Impact of ^{18}F-FDG PET Intensity Normalization on Radiomic Features of Oropharyngeal Squamous Cell Carcinomas and Machine Learning–Generated Biomarkers

Stefan P. Haider[1,2], Tal Zeevi[2], Kariem Sharaf[1], Moritz Gross[2,3], Amit Mahajan[2], Benjamin H. Kann[4], Benjamin L. Judson[5], Manju L. Prasad[6], Barbara Burtness[7], Mariam Aboian[2], Martin Canis[1], Christoph A. Reichel[1], Philipp Baumeister[1], and Seyedmehdi Payabvash[2]

[1]Department of Otorhinolaryngology, LMU Clinic of Ludwig Maximilians University of Munich, Munich, Germany; [2]Section of Neuroradiology, Yale School of Medicine, New Haven, Connecticut; [3]Charité Center for Diagnostic and Interventional Radiology, Charité Universitätsmedizin Berlin, Berlin, Germany; [4]Department of Radiation Oncology, Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts; [5]Division of Otolaryngology, Yale School of Medicine, New Haven, Connecticut; [6]Department of Pathology, Yale School of Medicine, New Haven, Connecticut; and [7]Section of Medical Oncology, Yale School of Medicine, New Haven, Connecticut

We aimed to investigate the effects of ^{18}F-FDG PET voxel intensity normalization on radiomic features of oropharyngeal squamous cell carcinoma (OPSCC) and machine learning–generated radiomic biomarkers. **Methods:** We extracted 1,037 ^{18}F-FDG PET radiomic features quantifying the shape, intensity, and texture of 430 OPSCC primary tumors. The reproducibility of individual features across 3 intensity-normalized images (body-weight SUV, reference tissue activity ratio to lentiform nucleus of brain and cerebellum) and the raw PET data was assessed using an intraclass correlation coefficient (ICC). We investigated the effects of intensity normalization on the features' utility in predicting the human papillomavirus (HPV) status of OPSCCs in univariate logistic regression, receiver-operating-characteristic analysis, and extreme-gradient-boosting (XGBoost) machine-learning classifiers. **Results:** Of 1,037 features, a high (ICC $\geq$ 0.90), medium (0.90 $>$ ICC $\geq$ 0.75), and low (ICC $<$ 0.75) degree of reproducibility across normalization methods was attained in 356 (34.3%), 608 (58.6%), and 73 (7%) features, respectively. In univariate analysis, features from the PET normalized to the lentiform nucleus had the strongest association with HPV status, with 865 of 1,037 (83.4%) significant features after multiple testing corrections and a median area under the receiver-operating-characteristic curve (AUC) of 0.65 (interquartile range, 0.62–0.68). Similar tendencies were observed in XGBoost models, with the lentiform nucleus–normalized model achieving the numerically highest average AUC of 0.72 (SD, 0.07) in the cross validation within the training cohort. The model generalized well to the validation cohorts, attaining an AUC of 0.73 (95% CI, 0.60–0.85) in independent validation and 0.76 (95% CI, 0.58–0.95) in external validation. The AUCs of the XGBoost models were not significantly different. **Conclusion:** Only one third of the features demonstrated a high degree of reproducibility across intensity-normalization techniques, making uniform normalization a prerequisite for interindividual comparability of radiomic markers. The choice of normalization technique may affect the radiomic features' predictive value with respect to HPV. Our results show trends that normalization to the lentiform nucleus may improve model performance, although more evidence is needed to draw a firm conclusion.

Key Words: PET; SUV; normalization; radiomics; machine learning

J Nucl Med 2024; 65:803–809
DOI: 10.2967/jnumed.123.266637

In clinical ^{18}F-FDG PET scans, voxel intensity values usually represent ^{18}F-FDG activity concentration in becquerels per milliliter. Apart from reflecting the metabolic properties of tissues, the measured activity depends on the amount of intravenously administered ^{18}F-FDG activity, distribution volume, and blood glucose level, among numerous other factors (*1*). Thus, many investigators advocate for voxel intensity value normalization to enable interindividually and interinstitutionally comparable quantitative PET assessment (*1*). In clinical settings, normalization is commonly performed with SUVs, with body-weight SUVs being widely used.

Radiomics can provide objective imaging biomarkers for quantifying intensity, texture, and shape features of PET findings (*2*). These biomarkers, whether assessed individually or incorporated into sophisticated models, demonstrate potential in predicting prognosis, treatment response, and molecular tumor traits, among other end points (*3*).

Since most radiomic features are by extension derived from voxel intensity values, the same factors impeding interindividual comparability of PET activity concentrations are expected to affect radiomic features. Consequently, and presumably because of its widespread clinical use, SUV normalization is performed by many groups before radiomic analysis (*4–6*). However, there is currently no empiric evidence supporting any normalization technique for PET radiomics specifically. Recent studies explored the impact of PET imaging, reconstruction, delineation, and feature extraction–associated parameters as well as image noise on radiomic features using clinical and phantom data (*7–10*). Yet, the extent to which voxel intensity normalization using the SUV or alternative techniques affects radiomic features remains elusive.

PET is a mainstay of diagnostic work-up, treatment planning, and surveillance for oropharyngeal squamous cell carcinoma (OPSCC). Human papillomavirus (HPV)–associated OPSCC is recognized as

Received Sep. 12, 2023; revision accepted Feb. 13, 2024.
For correspondence or reprints, contact Seyedmehdi Payabvash (sam.payabvash@yale.edu) or Stefan P. Haider (stefan.haider@yale.edu).
Published online Mar. 21, 2024.

a distinct entity with different etiopathogenetic, demographic, and prognostic characteristics (*11*). Using a multiinstitutional and multinational OPSCC cohort, we investigated the radiomic features' reproducibility across several PET voxel intensity-normalization techniques. To explore the impact of normalization on the radiomic markers' predictive capacity, we compared the performance of individual features and machine-learning models in predicting HPV association after different normalization techniques were applied. Finally, we probed into using feature reproducibility across normalization techniques as a feature selection criterion for machine-learning models.

In addition to the body-weight SUV, we normalized voxel intensities by calculating standardized uptake ratios to reference tissues, specifically the lentiform nucleus of the brain and the cerebellum (*12–14*). The lentiform nucleus and the cerebellum were selected because of their stable ^{18}F-FDG uptake, ease of localization and measurement, feasibility based on prior studies, the ability for imaging in the same bed position as the OPSCC primary, and the more reliable inclusion in dedicated head-and-neck reconstructions than the liver or mediastinal blood pool (*12–14*).

MATERIALS AND METHODS

Data Acquisition and Allocation

For this multicentric retrospective study, we used a previously reported dataset of 435 biopsy-proven nonrecurrent OPSCC patients with available pretreatment ^{18}F-FDG PET of the neck and known HPV status (*15*). In compliance with the Health Insurance Portability and Accountability Act, data were acquired from Yale School of Medicine and 4 publicly available collections in The Cancer Imaging Archive (https://www.cancerimagingarchive.net/) (*15–19*). Data collection from Yale School of Medicine was approved by the institutional review board, and the need for written informed consent was waived (*15*). After exclusion of 5 patients because of missing information for PET normalization, we allocated separate training (n = 325), independent validation (n = 79), and external validation cohorts (n = 26) for machine-learning analysis by retaining the HPV-stratified randomization from a previous study (*15*).

PET Intensity Normalization

PET normalization was performed with the body-weight SUV and by taking the standardized uptake ratio to the left lentiform nucleus and the cerebellum (*12*,*13*,*15*); the raw intensities (i.e., voxels represent becquerels/milliliter) were additionally analyzed. Supplemental Methods 1 provides a detailed description (supplemental materials are available at http://jnm.snmjournals.org) (*12*,*13*,*15*,*20*,*21*).

Tumor Segmentation and Radiomic Feature Extraction

After manual segmentation of the primary tumors, the 3-dimensional tumor masks and corresponding PET images (4 images from 4 types of intensity normalization) were fed into a customized radiomics pipeline for performing image preprocessing and radiomic extraction. Preprocessing included voxel dimension resampling and derivative image generation by Laplacian-of-gaussian filtering and via a coif-1 wavelet transform. We applied a fixed-bin-width method for voxel intensity discretization, which was shown to be more appropriate for inter- and intrapatient feature comparison in a clinical setting by Leijenaar et al. (*6*,*22*). Supplemental Methods 2 and 3 and Supplemental Tables 1 and 2 detail our segmentation and radiomics extraction pipeline, also specifying the features' compliance with the image biomarker standardization initiative (*6*,*15*,*20*,*22–24*). Finally, the pipeline extracted 14 shapes, 198 first-order features, and 825 texture features per patient per PET intensity–normalization technique.

Reproducibility of Radiomic Features Across PET Normalization Techniques

To investigate whether intensity normalization affects the values of radiomic features, we used the entire patient cohort to calculate a 2-way mixed effects absolute agreement single-rater and measurement intraclass correlation coefficient (ICC) (*25*) for each feature to quantify its reproducibility across the 4 normalization techniques. Features with ICC = 1, 1 > ICC ≥ 0.999, 0.999 > ICC ≥ 0.90, 0.90 > ICC ≥ 0.75, and ICC < 0.75 were considered to have perfect, nearly perfect, high, medium, and low degrees of reproducibility, respectively. Cutoffs were selected to obtain categories where normalization had no impact (ICC = 1) and no relevant impact (1 > ICC ≥ 0.999) on features. The ICC less than 0.75 cutoff was selected to exclude low outliers, and the remaining cutoffs were used to divide all other features into categories of similar feature counts on the basis of knowledge from preliminary analyses. Shape features by design reflect tumor morphology and size and therefore are not affected by voxel intensity normalization (unless voxel intensity–based segmentation is applied). We included them in our analyses as positive controls, expecting an ICC of 1 for all shape features.

Univariate Association Analysis with HPV Status

To assess the impact of intensity normalization on the radiomic features' utility as predictors of clinically relevant outcomes, we conducted univariate association analysis with HPV status in the total cohort. A series of logistic regressions was performed with HPV as the dependent variable and each feature from each intensity-normalized image type as the independent variable. In addition, the area under the receiver-operating-characteristic curve (AUC) was determined for each feature as a measure of univariate predictive ability.

Finally, we plotted the AUC of the features against their ICC values to investigate if the features' reproducibility across PET normalization methods impacts their association with HPV.

Machine-Learning Models for HPV Prediction

To determine the utility of the radiomic feature sets in differentiating the HPV status after PET intensity normalization, we devised, optimized, and validated 4 separate machine-learning models using features from a different intensity-normalized image type. An extreme-gradient-boosting (XGBoost) classifier combined with a minimum-redundancy-maximum-relevance (MRMR) feature selection formed the backbone of our machine-learning pipeline. The training cohort was used for model development on the basis of cross validation (CV) and Bayesian hyperparameter optimization (Supplemental Table 3) (*26*); final models were validated in the independent and external validation datasets. Notably, the machine-learning analyses outlined above were conducted independently from features' ICC values, and a detailed description is provided in Supplemental Methods 4 (*15*,*26–28*).

As an alternative to uniform normalization of PET images before machine learning, we investigated selective use of radiomic features with high robustness against differences in PET normalization as indicated by high ICC values. We designed an experiment based on the above-introduced CV framework using MRMR feature selection and the XGBoost classifier. We iteratively excluded the least robust features from the feature sets until a small set of highly robust features remained. Specifically, we iteratively excluded the 30 features with the lowest ICC scores from each intensity-normalized feature set and evaluated the retained features in the CV framework until fewer than 50 features remained. The MRMR algorithm was configured to select a fixed 20 features in every iteration, and XGBoost hyperparameters were set to the default recommendations. This study used a modified version of a previously reported machine-learning pipeline (*15*).

Statistical Analysis

Radiomic features were standardized before analysis by subtracting the cohort mean and dividing by the SD. To preclude information leakage, we used the training-cohort mean and SD to standardize the training, independent validation, and external validation cohorts for machine-learning analyses; in the CV experiments, the CV training data mean and SD were used to standardize both the training and the test data. The AUC quantified the XGBoost performance. The DeLong test was used to compare AUC scores. *P* values of less than 0.05 ascertained significance. Logistic regression *P* values corresponding to each intensity-normalized image type were adjusted for multiple testing using the Benjamini–Hochberg method (p.adjust function, stats R-package; R Project for Statistical Computing (*29*)). All statistical and machine-learning analyses were performed in R version 3.6.0 (*29*).

RESULTS

Patient Characteristics

Of the 430 patients included in this study, 73 (17.0%) were female and 313 (72.8%) had HPV-associated OPSCC. Detailed characteristics of patients and PET imagery are reported in Supplemental Table 4 and in previous works (*15–19*).

Reproducibility of Radiomic Features Across PET Normalization Techniques

Of the 1,037 radiomic features, 14 (1.4%), 47 (4.5%), 295 (28.4%), 608 (58.6%), and 73 (7%) had perfect, nearly perfect, high, medium, and low degrees of reproducibility across normalization techniques, respectively. Figure 1 summarizes the ICC scores and corresponding 95% CI; Supplemental Table 5 provides a comprehensive list (*19*,*22*). The 14 shape features achieved perfect reproducibility (ICC = 1), as anticipated. All skewness and kurtosis first-order features attained nearly perfect reproducibility (1 > ICC ≥ 0.999), indicating that PET intensity normalization had no relevant effect.

A breakdown of ICC values by image type (original, Laplacian-of-gaussian filtering, wavelet decompositions) and radiomic feature family shows that feature counts and fractions per reproducibility category were generally similar across image types and feature families (Supplemental Table 6). Wavelet decomposition high-pass filtering in 2 or more spatial directions led to slightly higher median ICC values and narrower ICC distributions, whereas low-pass filtering had the opposite effect, as compared with original image features (Supplemental Fig. 1.1.). The Laplacian-of-gaussian filtering sigma-value (3 mm vs. 6 mm) had no relevant effect on ICC distributions (Supplemental Fig. 1.1), and ICC distributions were similar across feature families (except shape features; Supplemental Fig. 1.2). Notably, the scatterplot reveals a cluster of 45 highly reproducible gray-level cooccurrence matrix features with an ICC greater than 0.98 (Supplemental Fig. 1.2), among which we identified 11 correlation features, 10 inverse difference moment–normalized features, and 10 inverse difference–normalized features.

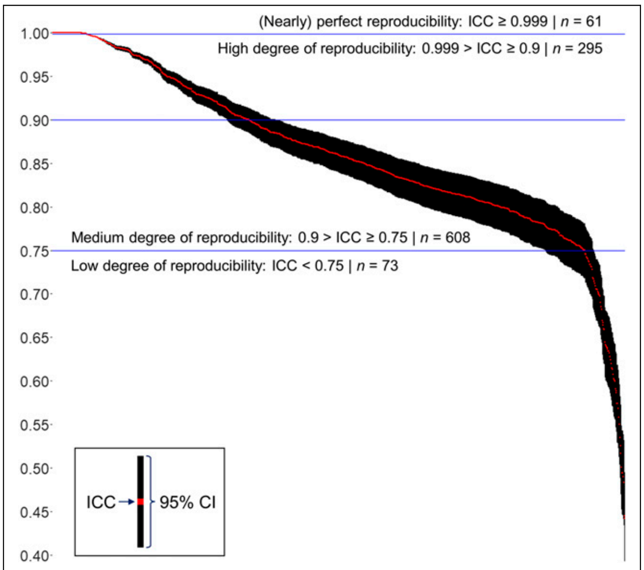

FIGURE 1. Reproducibility of radiomic features across PET normalization techniques. Each radiomic feature is represented by vertical bar, as shown in inset, with red central marker representing its ICC value and black vertical bar representing corresponding 95% CI. Features are ordered from left to right by decreasing ICC score.

Univariate Association Analysis with HPV Status

Features extracted from PET normalized to the lentiform nucleus exhibited the strongest association with HPV in univariate analysis (Table 1; Supplemental Figs. 2.1 and 2.2).

A breakdown by image type (original, Laplacian-of-gaussian filtering, wavelet decompositions) and radiomic feature family revealed that the effect of PET normalization on the features' predictive value was highly consistent across image types and feature families (Supplemental Figs. 2.3–2.6; Supplemental Table 7).

Plotting of the features' AUC (measuring their univariate association with HPV) against their ICC (measuring their reproducibility across normalization methods) revealed no clear association visually, which is confirmed in Spearman correlation analysis showing a trivial although statistically significant association ($\rho = -0.04$, $P = 0.005$; Supplemental Figs. 3.1–3.3). Supplemental Figure 3.1 again reveals a stratification of AUC scores by the PET normalization method.

Machine-Learning Models for HPV Prediction

Table 2 summarizes the optimized XGBoost classifiers' performance. The classifier using lentiform nucleus–normalized PET features attained the numerically highest averaged CV AUC ± SD of 0.72 ± 0.07. The AUC of all models in the independent and external validation datasets was similar to their CV AUC (Table 2). We detected no significant differences among the models' independent validation AUCs and among the models' CV AUCs (all $P > 0.05$; Supplemental Fig. 4 (*30*)). The *P* value from comparing the lentiform nucleus–normalized and SUV-normalized models' CV AUCs approached significance ($P = 0.052$; Supplemental Fig. 4.2). Supplemental Table 8 depicts the feature importance scores for all XGBoost classifiers (*19*,*22*,*26*); notably, most MRMR-selected features came from wavelet images.

Iterative exclusion of the features with the lowest ICC scores (i.e., the least reproducible features across normalization techniques) from models resulted in a convergence of XGBoost performance (i.e., models trained with increasingly more reproducible features yielded increasingly similar AUC values; Fig. 2). However, as the total number of features available to models decreased, the model performance deteriorated steadily. In Figure 2, no plateau providing a balance between classifier performance and feature reproducibility can be identified.

DISCUSSION

Despite the growing interest in quantitative radiomic analysis of PET imagery, the effect of voxel intensity value normalization on radiomic markers and machine-learning models has not been

TABLE 1
Univariate Association Analysis

Parameter	PET normalization method: SUV	None (raw intensities)	Reference tissue (lentiform nucleus)	Reference tissue (cerebellum)
Median of absolute standardized regression coefficients*	0.27 (0.17–0.36)	0.45 (0.30–0.54)	0.51 (0.36–0.62)	0.38 (0.26–0.48)
Number of significant features				
Before adjustment for multiple testing†	623	857	877	809
After adjustment for multiple testing†	591	843	865	785
Median AUC‡	0.59 (0.56–0.62)	0.62 (0.59–0.64)	0.65 (0.62–0.68)	0.63 (0.60–0.65)

*Standardized regression coefficients were converted to absolute values before median was calculated. Interquartile range is in parentheses.

†Adjustment by Benjamini–Hochberg method.

‡AUC values < 0.5 were substituted with 1 – AUC before median and interquartile range were calculated. Interquartile range is in parentheses.

studied before to the best of our knowledge. Using a dataset of 430 OPSCC patients with pretherapeutic ^{18}F-FDG PET scans, we investigated the reproducibility of radiomic features across PET normalization techniques (SUV, standardized uptake ratio to lentiform nucleus and cerebellum, raw intensities). We demonstrated that among 1,037 features extracted from OPSCC primary tumors, only 356 (34.3%) attained at least a high degree of reproducibility (ICC $\geq$ 0.90), suggesting that their values would be similar regardless of the normalization technique applied (Fig. 1). Given the lower reproducibility of most features, we concluded that uniform normalization or at least conversion of voxel intensities to identical units (e.g., Bq/mL) is a prerequisite for extraction of interindividually comparable radiomic features. We observed differential effects of PET normalization on the reproducibility of individual features (Supplemental Table 5). Highly voxel intensity–dependent features attained low ICC scores, because intensity normalization directly influences their feature value. For example, the 2 least reproducible features were minimum features, measuring the minimum intensity within the tumor. Most texture features, capturing patterns and relationships of intensity between adjacent voxels, were also markedly and to varying degrees affected by normalization. Conversely, we discovered features with mathematic definitions (*22*) that reduced or nullified the effects of normalization constants applied to all voxel values via multiplication and division. For instance, skewness and kurtosis features attained an ICC between 0.999 and 1 because, on the basis of their mathematic definitions, normalization constants cancel out. Presumably, their ICC was slightly decreased because of rounding in the complex feature extraction pipeline. Thus, the differential effects of normalization on feature values are partly explained by mathematic feature definitions. In addition, the effects of PET normalization on feature reproducibility are amplified or mitigated by image filters (Supplemental Fig. 1.1).

In addition, we studied the effects of intensity normalization on the features' predictive value with respect to the HPV status. Using the training cohort, we devised and optimized 4 XGBoost classifiers, each using a MRMR-selected feature subset from a

TABLE 2
Machine-Learning Classifier Performance

Parameter	PET normalization method: SUV	None (raw intensities)	Reference tissue (lentiform nucleus)	Reference tissue (cerebellum)
CV AUC*	0.65 (0.08)	0.67 (0.08)	0.72 (0.07)	0.69 (0.07)
Independent validation AUC†	0.64 (0.49–0.79)	0.74 (0.62–0.87)	0.73 (0.60–0.85)	0.72 (0.58–0.87)
External validation AUC†	0.68 (0.47–0.89)	0.74 (0.54–0.94)	0.76 (0.58–0.95)	0.78 (0.60–0.96)

*Model performance was quantified by AUC in each CV test fold and averaged across all CV iterations. Data are median with SD in parentheses.

†DeLong's method was used to estimate 95% CIs. 95% CIs are in parentheses.

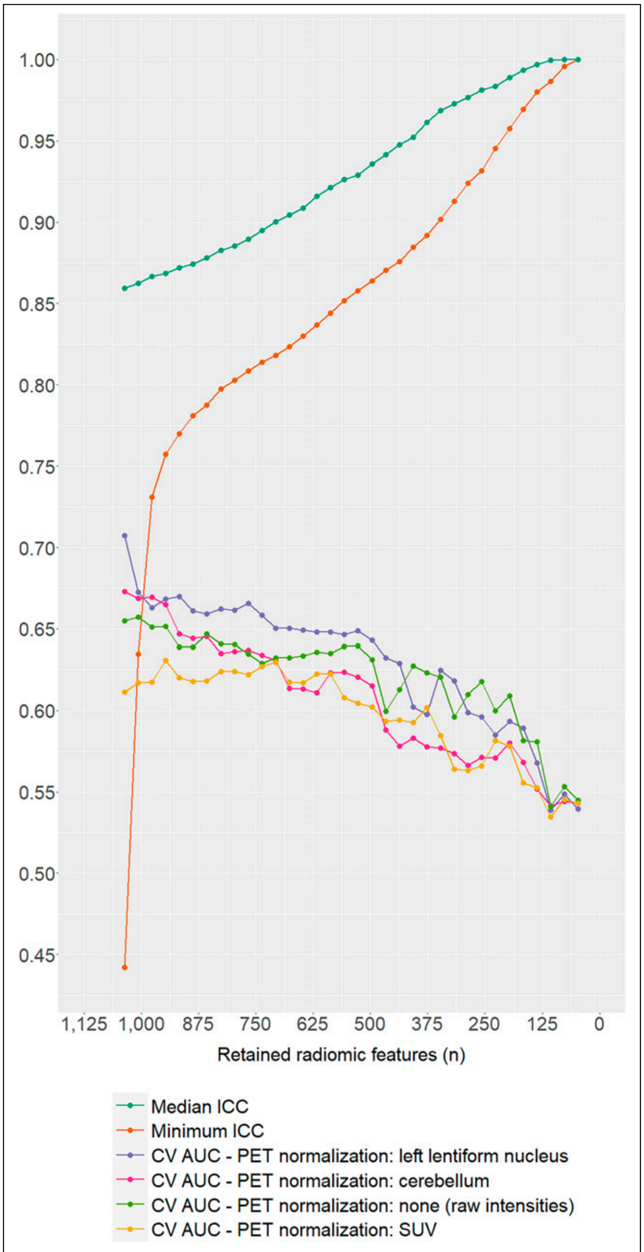

FIGURE 2. Iterative exclusion of least reproducible radiomic features from machine-learning models. We iteratively excluded 30 radiomic features with lowest ICC from radiomic feature sets and evaluated retained features in CV framework, until <50 features remained. On *y*-axis, median and minimum ICC values of radiomic feature sets and models' corresponding averaged CV AUC scores are plotted, with *x*-axis showing number of retained features.

different intensity-normalized image type. Considering that some variability in XGBoost performance is expected, we concluded that all machine-learning models generalized well to the independent and external validation cohorts (Table 2). Notably, all but 1 of the XGBoost validation AUCs lie within the range of ±1 SD from the corresponding CV AUC. The AUC differences between models based on different intensity-normalization techniques were not significant. However, tendencies from univariate analysis were generally replicated in machine-learning analysis in the training cohort, with the lentiform nucleus model attaining the numerically highest AUC in the CV (Table 2), and the *P* value from comparing the lentiform nucleus– and SUV-normalized models approached significance ($P = 0.052$; Supplemental Fig. 4.2). Notably, the model based on raw PET images (i.e., voxel intensities indicate Bq/mL) achieved an AUC ± SD of 0.67 ± 0.08 in the CV and the numerically highest AUC in independent validation. This finding suggests that machine learning–generated radiomic biomarkers are relatively robust even in the absence of ^{18}F-FDG PET voxel intensity normalization. For exploratory radiomics-based machine-learning research, it may therefore be appropriate to forgo intensity normalization under certain circumstances (provided that voxel intensities are converted to identical units, for example, Bq/mL).

Univariate association analysis with the HPV status revealed a limited effect of intensity normalization on the predictive capacity of individual features, with normalization to the lentiform nucleus yielding the numerically highest median regression coefficient and AUC score and the largest fraction of significant features in a logistic regression analysis (Table 1; Supplemental Figs. 2.1–2.2). This effect was pervasive across feature families and image types (Supplemental Figs. 2.3–2.6; Supplemental Table 7). These tendencies suggest that the choice of normalization technique may affect the radiomic features' predictive value with respect to HPV and that normalizing ^{18}F-FDG PET voxel intensities to the lentiform nucleus may be preferable for head-and-neck radiomics. However, more evidence is needed to draw a firm conclusion, including investigation of additional reference tissues and studies with higher statistical power. We found no relevant impact of the features' reproducibility across normalization techniques on their predictive capacity with respect to HPV (Supplemental Fig. 3).

We hypothesized that a measure of feature robustness against differences in PET normalization could be a useful feature selection criterion for machine-learning models. If models achieved good performance with only features known to be highly robust, uniform image normalization as a preprocessing step may be omitted. To investigate this, we iteratively excluded features with the lowest ICC scores from machine-learning models. We observed a steady decline in XGBoost performance as the number of available candidate features decreased, with no plateau identified that provided a satisfactory balance between classifier performance and feature reproducibility (Fig. 2). Consequently, we regarded the hypothesis as false.

The diversity of our dataset, which was acquired from multiple institutions including Yale School of Medicine (*15*), comes with inherent advantages and disadvantages. The variability in PET scanners and in imaging and reconstruction protocols may bolster the generalizability of our results, whereas the noise introduced by this variability may reduce the granularity of some findings. Future studies may gather larger, more homogeneous datasets—ideally by following clinical practice guidelines designed to improve the repeatability and reproducibility of oncologic PET (*31*)—to enable a clearer understanding of the impact of PET normalization on radiomic features. Future work should additionally investigate the liver and mediastinal blood pool, given their widespread use as ^{18}F-FDG uptake references. In addition, further tumor entities, radiotracers, SUV variants, and segmentation techniques should be considered. The extent to which normalization affects individual feature values is partly inherent in the features' mathematic definitions. Therefore, the rank order of the features' reproducibility across normalization techniques is generalizable to other reference tissues, SUV variants, tissues of interest, or radiotracers, if similar effects of image filtering are assumed. Notably, absolute ICC values are more dependent on absolute voxel intensities and are therefore less generalizable. We expect similar effects of PET normalization on the features' predictive value in future studies on other cancers and tissues of interest, reference tissues, clinical end points, and tracers. However, the generalizability of these analyses depends on an array of additional factors, including the predictability and steadiness of tracer uptake in new reference tissues, feature selection for predictive models, and the dynamics and interaction of radiotracer uptake in the reference tissue and tissues of interest, which may include malignant and benign lesions or healthy tissue. This complexity impedes prognoses and warrants further studies.

CONCLUSION

To the best of our knowledge, our study is the first to investigate the effects of ^{18}F-FDG PET voxel intensity normalization—using the SUV and reference tissue–standardized uptake ratios—on radiomic features and machine learning–generated radiomic biomarkers. Although radiomics researchers thus far have converted PET imagery to SUV maps, reasoning that a clinically established normalization technique would be equally effective in radiomics pipelines, our study is the first to our knowledge to provide data and empiric evidence. We found that few features were reproducible across intensity-normalization techniques, making uniform normalization (or at least conversion to identical units, for example, Bq/mL) a methodologic prerequisite for interindividual comparability of radiomic biomarkers. In machine-learning models and univariate association analyses with the HPV status, we discovered trends suggesting that the choice of normalization technique may affect the radiomic features' predictive values and that normalization to a central nervous system reference, specifically the lentiform nucleus of the brain, may be preferable. However, more evidence is needed before a definitive recommendation can be made.

KEY POINTS

QUESTION: How does ^{18}F-FDG PET voxel intensity value normalization (SUV vs. reference tissue standardized uptake ratios vs. raw intensities) affect the radiomic features of OPSCC and machine learning–generated radiomic biomarkers?

PERTINENT FINDINGS: Using pretreatment ^{18}F-FDG PET scans of 430 OPSCC patients, we show that only approximately one third of the 1,037 radiomic features attains a high degree of reproducibility across normalization techniques, highlighting the necessity for uniform intensity normalization to ensure interindividual comparability of radiomic markers. In addition, normalization may affect the features' predictive value. We observed tendencies suggesting that normalization to a neural reference tissue (lentiform nucleus) may be preferable for HPV prediction in univariate and XGBoost machine-learning analyses, but more evidence is needed to support this conclusion.

IMPLICATIONS FOR PATIENT CARE: Our study begins to elucidate the impact of ^{18}F-FDG PET intensity normalization on radiomic markers of OPSCC and machine learning–generated radiomic biomarkers, potentially expediting the translation of exploratory radiomic tools into clinical applications.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We thank Professor Andre Dekker, PhD, and Associate Professor Leonard Wee, PhD (MAASTRO Clinic, Maastricht, The Netherlands), for providing data allowing us to generate SUV maps of PET imagery from the MAASTRO cohort.

REFERENCES

1. Vriens D, Visser EP, de Geus-Oei LF, Oyen WJ. Methodological considerations in quantification of oncological FDG PET studies. *Eur J Nucl Med Mol Imaging.* 2010;37:1408–1425.
2. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to radiomics. *J Nucl Med.* 2020;61:488–495.
3. Tortora M, Gemini L, Scaravilli A, et al. Radiomics applications in head and neck tumor imaging: a narrative review. *Cancers (Basel).* 2023;15:1174.
4. Yusufaly TI, Zou J, Nelson TJ, et al. Improved prognosis of treatment failure in cervical cancer with nontumor PET/CT radiomics. *J Nucl Med.* 2022;63:1087–1093.
5. Eertink JJ, van de Brug T, Wiegers SE, et al. ^{18}F-FDG PET baseline radiomics features improve the prediction of treatment outcome in diffuse large B-cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2022;49:932–942.
6. Leijenaar RT, 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.
7. 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.
8. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *Eur J Nucl Med Mol Imaging.* 2019;46:2638–2655.
9. Ger RB, Meier JG, Pahlka RB, et al. Effects of alterations in positron emission tomography imaging parameters on radiomics features. *PLoS One.* 2019;14:e0221877.
10. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys.* 2018;102:1143–1158.
11. Gillison ML, Chaturvedi AK, Anderson WF, Fakhry C. Epidemiology of human papillomavirus–positive head and neck squamous cell carcinoma. *J Clin Oncol.* 2015;33:3235–3242.
12. Britz-Cunningham SH, Millstine JW, Gerbaudo VH. Improved discrimination of benign and malignant lesions on FDG PET/CT, using comparative activity ratios to brain, basal ganglia, or cerebellum. *Clin Nucl Med.* 2008;33:681–687.
13. Helsen N, Van den Wyngaert T, Carp L, et al. Quantification of ^{18}F-fluorodeoxyglucose uptake to detect residual nodal disease in locally advanced head and neck squamous cell carcinoma after chemoradiotherapy: results from the ECLYPS study. *Eur J Nucl Med Mol Imaging.* 2020;47:1075–1082.
14. van den Bosch S, Dijkema T, Philippens MEP, et al. Tumor to cervical spinal cord standardized uptake ratio (SUR) improves the reproducibility of ^{18}F-FDG-PET based tumor segmentation in head and neck squamous cell carcinoma in a multicenter setting. *Radiother Oncol.* 2019;130:39–45.
15. Haider SP, Mahajan A, Zeevi T, et al. PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma. *Eur J Nucl Med Mol Imaging.* 2020;47:2978–2991.
16. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep.* 2017;7:10117.
17. Grossberg AJ, Mohamed ASR, Elhalawani H, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Sci Data.* 2018;5:180173.
18. Zuley ML, Jarosz R, Kirk S, et al. The cancer genome atlas head-neck squamous cell carcinoma collection (TCGA-HNSC). Cancer Imaging Archive website. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=11829589. Updated August 3, 2023. Accessed February 29, 2024.

19. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006.
20. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging.* 2012;30:1323–1341.
21. Kinahan P, Clunie D, Boellaard R, et al. Vendor-neutral pseudo-code for SUV calculation. QIBAwiki website. https://qibawiki.rsna.org/index.php/Standardized_Uptake_Value_(SUV). Updated June 26, 2018. Accessed February 29, 2024.
22. Pyradiomics documentation release v3.0.1. Pyradiomics Community website. https://pyradiomics.readthedocs.io/_/downloads/en/v3.0.1/pdf/. Published April 12, 2021. Accessed February 29, 2024.
23. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 2017;77:e104–e107.
24. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* 2020;295:328–338.
25. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15:155–163.
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Cornell University website. https://arxiv.org/abs/1603.02754. Published March 9, 2016. Accessed February 29, 2024.
27. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics.* 2013;29:2365–2368.
28. rBayesianOptimization, version 1.1.0. R Archive Network website. https://cran.r-project.org/web/packages/rBayesianOptimization/rBayesianOptimization.pdf. Published October 14, 2022. Accessed February 29, 2024.
29. R Core Team. R: a language and environment for statistical computing; R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Published April 26, 2019. Accessed February 29, 2024.
30. Bouckaert RR, Frank E. Evaluating the replicability of significance tests for comparing learning algorithms. In: Karlapalem K, Cheng H, Ramakrishnan N, et al., eds. *Advances in Knowledge Discovery and Data Mining.* Springer; 2004:3–12.
31. Boellaard R, Delgado-Bolton R, Oyen WJ, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging.* 2015;42:328–354.

Errata

In the article "^{177}Lu-PSMA SPECT Quantitation at 6 Weeks (Dose 2) Predicts Short Progression-Free Survival for Patients Undergoing ^{177}Lu-PSMA-I&T Therapy," by John et al. (*J Nucl Med.* 2023;64:410–415), the first sentence of the Discussion on page 413 currently reads "… between baseline and 6 wk of ^{177}Lu-PSMA-617 therapy … " The statement should read "… between baseline and 6 wk of *^{177}Lu-PSMA-I&T* therapy … " The authors regret the error.

In the article "[^{18}F]FDG PET/CT in the Initial Staging and Restaging of Soft-Tissue or Bone Sarcoma in Patients with Negative or Equivocal Findings for Metastases or Limited Recurrence on Conventional Work-up: Results of a Prospective Multicenter Registry," by Metser et al. (*J Nucl Med*. 2023;64:1371–1377), the last author's surname was spelled incorrectly. Singunkar should be *Singnurkar*. The authors regret the error.