM: Grey-Level-Size-Zone-Matrix, GLDZM: Grey-Level-Distance-Zone-Matrix, NGTDM: Neighboring-Grey-Tone-Difference-Matrix).**

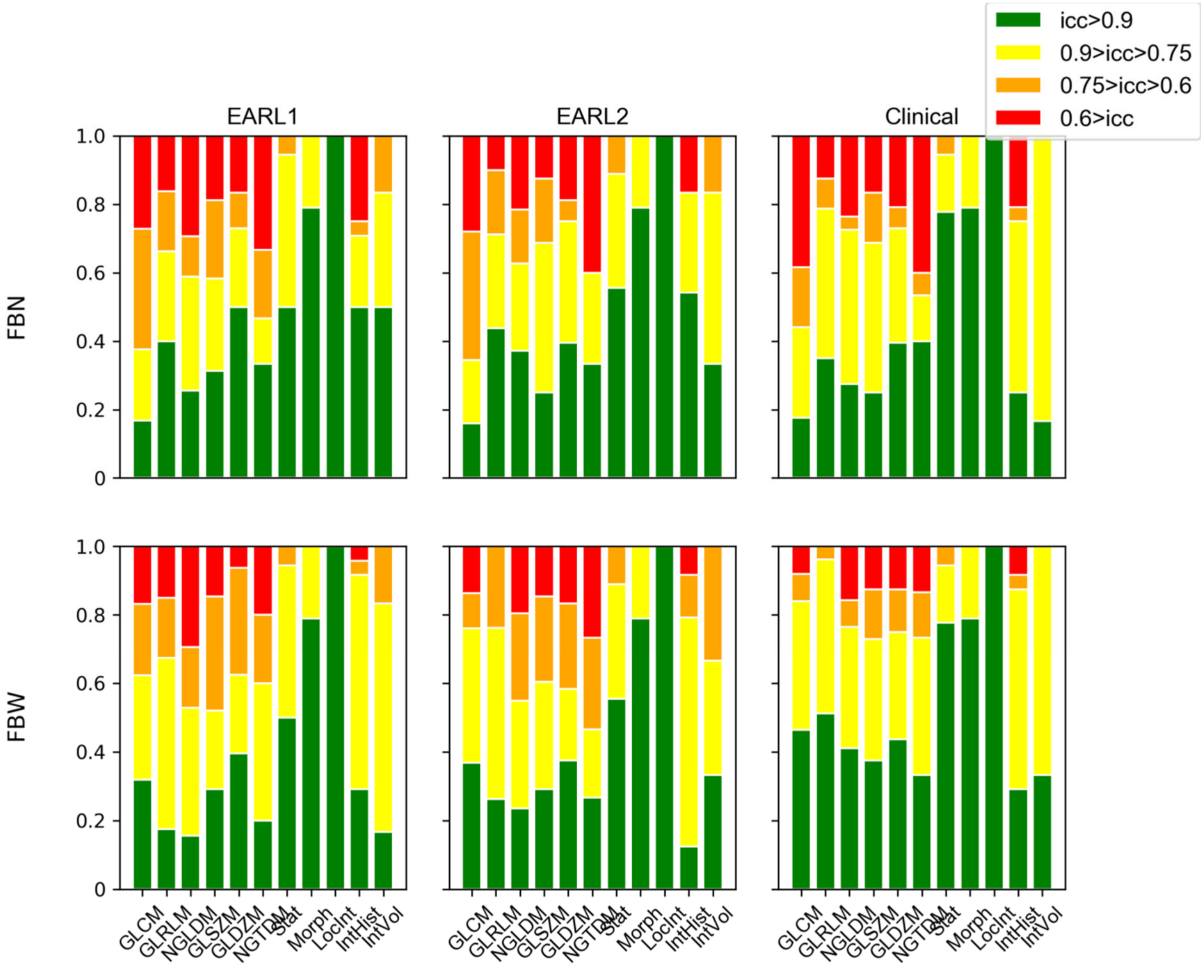

**Figure 4: Percentage of features extracted from the four scans acquired on the same PET/CT system yielding an excellent, good, moderate, or bad ICC for FBN (first row) and FBW discretization (second row)**

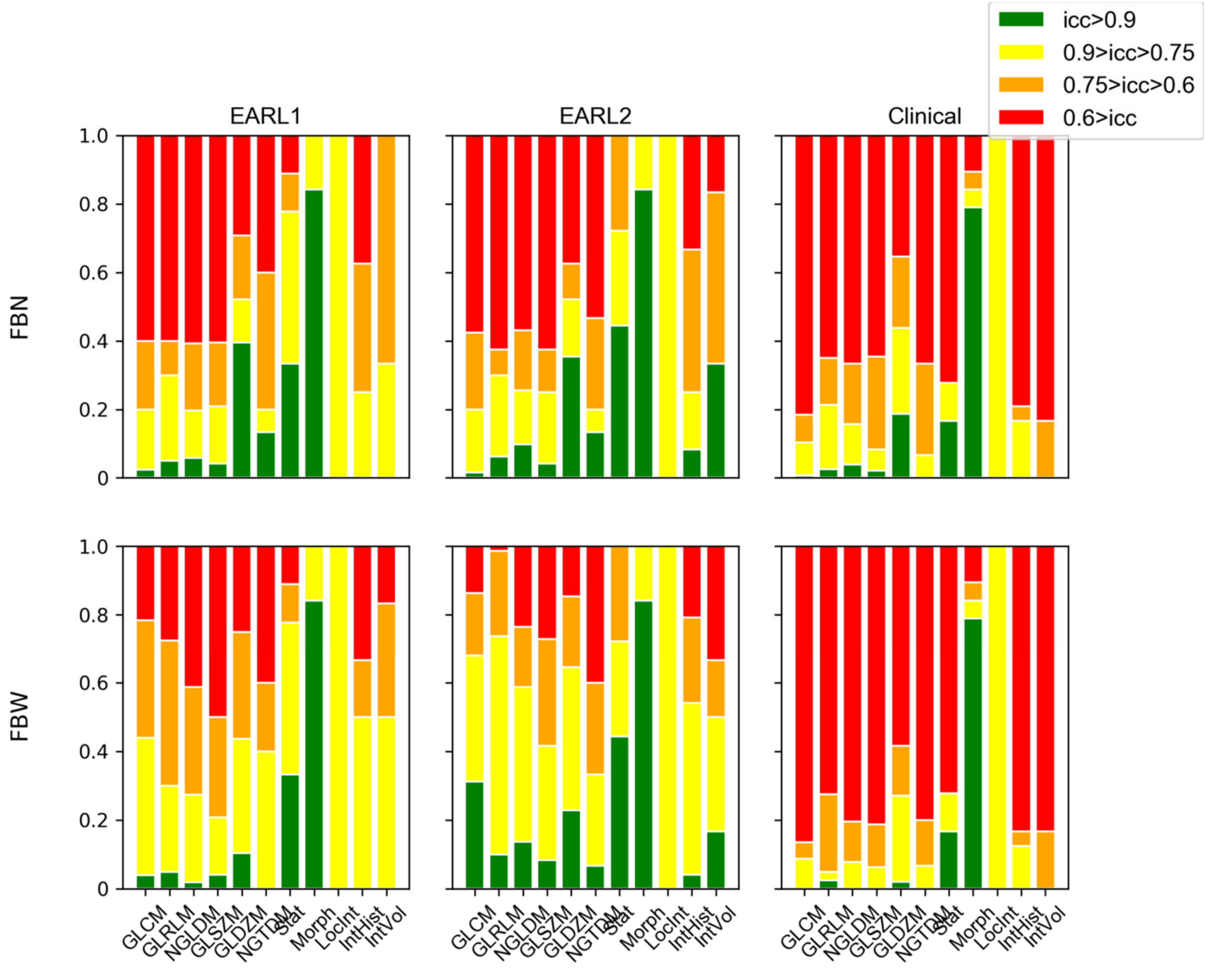

**Figure 5: Percentage of features extracted from the multicenter setting yielding an excellent, good, moderate, or bad ICC for FBN (first row) and FBW discretization (second row)**

# Tables

| | EARL1 | EARL2 | Clinical preferred |
|---|---|---|---|
| High grey level run emphasis$_{3DAVG}$ | 142.24 | 175.07 | 130.10 |
| Busyness$_{3D}$ | 0.34 | 0.30 | 0.50 |
| Contrast$_{2DAVG}$ | 11.21 | 14.07 | 7.64 |

**Table 1: Radiomic features of patient displayed in Figure 1 found to give valuable information about survival in lung cancer patients (*31*) for different reconstruction settings**

|  | Size | Volume |
|---|---|---|
| Tumor 1 | 40.3 mm x 44 mm x 54.5 mm | 46.05 ml |
| Tumor 2 – upper part | 33.9 mm x 37 mm x 30 mm | 10.75 ml |
| Tumor 2 – lower part | 24.3 mm x 40.5 mm x 36.6 mm | 13.12 ml |
| Tumor 3 – outer part | 56 mm x 54 mm x 65.1 mm | 65.35 ml |
| Tumor 3 – necrotic core | 25 mm x 24 mm x 31 mm | 7.8 ml |

**Table 2: Size of 3D printed inserts**

| | Background filling in kBq/ml | Tumor filling in kBq/ml (parts 10:1/5:1) |
|---|---|---|
| Biograph mCT64 Scan 1 | 2.2 | 21.8/15.8 |
| Biograph mCT64 Scan 2 | 2.3 | 22.6/15.5 |
| Biograph mCT64 Scan 4 | 1.9 | 2.1/14.5 |
| Biograph mCT64 Scan 4 (included in multicenter study) | 1.4 | 14.3/9.0 |
| Horizon | 2.2 | 20.0/10.0 |
| Vereos | 1.2 | 12.1/4.6 |
| Biograph mCT40 | 1.9 | 19.4/10.0 |
| Vision | 2.6 | 23.1/11.9 |
| Discovery MI | 1.5 | 14.6/6.9 |

**Table 3: Amount of activity in phantom background and tumor inserts for the four scans acquired on same scanner and the multi-center setting**

|  | EARL1 | EARL2 | Clinical |
|---|---|---|---|
| Horizon | 'TOF', 'M256', '5mm' | 'PSFTOF', 'M256', '5mm' | 'PSFTOF', 'M256', '5mm' |
| Vereos | 'TOF', 'M144', '6mm' | 'PSFTOF', 'M144', '5mm' | 'TOF', 'M144', '4mm' |
| Biograph mCT40 | 'TOF', 'M256', '5mm' | 'PSFTOF', 'M256', '5mm' | 'PSFTOF', 'M256', '7mm' |
| Biograph mCT64 | 'TOF', 'M256', '5mm' | 'PSFTOF', 'M256', '5mm' | 'PSFTOF', 'M256', '7mm' |
| Vision | 'TOF', 'M256', '5mm' | 'PSFTOF', 'M256', '5mm' | 'PSFTOF', 'M256', '**0** mm' |
| Discovery MI | 'TOF', 'M192', '7mm' | ‘VPFXS’, 'M192', '7 mm' | 'VPHD', 'M192', '0mm', |

**Table 4: Applied reconstruction algorithm, matrix size, and smoothing factor for each scanner. The reconstruction settings of the Discovery MI scanner are comparable to: VPFXS is equivalent to a PSF+TOF reconstruction, while VPHD is equivalent to a PSF reconstruction**

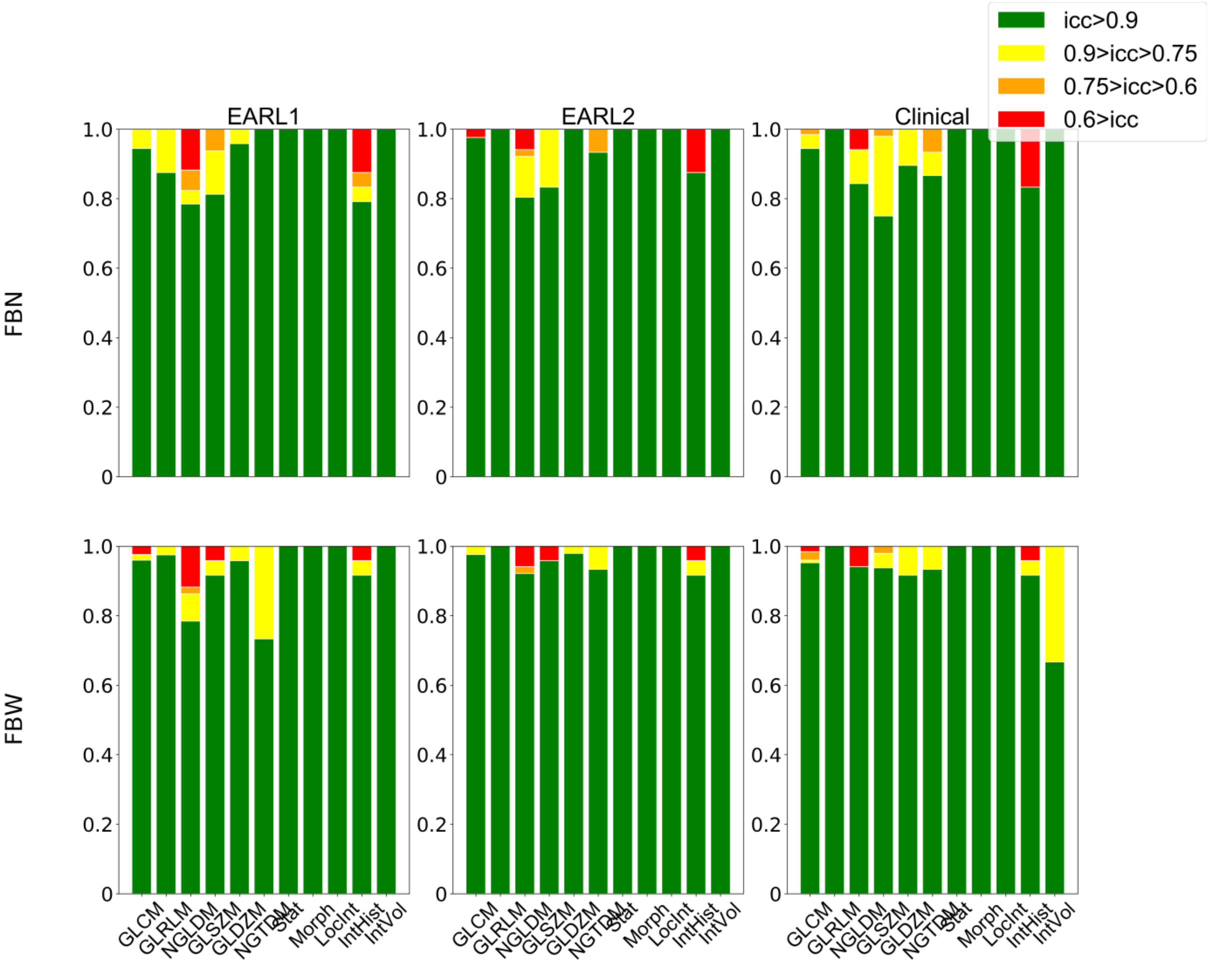

**Figure 1: Percentage of features extracted from the five statistical equal replicates after resampling to cubic voxels yielding an excellent, good, moderate, or bad ICC for FBN (first row) and FBW discretization (second row)**

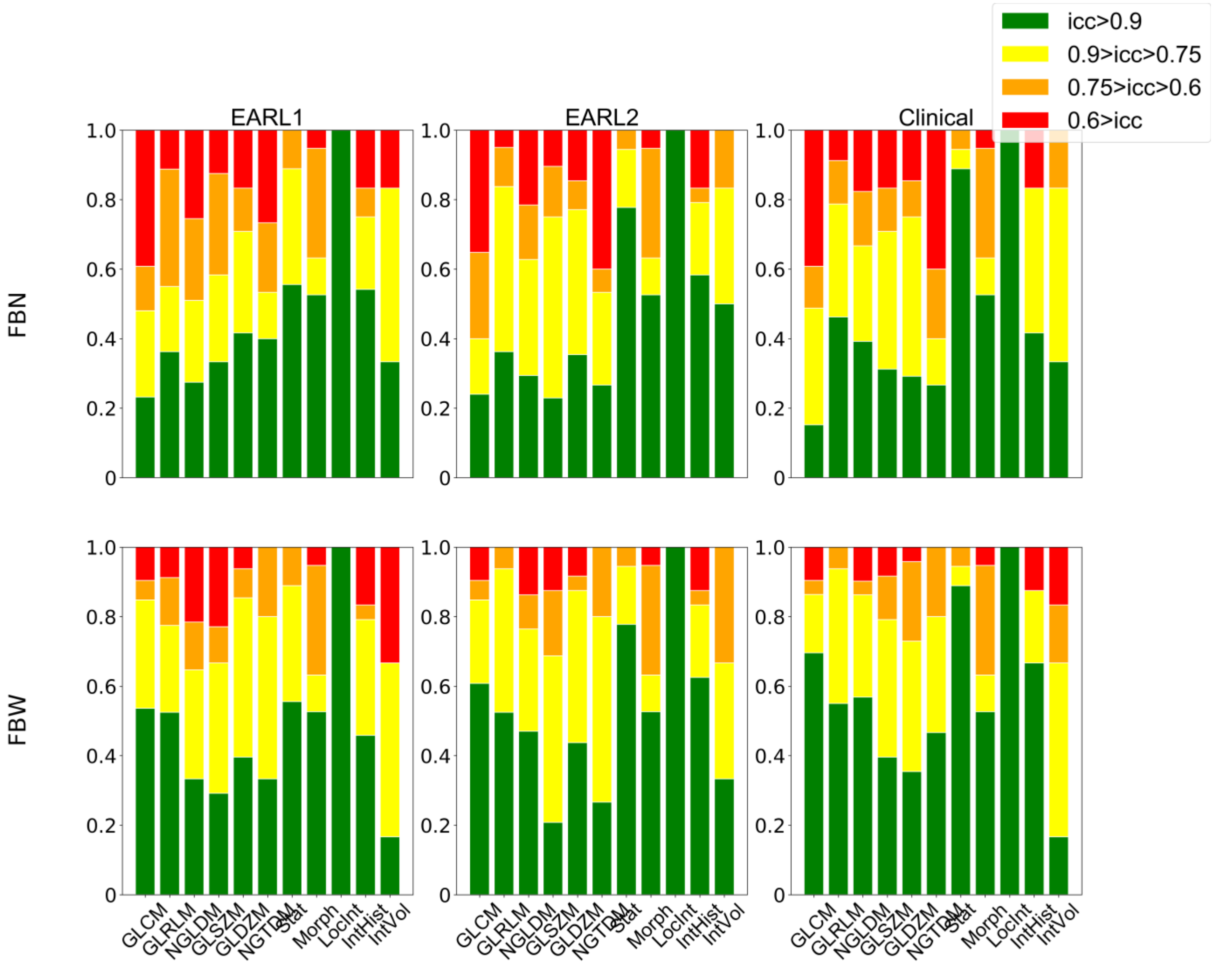

**Figure 2: Percentage of features extracted from the four scans acquired on the same PET/CT system after resampling to cubic voxels yielding an excellent, good, moderate, or bad ICC for FBN (first row) and FBW discretization (second row)**

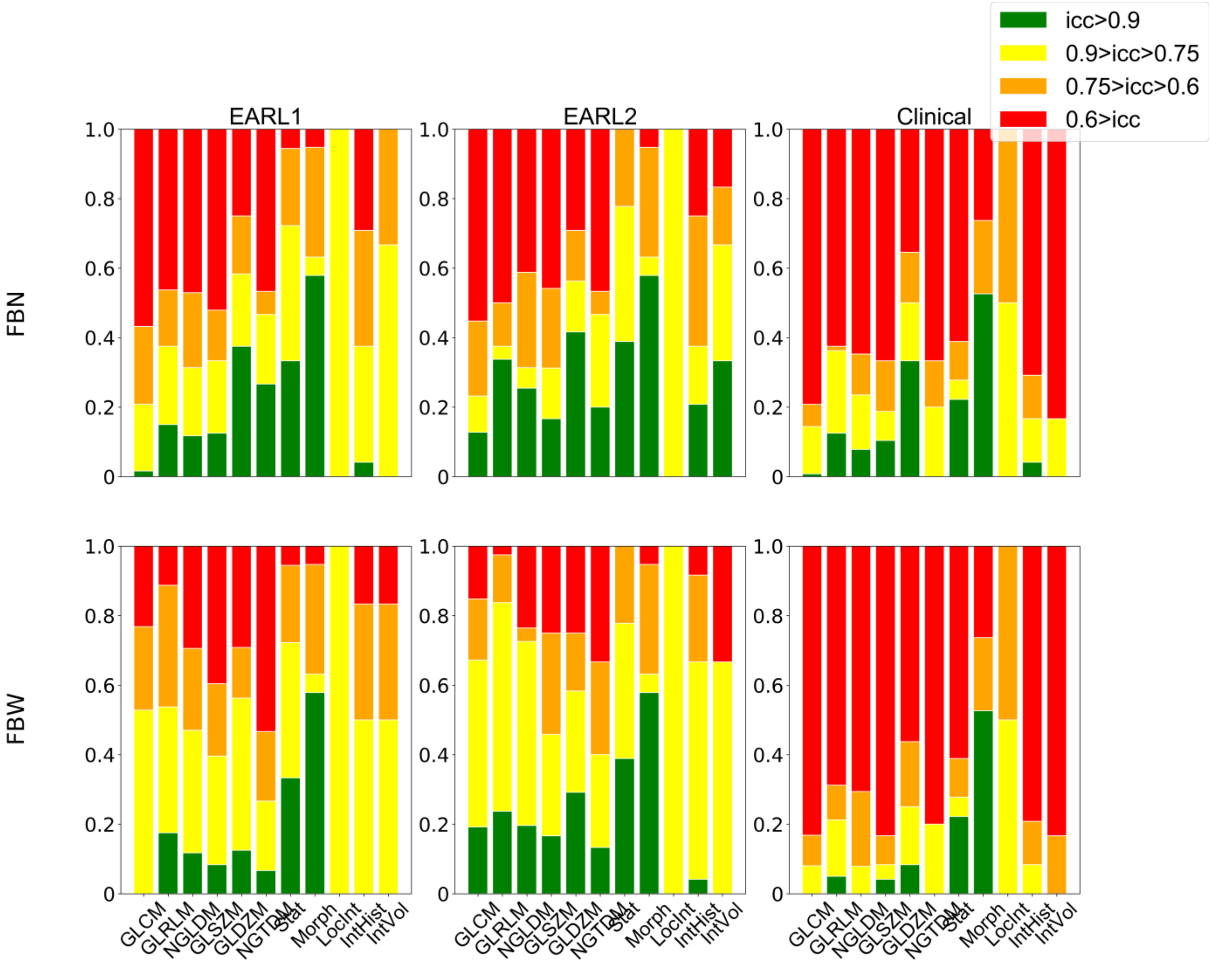

**Figure 3: Percentage of features extracted from the multicenter setting after resampling to cubic voxels yielding an excellent, good, moderate, or bad ICC for FBN (first row) and FBW discretization (second row)**

| Discretization | Recon. setting | Excellent ICC | Good ICC | Moderate ICC | Bad ICC |
|---|---|---|---|---|---|
| FBN | EARL1 | 0,587156 | 0,116972 | 0,12844 | 0,167431 |
|  | EARL2 | 0,651376 | 0,188073 | 0,103211 | 0,057339 |
|  | Clin | 0,786697 | 0,130734 | 0,029817 | 0,052752 |
| FBW | EARL1 | 0,759174 | 0,142202 | 0,066514 | 0,03211 |
|  | EARL2 | 0,883028 | 0,077982 | 0,013761 | 0,025229 |
|  | Clin | 0,926606 | 0,043578 | 0,009174 | 0,020642 |
| FBN - resampled | EARL1 | 0,598624 | 0,144495 | 0,149083 | 0,107798 |
|  | EARL2 | 0,697248 | 0,18578 | 0,06422 | 0,052752 |
|  | Clin | 0,743119 | 0,162844 | 0,050459 | 0,043578 |
| FBW - resampled | EARL1 | 0,855505 | 0,087156 | 0,029817 | 0,027523 |
|  | EARL2 | 0,908257 | 0,043578 | 0,022936 | 0,025229 |
|  | Clin | 0,928899 | 0,036697 | 0,009174 | 0,025229 |

**Table 1: Total percentage of features yielding an excellent, good, moderate, or bad ICC for the five statistical replicates**

| Discretization | Recon. setting | Excellent ICC | Good ICC | Moderate ICC | Bad ICC |
|---|---|---|---|---|---|
| FBN | EARL1 | 0,28211 | 0,275229 | 0,224771 | 0,21789 |
| | EARL2 | 0,300459 | 0,307339 | 0,213303 | 0,178899 |
| | Clin | 0,220183 | 0,412844 | 0,137615 | 0,229358 |
| FBW | EARL1 | 0,291284 | 0,474771 | 0,112385 | 0,12156 |
| | EARL2 | 0,357798 | 0,449541 | 0,09633 | 0,09633 |
| | Clin | 0,488532 | 0,348624 | 0,080275 | 0,082569 |
| FBN - resampled | EARL1 | 0,288991 | 0,231651 | 0,194954 | 0,284404 |
| | EARL2 | 0,305046 | 0,300459 | 0,211009 | 0,183486 |
| | Clin | 0,309633 | 0,323394 | 0,172018 | 0,194954 |
| FBW - resampled | EARL1 | 0,511468 | 0,259174 | 0,12156 | 0,107798 |
| | EARL2 | 0,550459 | 0,247706 | 0,12156 | 0,080275 |
| | Clin | 0,56422 | 0,259174 | 0,103211 | 0,073394 |

**Table 2: Total percentage of features yielding an excellent, good, moderate, or bad ICC for the four scans acquired on same scanner**

| Discretization | Recon. setting | Excellent ICC | Good ICC | Moderate ICC | Bad ICC |
|---|---|---|---|---|---|
| FBN | EARL1 | 0,114679 | 0,21789 | 0,16055 | 0,506881 |
| | EARL2 | 0,165138 | 0,165138 | 0,165138 | 0,504587 |
| | Clin | 0,08945 | 0,208716 | 0,082569 | 0,619266 |
| FBW | EARL1 | 0,08945 | 0,323394 | 0,305046 | 0,28211 |
| | EARL2 | 0,21789 | 0,394495 | 0,236239 | 0,151376 |
| | Clin | 0,061927 | 0,149083 | 0,112385 | 0,676606 |
| FBN - resampled | EARL1 | 0,151376 | 0,222477 | 0,201835 | 0,424312 |
| | EARL2 | 0,268349 | 0,103211 | 0,18578 | 0,442661 |
| | Clin | 0,12156 | 0,176606 | 0,082569 | 0,619266 |
| FBW - resampled | EARL1 | 0,110092 | 0,405963 | 0,252294 | 0,231651 |
| | EARL2 | 0,220183 | 0,454128 | 0,178899 | 0,146789 |
| | Clin | 0,073394 | 0,139908 | 0,137615 | 0,649083 |

**Table 3: Total percentage of features yielding an excellent, good, moderate, or bad ICC for the multi-center setting**