Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-dimensionally printed phantom inserts

Pfaehler, E.; Sluis, J. van; Merema, B.B.J.; Ooijen, P. van; Berendsen, R.C.M.; Velden, F.H.P. van; Boellaard, R.

Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally 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 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, Groningen, The Netherlands; [4]Department of Medical Physics, Zuyderland Medical Center, Heerlen, The Netherlands; [5]Section of Nuclear Medicine, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands; and [6]Department of Radiology and Nuclear Medicine, VU University Medical Center, Amsterdam, The Netherlands

The sensitivity of radiomic features to several confounding factors, such as reconstruction settings, makes clinical use challenging. To investigate the impact of harmonized image reconstructions on feature consistency, a multicenter phantom study was performed using 3-dimensionally 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 3-dimensionally printed inserts. Different heterogeneity pattern were realized by printing separate compartments that could be filled with different activity solutions. The inserts were placed in the National Electrical Manufacturers Association image-quality phantom and scanned various times. First, a list-mode scan was acquired and 5 statistically equal replicates were reconstructed. Second, the phantom was scanned 4 times on the same scanner. Third, the phantom was scanned on 6 PET/CT systems. All images were reconstructed using EANM Research Ltd. (EARL)–compliant and locally clinically preferred reconstructions. EARL-compliant reconstructions were performed without (EARL1) or with (EARL2) point-spread function. 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. 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 largest for the statistically equal replicates (70%–91% for FBN; 90%–96% for FBW discretization). For scans acquired on the same system, the percentage decreased, but most 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: for example, 40% for EARL1 and 60% for EARL2 versus 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 enhances the benefits of EARL-compliant reconstructions.

Key Words: ^{18}F-FDG PET/CT radiomic features; feature harmonization; image reconstruction

J Nucl Med 2020; 61:469–476
DOI: 10.2967/jnumed.119.229724

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 therefore has the potential to help in clinical decision making (*1*). One part of this approach is the rapidly growing field of radiomics, which aims to extract a large number of feature values from medical images describing tumor phenotype and tumor inter- and intraheterogeneity (*2–4*). In 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*). To make radiomic studies comparable over patients, institutions, and scanners, it is essential that radiomic features be 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 SUV features based on the SUV_{max}, SUV_{mean}, and SUV_{peak} by comparing phantom scans of the National Electrical Manufacturers Association (NEMA) NU2-2012 image-quality phantom. For this purpose, centers choose 1 reconstruction setting that is in line with the standards provided by EARL and uses an iterative reconstruction algorithm (EARL1). It has been shown that 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 usually also

Received Apr. 11, 2019; revision accepted Jul. 24, 2019.
For correspondence or reprints contact: Elisabeth Pfaehler, Department of Nuclear Medicine and Molecular Imaging, University Medical Center Groningen, Hanzeplein 1, Groningen, 9713GZ, The Netherlands.
E-mail: e.a.g.pfaehler@umcg.nl
Published online Aug. 16, 2019.

applies 1 reconstruction with settings leading to optimal lesion detection, which is used for clinical reads. As illustrated in Figure 1, the quality of a PET/CT image differs across these 3 reconstruction settings, which therefore have a high impact on the extracted radiomic features (Table 1).

The EARL harmonization is based on basic SUV features. To the best of our knowledge, no multicenter experimental study has yet investigated the effect of EARL harmonization on the variability of complex radiomic features. For this purpose, 1 object that reflects realistic heterogeneity uptake has to be scanned at multiple centers, and the feature values across centers have to be compared. Commercially available phantoms such as the NEMA image-quality phantom are not optimal, as they contain only spheric and homogeneous-uptake objects. Therefore, in this study, 3-dimensionally 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 3 institutions on 6 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 were reported.

MATERIALS AND METHODS

Phantom Design and 3-Dimensional Printing

Three 3-dimensionally printed phantom inserts were used in this study. PET scans of patients with non–small cell lung cancer served as models for the inserts. For this purpose, several non–small cell lung cancer tumors showing various heterogeneity uptake pattern were visually checked. Three tumors with different shapes and uptake characteristics were selected as models for the 3-dimensional printing. These tumors were segmented, slightly smoothed, scaled, and converted to a stereolithography file to make the printing possible. Differences in heterogeneity uptake were realized by printing 2 separate compartments that could 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 sizes of the inserts are displayed in Table 2. The printing was performed by a Form 2 printer (Formlabs Inc.), which relies on a stereolithography technique to cure its photopolymeric clear resin (FLGPCL02; Formlabs Inc.). A picture of the 3-dimensional inserts and the corresponding tumors is displayed in Figure 2. The inserts were placed at equal distances in the NEMA NU-2 image-quality phantom. The feature values of the phantom inserts were verified to be within the range of radiomic feature values extracted from 10 ^{18}F-FDG PET/CT studies of non–small cell lung cancer patients (*12*). More than 82% of the features are well within the clinically expected range, and only 1.6% show a large variation from the clinical data. Therefore, the inserts generate feature values that are representative of clinical data.

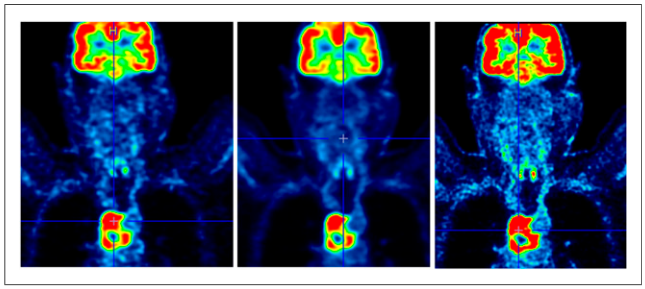

FIGURE 1. In patient with non–small cell lung cancer, Biograph Vision PET scan reconstructed with EARL2, EARL1, and clinically preferred reconstruction (from left to right).

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

Parameter	EARL2	EARL1	Clinically preferred
High gray level run emphasis, 3D average	142.24	175.07	130.10
Busyness, 3D	0.34	0.30	0.50
Contrast, 2D average	11.21	14.07	7.64

Phantom Scans

To obtain features comparable across institutions and PET/CT systems, only features that are reliable, repeatable, and reproducible should be used. Reliable features are defined as those yielding only marginal differences when extracted from images obtained under exactly the same conditions, and 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 settings, and reconstruction settings.

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

To measure repeatability, the phantom was scanned 4 times on the same system (Biograph mCT64) independently. That is, for every scan, the phantom was filled with an activity solution and placed at a slightly different position in the scanner. For differences in phantom

TABLE 2
Size of 3-Dimensionally Printed Inserts

Tumor	Size	Volume (mL)
1	40.3 × 44 × 54.5 mm	46.05
2, upper part	33.9 × 37 × 30 mm	10.75
2, lower part	24.3 × 40.5 × 36.6 mm	13.12
3, outer part	56 × 54 × 65.1 mm	65.35
3, necrotic core	25 × 24 × 31 mm	7.8

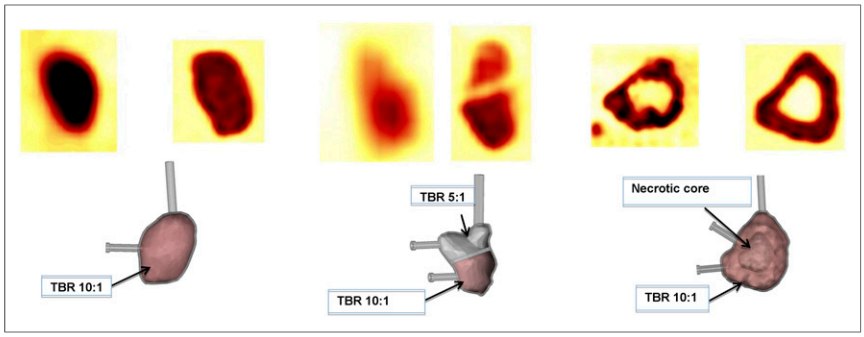

FIGURE 2. (Top) PET/CT images of original tumor (left) and phantom insert (right) for tumors 1, 2, and 3 (from left to right). (Bottom) Corresponding stereolithographed models with tumor-to-background ratio (TBR).

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, which could lead to slightly different delineations. Therefore, this scenario reflects a more realistic clinical setup.

Furthermore, a multicenter study was performed to measure reproducibility. The inserts were scanned at 3 institutions on 6 PET/CT systems including 4 manufactured by Siemens Healthcare (Biograph mCT40, Biograph mCT64, Horizon with an extra ring of detectors [TrueV option], and Biograph Vision), 1 by Philips Healthcare (Vereos), and 1 by GE Healthcare (Discovery MI 4 ring). The data were reconstructed with a clinically relevant scan duration of 60 s. The scan duration was adjusted for differences in phantom filling across centers. Table 3 lists the phantom fillings for each scan. Also, images were reconstructed using the scanner-defined reconstruction settings complying with the EANM standards (EARL1 and EARL2), as well as using the locally clinically preferred settings of each institution. The applied reconstruction algorithm, matrix size, and smoothing kernel for the reconstructed images are listed in Table 4. The inserts were segmented separately for each scan.

TABLE 3
Activity in Phantom Background and Tumor Inserts for 4 Scans Acquired on Same Scanner and Multicenter Setting

Scanner	Background activity (kBq/mL)	Tumor activity (kBq/mL) (parts 10:1/5:1)
Biograph mCT64		
Scan 1	2.2	21.8/15.8
Scan 2	2.3	22.6/15.5
Scan 4	1.9	2.1/14.5
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

PET Analysis

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

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 and have been 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 SUVs so that the phantom background had an SUV_{mean} 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 masks were resampled using trilinear interpolation. Before the extraction of textural features, images were discretized using a fixed bin number (FBN) of 64 and a fixed bin width (FBW) of 0.25.

Statistical Analysis

Data analysis was performed with Python, version 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. To measure feature consistency (i.e., reliability, repeatability, and reproducibility) for the 3 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 2-way single-measure model was used to evaluate the consistency of features for all scans. Every 3-dimensionally printed insert was regarded as a tumor in a patient, and each scan was regarded as 1 observer. The ICC is defined as the ratio of intercluster variability and the sum of intercluster and intracluster variability. Therefore, ICCs vary from 0 to 1, with 1 representing perfect agreement. Furthermore, a high ICC implies that the intracluster variability is low when compared with the intercluster 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*).

ICCs were compared between reconstruction settings, discretization methods, and original versus resampled data using a nonparametric permutation test. A permutation test compares 2 groups by checking differences in test statistics for the groups. The test randomly swaps 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 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; supplemental materials are available at http://jnm.snmjournals.org), including their ICCs for each reconstruction setting and discretization method.

TABLE 4
Applied Reconstruction Algorithm, Matrix Size, and Smoothing Factor for Each Scanner

Scanner	EARL1	EARL2	Clinical
Horizon	TOF, M256, 5 mm	PSF TOF, M256, 5 mm	PSF TOF, M256, 5 mm
Vereos	TOF, M144, 6 mm	PSF TOF, M144, 5 mm	TOF, M144, 4 mm
Biograph mCT40	TOF, M256, 5 mm	PSF TOF, M256, 5 mm	PSF TOF, M256, 7 mm
Biograph mCT64	TOF, M256, 5 mm	PSF TOF, M256, 5 mm	PSF TOF, M256, 7 mm
Vision	TOF, M256, 5 mm	PSF TOF, M256, 5 mm	PSF TOF, M256, 0 mm
Discovery MI	TOF, M192, 7 mm	VPFXS, M192, 7 mm	VPHD, M192, 0 mm

GE Healthcare's VPFXS is equivalent to PSF + TOF and VPHD is equivalent to PSF.

Figure 3 displays the percentage of features resulting in an excellent, good, moderate, or bad ICC sorted by feature groups for the statistically equal replicates and both discretization methods. The total percentage of excellent, good, and moderate ICCs was comparable across all reconstruction settings, with the highest values being 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%, respectively, for FBN discretization) (Supplemental Table 1). The EARL1 setting yielded the lowest percentage of features with an excellent ICC. When the feature groups were compared, the differences in ICCs were significant only for gray-level run-length matrix features ($P < 0.01$). A discretization with FBW resulted in more reliable features than FBN discretization, but the ICCs resulted in significant differences only for gray-level cooccurrence matrix features. Resampling to cubic voxels had almost no effect on reliability, although it led to a slight increase in the number of reliable features (Supplemental Fig. 1) with no significant differences in ICCs.

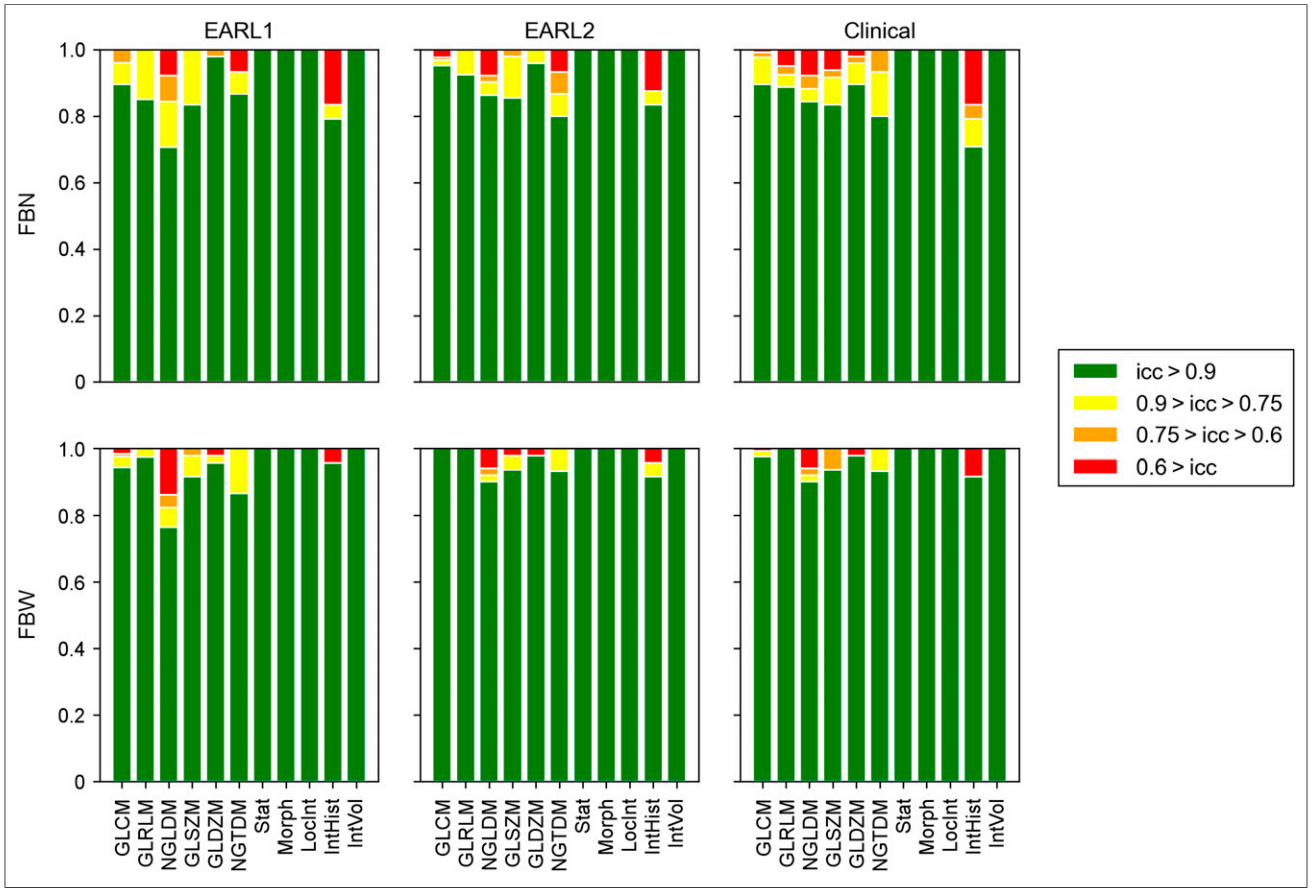

FIGURE 3. Percentage of features extracted from 5 statistically equal replicates yielding excellent, good, moderate, or bad ICC for FBN and FBW discretization for different feature groups. GLCM = gray-level cooccurrence matrix; GLRLM = gray-level run-length matrix; NGLDM = neighboring gray-level dependence matrix; GLSZM = gray-level size-zone matrix; GLDZM = gray-level distance-zone matrix; NGTDM = Neighboring gray-tone difference matrix; Stat = intensity-based statistics; Morph = morphology; LocInt = local intensity; IntHist = intensity histogram; IntVol = intensity volume.

By comparison, the percentages of features yielding excellent, good, moderate, or bad ICCs for the 4 scans acquired on the same system are displayed in Figure 4. The number of features yielding an excellent ICC decreased when compared with the 5 statistically equal replicates. However, most features still resulted in a good or moderate ICC. Also, 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%, respectively, for FBN discretization), a slight increase after resampling (Supplemental Table 2), and significant differences for gray-level cooccurrence matrix features ($P < 0.01$). The differences between clinically preferred and EARL-compliant reconstructions also were not significant, but the clinically preferred reconstruction yielded the highest percentage, and the EARL1 setting the lowest percentage, of repeatable features. The only feature group whose features were less repeatable after resampling were the morphologic features (Supplemental Fig. 2).

In the multicenter setting, the percentage of features yielding a moderate or better ICC was low when compared with the other scan settings (Fig. 5). Also, 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, and 32.3% for the clinically preferred setting vs. 49.3%, 49.5%, and 38%, respectively, for FBN discretization). Significant differences in ICCs between the 2 discretization methods were found only for the EARL-compliant reconstructions and some textural feature groups (gray-level cooccurrence matrix and gray-level run-length matrix features for both EARL-compliant reconstructions, neighboring gray-level dependence matrix and gray-level size-zone matrix for EARL2). For discretization with FBN, only small and nonsignificant discrepancies could be observed between the reconstruction settings. However, for FBW discretization, the difference between EARL-compliant reconstructions and clinically preferred reconstructions led to significant differences for most textural feature groups. In the multicenter setting, the local clinically preferred reconstructions differed substantially between sites and scanners, whereas this was not the case in the single-scanner experiments described. Significant differences in ICCs between EARL1 and EARL2 were observed only for gray-level cooccurrence matrix features and gray-level run-length matrix features when discretized with FBW. A resampling to cubic voxels was beneficial, especially for textural feature groups, although the differences were not significant (Supplemental Fig. 3). In addition, the only feature group resulting in less reproducible features after resampling was the group of morphologic features, for which a significant difference was observed (Supplemental Table 3).

DISCUSSION

To the best of our knowledge, this was the first multicenter and multivendor experimental study to investigate the impact of

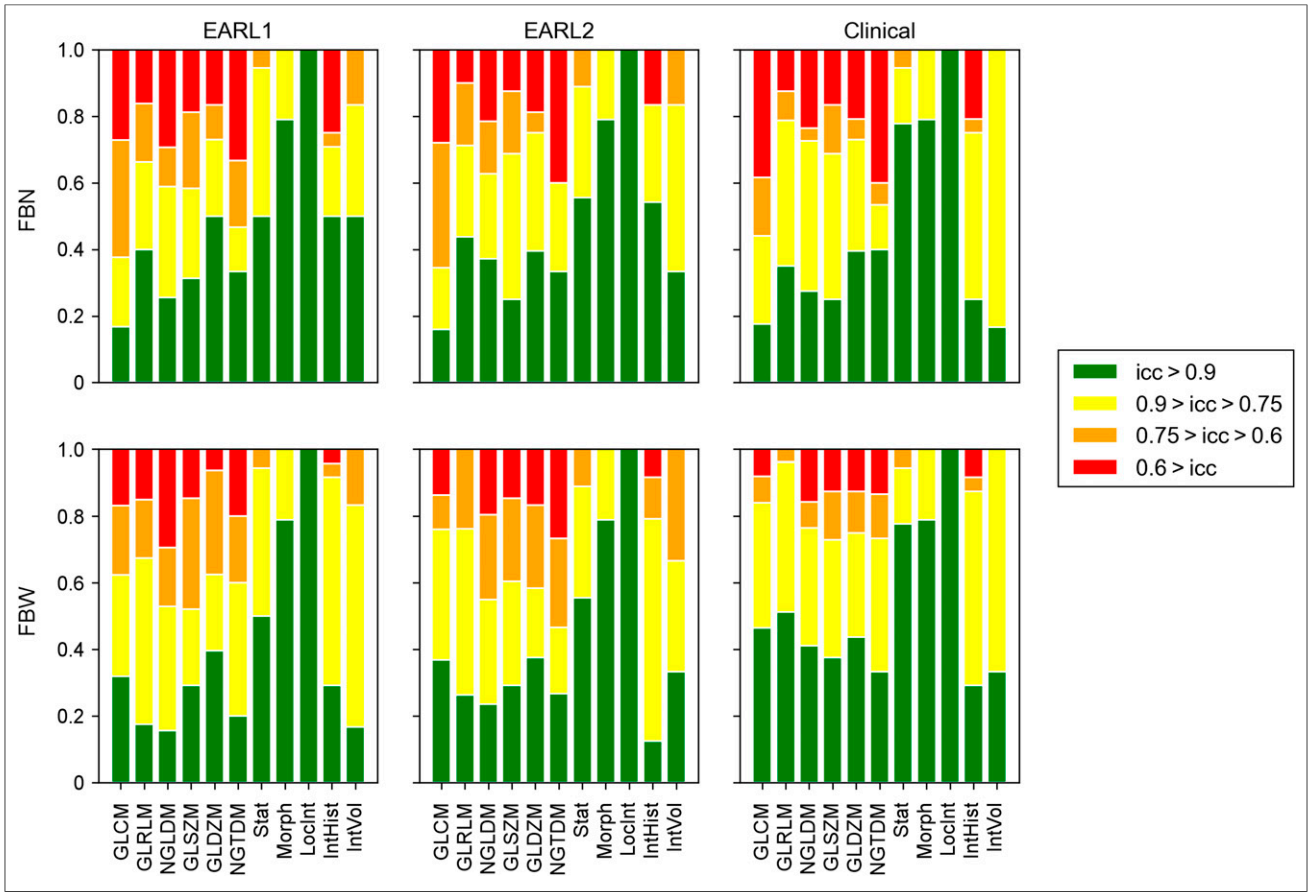

FIGURE 4. Percentage of features extracted from 4 scans acquired on same PET/CT system yielding excellent, good, moderate, or bad ICC for FBN and FBW discretization. GLCM = gray-level cooccurrence matrix; GLRLM = gray-level run-length matrix; NGLDM = neighboring gray-level dependence matrix; GLSZM = gray-level size-zone matrix; GLDZM = gray-level distance-zone matrix; NGTDM = Neighboring gray-tone difference matrix; Stat = intensity-based statistics; Morph = morphology; LocInt = local intensity; IntHist = intensity histogram; IntVol = intensity volume.

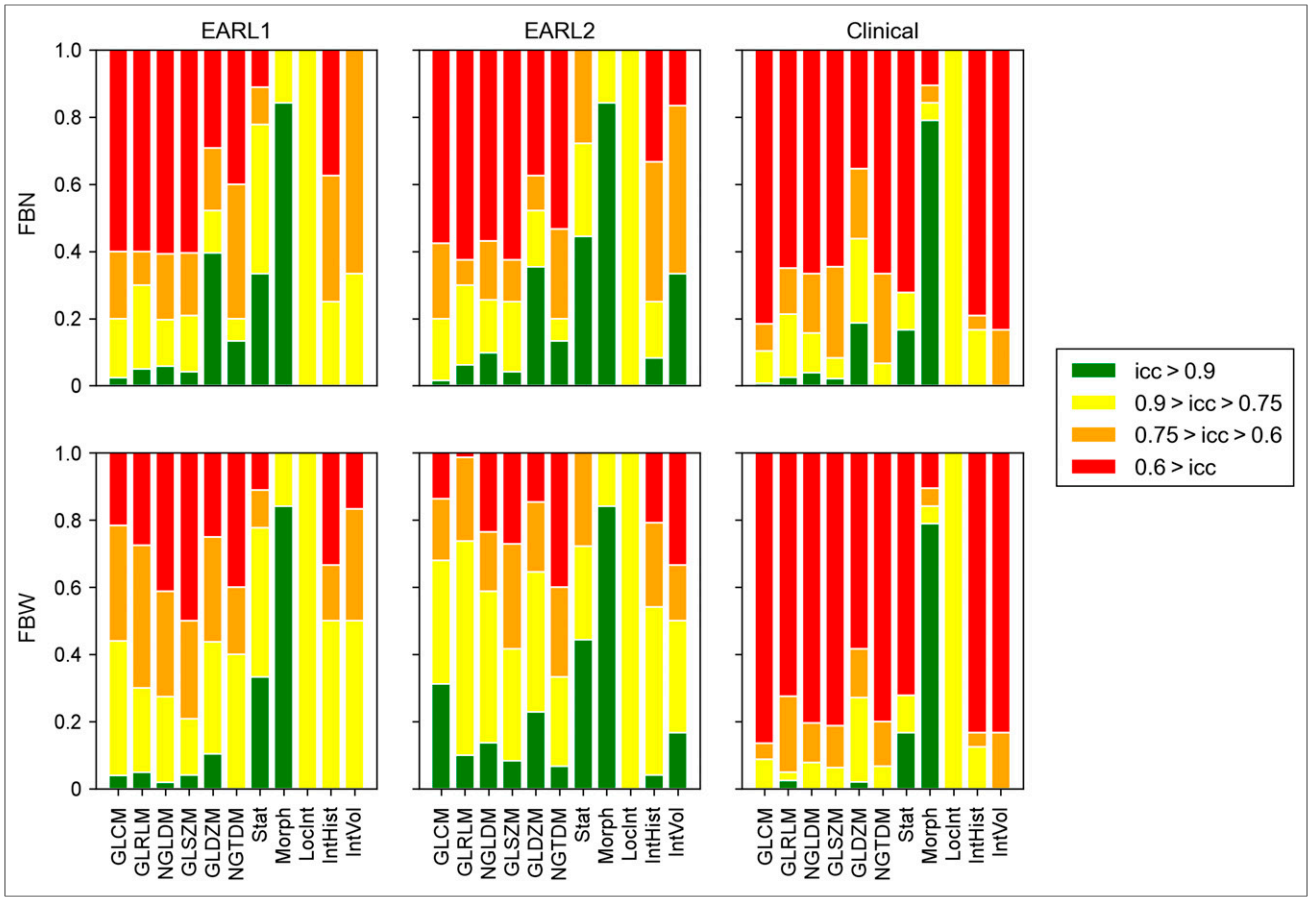

FIGURE 5. Percentage of features extracted from multicenter setting yielding excellent, good, moderate, or bad ICC for FBN and FBW discretization. GLCM = gray-level cooccurrence matrix; GLRLM = gray-level run-length matrix; NGLDM = neighboring gray-level dependence matrix; GLSZM = gray-level size-zone matrix; GLDZM = gray-level distance-zone matrix; NGTDM = Neighboring gray-tone difference matrix; Stat = intensity-based statistics; Morph = morphology; LocInt = local intensity; IntHist = intensity histogram; IntVol = intensity volume.

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 might be that the clinically preferred reconstructions varied widely in spatial resolution and contrast recovery across PET/CT systems. Because radiomic features are sensitive to resolution and image noise, these variations could be the reason for a higher variation in radiomic features (*18*). This possibility is in line with the fact that differences in feature consistency between reconstruction settings were not visible in the 5 statistically equal replicates and the 4 scans acquired on the same scanner, for which the same local clinically preferred reconstruction was applied.

In the multicenter setting, EARL-compliant images yield comparable image quality. This might be the reason for the low differences in reliability, repeatability, and reproducibility for these 2 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*).

Although 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 that is visible in some of the clinically preferred reconstructions. This effect is lower in the updated EARL standards (EARL2), which yield higher contrast recoveries and spatial resolution and are therefore preferred for future multicenter studies. One limitation of this study is that we do not report the accuracy of feature values. Because it was demonstrated before that radiomic features are biased as a function of acquisition parameters, image reconstruction settings, and noise (*18,20,21*), there is an urgent need for standardization of feature values 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 to improve the reproducibility of radiomic features. Nonetheless, because a 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 has limitations, as the 3-dimensionally printed inserts reflect only 3 coarse heterogeneity patterns. However,

they provide a more realistic scenario than publicly available phantoms containing only spheres. Furthermore, phantoms have the advantage of providing a more reproducible setting than patient scans, because the activity solution within the 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 better repeatability and less sensitivity to differences in delineations for FBW discretization (*7,10,23*). Furthermore, Orlhac et al. demonstrated that discretization with FBW led to more meaningful features—that is, features that can distinguish well between tumor types (*23*). Our results also confirm the benefit of 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 better reproducibility of radiomic features. A possible explanation might be that a common voxel size might lead to more comparable features because a large number of features are sensitive to differences in slice thickness and voxel size (*26,27*). The only feature group not benefiting from resampling were the morphologic features. This effect was observed only in the scan setups in which 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 5 statistically equal replicates to the 4 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}, and SUV_{mean}) are affected by phantom repositioning (*30*), so it is likely that repositioning also affects more complex textural features. However, as 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 for use in radiomic analysis in the clinic.

CONCLUSION

This study reports on the impact of EARL-compliant reconstructions on the reliability, repeatability, and reproducibility of radiomic features in comparison with clinically 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. Discretization with FBW and resampling to cubic 2-mm 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 they have higher contrast recovery and spatial resolution while providing radiomic performance similar to the EARL1 standards (*11*).

DISCLOSURE

This work is part of the STRaTeGy research program (project 14929), which is (partly) financed by The Netherlands Organisation for Scientific Research (NWO). This study was financed by the POINTING project of the Dutch Cancer Society (grant 10034). No other potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We thank Hinke Schokker and Johan R. de Jong for help with the phantom scans.

KEY POINTS

QUESTION: Which reconstruction algorithm leads to the most stable radiomic features in a multicenter and multivendor 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 FBW and resampled to isotropic voxels before feature extraction.

IMPLICATIONS FOR PATIENT CARE: To make radiomic features comparable across multiple centers, multicenter radiomic studies should be performed using harmonized (EARL-compliant) reconstructions, and images should be discretized using a FBW 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 Med.* 2017;38:122–139.
3. Lambin P, Rios-Velazquez E, Leijenaar R. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* 2012;48: 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. *Sci Reports.* 2015;5: 11044.
7. van Velden FHP, 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;18:788–795.
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.* 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 [^{18}F]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. *PLoS One.* 2019;14:e0212223.
14. Zwanenburg A, Leger S, Vallières M, Löck S. The image biomarker standardisation initiative. arXiv.org website. https://arxiv.org/pdf/1612.07003.pdf. Published 2016. Accessed October 16, 2019.
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.* 2019;46:665–678.
19. Lasnon C, Majdoub M, Lavigne B, et al. ^{18}F-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 (Bellingham).* 2015;2:041002.
21. Yan J, Chu-Shern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in ^{18}F-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. ^{18}F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. *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 ^{18}F-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.* 2019;60:864–872.
28. Bashir U, Azad G, Siddique MM, et al. The effects of segmentation algorithms on the measurement of ^{18}F-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;7:358.