Journal of Nuclear Medicine, published on July 16, 2021 as doi:10.2967/jnumed.121.262117

Quantitative radiomics features in diffuse large B-cell lymphoma: does segmentation method matter?

Jakoba.J. Eertink[1], Elisabeth.A.G. Pfaehler[2], Sanne.E. Wiegers[1], Tim van de Brug[3], Pieternella.J. Lugtenburg[4], Otto.S. Hoekstra[5], Josée.M. Zijlstra,[1] Henrica.C.W. de Vet[3], Ronald Boellaard[5]

[1]Amsterdam UMC, Vrije Universiteit Amsterdam, department of Hematology, Cancer Center Amsterdam, Amsterdam, Netherlands
[2]Department of Nuclear Medicine, University Hospital Augsburg, Augsburg, Germany
[3]Amsterdam UMC, Vrije Universiteit Amsterdam, department of Epidemiology and Data Science, Amsterdam Public Health research institute, Amsterdam, Netherlands
[4]Erasmus MC Cancer Institute, University Medical Center Rotterdam, department of Hematology, Rotterdam, Netherlands
[5]Amsterdam UMC, Vrije Universiteit Amsterdam, department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Amsterdam, Netherlands

Corresponding author:
Ronald Boellaard
r.boellaard@amsterdamumc.nl

First author:
Jakoba Eertink, PhD-student
j.eertink@amsterdamumc.nl

Amsterdam UMC-location VUmc
De Boelelaan 1117, 1081HV Amsterdam, Netherlands
Tel:+31(0)204449638
fax:+31(0)204444329

Funding:
Dutch Cancer Society (#VU-2018-11648) and partially by the research program STRaTeGy (14929)

word count: 5322
short running title: Influence of segmentation on radiomics

ABSTRACT

Introduction: Radiomics features may predict outcome in diffuse large B-cell lymphoma (DLBCL). Currently, multiple segmentation methods are used to calculate metabolic tumor volume (MTV). We assessed the influence of segmentation method on the discriminative power of radiomics features in DLBCL at patient level and for the largest lesion. **Methods:** 50 baseline ^{18}F-fluorodeoxyglucose positron emission tomography computed tomography (PET/CT) scans of DLBCL patients who progressed or relapsed within 2 years after diagnosis were matched on uptake time and reconstruction method with 50 baseline PET/CT scans of DLBCL patients without progression. Scans were analysed using 6 semi-automatic segmentation methods (standardized uptake value (SUV)4.0, SUV2.5, 41% of the maximum SUV, 50% of the SUV_{peak}, majority vote (MV)2 and MV3, respectively). Based on these segmentations, 490 radiomics features were extracted at patient level and 486 features for the largest lesion. To quantify the agreement between features extracted from different segmentation methods, the intra-class correlation (ICC) agreement was calculated for each method compared to SUV4.0. The feature space was reduced by deleting features that had high Pearson correlations (≥0.7) with the previously established predictors MTV and/or SUV_{peak}. Model performance was assessed using stratified repeated cross-validation with 5 folds and 2000 repeats yielding the mean receiver-operating characteristics curve integral (CV-AUC) for all segmentation methods using logistic regression with backward feature selection. **Results:** The percentage of features yielding an ICC ≥0.75 compared to the SUV4.0 segmentation was lowest for A50P both at patient level and for the largest lesion, with 77.3% and 66.7% of the features yielding an ICC ≥0.75, respectively. Features were not highly correlated with MTV, with at least 435 features at patient level and 409 features for the largest lesion for all segmentation methods with a correlation coefficient <0.7. Features were highly correlated with SUV_{peak} (at least 190 and 134 were uncorrelated, respectively). CV-AUCs ranged between 0.69±0.11 and 0.84±0.09 at patient level, and between 0.69±0.11 and 0.73±0.10 for

lesion level. **Conclusion:** Even though there are differences in the actual radiomics feature values derived and selected features between segmentation methods, there is no substantial difference in the discriminative power of radiomics features between segmentation methods.

Keywords

Diffuse-large-B-cell-lymphoma, Segmentation methods, Radiomics, ^{18}F-FDG-PET/CT.

INTRODUCTION

Diffuse large B cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin lymphoma. To improve outcome of patients with DLBCL, early identification of patients at risk of treatment failure is of utmost importance, as 25-40% of patients relapse or progress in the first years after diagnosis (*1*). Recent data suggest that baseline radiomic features are promising biomarkers to predict treatment outcome in DLBCL (*2-4*), as they can predict outcome beyond metabolic tumor volume (MTV), and the international prognostic index (IPI) (*5*).

Radiomic features can be calculated from the baseline ^{18}F-fluorodeoxyglucose positron emission tomography computed tomography (^{18}F-FDG-PET/CT) scans and capture detailed and quantitative information on e.g. texture, intensity and shape of lesions. Currently, radiomics analyses in lymphoma are based on predefined tumor segmentations. Segmentations are usually performed using an absolute Standardized Uptake Value (SUV) thresholds (*6*) or percentages of SUV_{max} or SUV_{peak} (*2,7*). For the calculation of radiomics features, some studies use the hottest lesion (*4*), whereas others use the largest lesion (*3,8*) or tumor segmentations at patient level (*2,9*). The largest lesion and MTV at patient level had the highest predictive value (*9*). Therefore, in this study we concentrated on the largest lesion and radiomic features extracted from tumor segmentations at patient level.

One of the main problems with generating a multitude of features is the high false detection rate caused by multiple testing. Moreover, several features may represent similar characteristics, that are often highly correlated and therefore redundant (*10*). Redundant features may induce a correlation bias (*11*) and models become difficult to interpret (*12*).

Therefore, reducing the feature space to a degree feasible for clinical use without losing important information is essential. One method to reduce feature space is hierarchical clustering, based on correlation analysis or distance metrics (*13*).

Previous DLBCL studies showed that MTV measured with different segmentation methods, albeit at different cut-offs, showed comparable discriminative power to predict survival (*6,7*). However, it is unclear to which extent the discriminative power of other radiomic features is affected by the method used to segment the lesions. Therefore, our main objective was to assess the effects of using six frequently used segmentation methods on the discriminative power for 2-year time to progression (TTP) of baseline PET/CT radiomics features in DLBCL both at patient level and for the largest lesion.

MATERIALS AND METHODS

Study Population

For this case-control study 100 newly diagnosed DLBCL patients from the HOVON-84 study (EudraCT: 2006-005174-42) with a baseline PET/CT-scans available were included. 50 patients with progressive disease or relapse within 2 years after diagnosis were matched on scan interval and reconstruction method (EARL/non-EARL)(*14*) with 50 patients without progression. For this analysis we combined R-CHOP14 and RR-CHOP14, as outcomes were similar between treatment arms (*15*). The HOVON-84 study was approved by the institutional review board and all participants gave informed consent.

Quantitative Analysis

Quantitative PET/CT analysis was performed using the quantitative oncology molecular analysis suite (ACCURATE) (*16*). To match quality criteria, PET and low dose CT scans should

be complete, and liver SUV_{mean} and plasma glucose within ranges suggested by the European association of Nuclear Medicine guidelines (*14*). If liver SUV_{mean} was outside suggested ranges, but total image activity was between 50-80% of the injected activity scans were still included. All scans were reviewed by nuclear medicine physicians and delineations were performed under their supervision. The following frequently used semi-automatic segmentation methods were applied to delineate lesions:

1. SUV threshold of 2.5 (SUV2.5)
2. SUV threshold of 4.0 (SUV4.0)
3. 50% of SUV_{peak} (A50P) (*17*)
4. 41% of SUV_{max} (41%max)
5. Majority vote segmenting voxels detected by ≥2 methods (MV2)
6. Majority vote segmenting voxels detected by ≥3 methods (MV3) (Supplemental data)

Lesions were delineated with a fully automated preselection of lesions with a volume threshold of ≥3mL. Lymphoma lesions <3mL were added by observer selection and non-tumor regions were deleted with single mouse-clicks for all 6 segmentation methods (*18*). Automatically successfully segmented lesions were added to the patient level volume of interest (VOI). If lesion selection resulted in flooding (i.e. selection of large parts of non-tumor regions: e.g. liver, spleen and/or skeleton), the lesion was not added. Adjacent non-tumor ^{18}F-FDG avid regions (e.g. bladder, kidney) were manually removed. For the fixed SUV4.0 method, we also generated segmentations with a volume threshold of ≥ 3mL (SUV4.0(≥3mL)). Two observers selected the method with the highest visual agreement (best method) for each patient, resolving initial discrepancies in consensus meetings.

Feature Extraction

480 radiomics features (texture (n=408), morphology (n=22), intensity-based statistics (n=18), intensity histogram (n=24), intensity-volume histogram (n=6) and local intensity (n=2)) and 6 conventional PET uptake metrics before rebinning were extracted for both patient level, and the largest lesion for each segmentation method. The patient level VOI included all segmented lesions and was generated by assigning all voxels within the individual lesions to one and all voxels outside any of the segmented individual lesions to zero. At patient level, 4 additional dissemination features were calculated. All image-processing and feature calculations were performed using RaCat software (*19*), which complies with the imaging biomarker standardisation initiative criteria (*20*). Details regarding feature calculation are presented in the supplemental data.

Statistical Analysis

All statistical analyses were performed for radiomics features at patient level and for the largest lesion using R (version 4.0.3). The paired student t-test was used to compare the MTV and SUV_{peak} of all segmentation methods compared to the best segmentation. Based on recent studies, the SUV4.0 segmentation was chosen as reference (*7,18*). Firstly, if the distribution of the radiomics feature values had skewness >0.5 for the SUV4.0 segmentation method, they were log-transformed for all segmentations using the natural logarithm. The agreement between radiomics features extracted from different segmentations was quantified by calculating the intra-class correlation (ICC) agreement compared to the SUV4.0 segmentation. ICCs were categorized as poor (ICC<0.5), moderate (ICC:0.5-0.74), good (ICC:0.75-0.89) or excellent reliability (ICC≥0.90) (*21*). Two texture features at patient level, and three texture features at lesion level did not show any variation and were therefore excluded.

MTV and SUV_{peak} have shown to be predictive in DLBCL (*9*). To avoid overfitting and to remove redundancy, the feature space was reduced by deleting features that highly correlated with either MTV and/or SUV_{peak}. The Pearson correlation coefficient between MTV and other radiomics features, and between SUV_{peak} and other radiomics features was calculated for each segmentation method. A correlation was considered high if the Pearson correlation coefficient was ≥0.7 (*22*).

For each segmentation method the mutual correlations between features that were not correlated with MTV and SUV_{peak} were calculated using Pearson correlation. For clusters of features with high mutual correlations, as identified with hierarchical clustering using Euclidian distance as distance measure, the feature with the lowest correlation to MTV and/or SUV_{peak} was preserved.

Discriminative power (progression versus non-progression) was assessed using logistic regression with backward feature selection based on the Akaike Information Criteria (*23*). We included all independent features, MTV and SUV_{peak} for all segmentations. Stratified repeated cross-validation with 5 folds and 2000 repeats was applied, yielding the mean receiver-operating-characteristics curve integral (CV-AUC), and the standard deviation of AUCs between repeats. Comparing CV-AUCs is a known difficulty due to the inherent dependency of train-test iterations and complex relations between the trained models (*24*). Currently, there is no valid statistical approach to compare CV-AUCs.

As a sensitivity analysis, the entire analyses were repeated for features that were reliable, repeatable and reproducible in a multi-center setting (*25*).

RESULTS

Patient characteristics are summarized in Table 1. 64 Scans were semi-automatically analysed and adapted with a single mouse-clicks only. 36 scans required manual editing because tumor and non-tumor regions were adjacent. SUV4.0 was selected most frequently as best method for both patient level and lesional level (49% and 64%, respectively).

MTV Analysis

The SUV2.5 method resulted in MTV flooding for 44 patients, leading to exclusion of this method for further analysis. At patient and lesion levels, MTV was highest for the MV2, and lowest for the A50P segmentation method (Table 2). Using the best visual segmentation as reference, MTV was significantly higher using the MV2 segmentation, and significantly lower using all other segmentation methods (all: $p<0.05$, Table 2; Figure 1). SUV_{peak} was comparable between segmentation methods (all: $p>0.05$).

Patient Level

Radiomics features based on a SUV4.0 preselection with a 3mL volume threshold (SUV4.0(≥3mL)) resembled the features of the SUV4.0 segmentation most, with excellent reliability for 414 features (84.8%), followed by the best segmentation. For the A50P segmentation similarity was lowest, with only 218 features (44.7%) with excellent reliability (Figure 2, Supplemental Table 1).

For all segmentation methods, at least 435 features (89.3%) were not highly correlated with MTV (Table 3), of which 433 (88.9%) were not highly correlated with MTV for all segmentation methods. At least 190 features (38.9%) were not highly correlated with SUV_{peak}, of which 175 (35.9%) were not highly correlated with SUV_{peak} for all segmentations. 197 features (40.5%) were

not correlated with MTV and SUV_{peak} for at least one method, of which 125 (25.7%) were neither correlated with MTV nor with SUV_{peak} for all segmentation methods. For each segmentation method, at least 25 features (5.1%) did not show high mutual correlations, and were not correlated with MTV and SUV_{peak}. After backward feature selection, the SUV4.0 segmentation method yielded a CV-AUC 0.74±0.10; 41%max had the highest CV-AUC (0.84±0.09), the visually best segmentation method had the lowest CV-AUC (0.69±0.11). Selected features after backward selection differed between segmentation methods and varied between 4-20 features (Table 3, Supplemental Table 2). For all segmentation methods, the morphological feature 'center of mass shift' and the texture feature 'first measure of information correlation' were retained in the linear regression model.

Largest Lesion

Radiomics features of the MV2 segmentation resembled those of the SUV4.0 method most, with excellent reliability for 389 features (80.5%). For the A50P segmentation similarity was lowest, at only 83 features (17.2%) with excellent reliability (Figure 3, Supplemental Table 3).

For all segmentations, at least 409 features (84.9%) were not highly correlated with MTV (Table 4), of which 404 (83.8%) were not highly correlated with MTV for all segmentation methods. At least 134 features (27.8%) were not highly correlated with SUV_{peak}, of which 130 features (27.0%) were not highly correlated with SUV_{peak} for all segmentations. 149 (31.0%) features were not correlated with MTV and SUV_{peak} for at least one method, of which 61 features (12.7%) were neither correlated with MTV nor with SUV_{peak} for all segmentation methods. For each segmentation method, at least 19 features (4.0%) did not show high mutual correlations and were not correlated with MTV and SUV_{peak}. After backward feature selection, SUV4.0 had the highest CV-AUC (0.73±0.10), MV3 and the best segmentation method had the lowest CV-AUC (0.69±0.11). Selected features after backward selection differed between segmentation methods

and varied between 5-11 features (Table 4, Supplemental Table 4). For all segmentation methods, the texture feature 'first measure of information correlation' was retained in the linear regression model, and the intensity histogram feature 'minimum histogram gradient' was retained in all models except for the SUV4.0 segmentation method.

When starting from a selection with reliable, repeatable and reproducible features, similar results were found both at patient level and for the largest lesion (Table 3, Table 4).

DISCUSSION

This study showed that the discriminative power is largely independent of segmentation method. However, there are large differences in radiomics feature values derived using different segmentation methods, as shown by $ICC_{agreement}$ values.

Both MTV and SUV_{peak} have shown to be predictive in DLBCL (*9*). Our study showed that most radiomics features are independent of MTV for both patient level and the largest lesion. Hatt et al (*26*) showed that textural features, which comprise >80% of our radiomics features, already provide clinical complementary information in addition to MTV in lesions larger than 10mL, with increasing complementary prognostic value for larger MTVs, disputing the threshold for texture features of 45mL (*27*). With only four patients with MTVs <10mL for the largest lesion, and one patient with a MTV <10mL at patient level it is to be expected that most features are independent of MTV. However, many features were correlated with SUV_{peak}, c.q. redundant.

Currently, there is no consensus on the best segmentation method for delineating lesions in DLBCL ^{18}F-FDG PET/CT studies. Therefore, it is essential to study the sensitivity of radiomics features in relation to segmentation method. In several solid cancers radiomics features, especially morphological and texture features, are influenced by the delineation method (*28-31*). The number of extracted features in these studies varied widely, between 9 and 480 features. We

extend these findings by showing that for the largest lesion in DLBCL, up to 31% of the texture features, and 68% of the morphological features were highly sensitive to the segmentation method, as shown by the reliability of features compared to SUV4.0 segmentation. DLBCL lesions usually are large, heterogeneous and bulky. Larger lesions are known to exhibit higher hypoxia, necrosis, or anatomical and physiological complexity which logically translates to higher complexity in the spatial ^{18}F-FDG distribution and hence sensitivity to segmentation method leading to lower reliability of features between applied methods. Furthermore, as variations in segmentation methods have a strong effect on the outercontour of the segmentation, thus influencing the shape of the segmentation, a high sensitivity to segmentation methods for morphological features could be expected. Due to the higher MTV, the radiomics features at patient level were less influenced by segmentation method, with up to 20% of the texture features, and 32% of the morphology features being sensitive to segmentation method. Because of low similarity of part of the features between segmentations, it is not advised to use regression coefficients from other studies that applied other segmentation methods.

However, even though values are not interchangeable, in our study the discriminative power at lesional and patient levels was comparable between segmentations. Contrary to what we expected, selecting the segmentation method that visually selected the tumors best, did not result in a higher CV-AUC. These results are in line with previous studies exploring the predictive value of radiomics features using different segmentations for other cancer types. These studies all found no significant differences in predicting outcome (*28,32*), metastasis or lymph node invasion (*30*) using different segmentation methods. However, $ICC_{agreement}$ values, correlations with MTV, SUV_{peak} and mutual correlations differed between segmentation methods, resulting in different preselections of features for the logistic regression model. Even though discriminative power is comparable, different features are predictive of outcome when applying different segmentation methods.

When only using previously defined reliable, repeatable and reproducible features, discriminative power was slightly lower for all segmentation methods. However, confidence intervals of CV-AUCs overlapped with the CV-AUCs using all features. Therefore, using only reproducible features does not affect discriminative power. In clinical practice and multicenter studies variable image qualities are encountered. Therefore some features that have high predictive values may in reality be difficult to measure reliably. Therefore, it is advised to only use reproducible features, especially in multicenter settings.

To our knowledge, this is the first study that assessed the influence of segmentation methods on PET radiomic features and their predictive power, other than MTV, in DLBCL. By applying multiple frequently used methods on the same patients, we could directly compare the effect of segmentation methods on quantitative PET radiomic features. We chose to calculate linear relations between radiomics features using Pearson because we used logistic regression as classifier, and logistic regression model calculates linear relations with included features. By using Pearson for data reduction, probably more features were included in the logistic regression analyses compared to the number of features we would have included when using Spearman. One of the limitations of this study was that not all scans were scanned according to EARL protocol, which might affect the discriminative power and repeatability of features (*25*). However, by matching events in this study there were no differences in EARL compliance between groups, but this matching still precludes an effect of reconstruction method on the discriminative power. Harmonization methods such as ComBat have shown to be definitely worthwhile to retrospectively increase uniformity in large datasets (*33,34*). Therefore, ComBat based data-alignment would be a very successful approach to harmonize these differences. Unfortunately, in our study the number of patients per center was too small to apply ComBat. Moreover, based on the equivalent discriminative power seen in our data between various segmentation methods ComBat based data-alignment would be a very successful approach to harmonize databases of radiomics

features analysed using different segmentation methods. Moreover, it should be noted that in our cohort patients presented with high MTVs. Therefore these results need to be validated for other cohorts with smaller lesion sizes.

CONCLUSION

This study showed that there is no substantial difference in the discriminative performance of radiomics features extracted using different segmentation methods. However, there are differences in the actual radiomics feature values derived and selected features between segmentation methods. Until consensus on a segmentation method for DLBCL is reached, it is advised to only use prediction models that are built using data with the same segmentation methods.

DISCLOSURE

This work was financially supported by the Dutch Cancer Society (#VU-2018-11648). This work is partially supported by the research program STRaTeGy (14929) which is financed by the Netherlands Organisation for Scientific Research. No potential conflicts of interest relevant to this article exist

KEY POINTS

Question: What is the influence of segmentation methods on the discriminative power value of baseline radiomics features in DLBCL?

Pertinent findings: There is no difference in the discriminative power of radiomics features between segmentation methods. However, different features are selected when applying different segmentation methods.

Implications for patient care: It is advised to only use prediction models that are build using data with the same segmentation methods.

REFERENCES

1. Crump M, Neelapu SS, Farooq U, et al. Outcomes in refractory diffuse large B-cell lymphoma: results from the international SCHOLAR-1 study. *Blood.* 2017;130:1800-1808.

2. Cottereau AS, Nioche C, Dirand AS, et al. (18)F-FDG PET dissemination features in diffuse large B-Cell lymphoma are predictive of outcome. *J Nucl Med.* 2020;61:40-45.

3. Aide N, Fruchart C, Nganoa C, Gac AC, Lasnon C. Baseline (18)F-FDG PET radiomic features as predictors of 2-year event-free survival in diffuse large B cell lymphomas treated with immunochemotherapy. *Eur Radiol.* 2020;30:4623-4632.

4. Ceriani L, Gritti G, Cascione L, et al. SAKK38/07 study: integration of baseline metabolic heterogeneity and metabolic tumor volume in DLBCL prognostic model. *Blood Adv.* 2020;4:1082-1092.

5. International Non-Hodgkin's Lymphoma Prognostic Factors P. A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med.* 1993;329:987-994.

6. Ilyas H, Mikhaeel NG, Dunn JT, et al. Defining the optimal method for measuring baseline metabolic tumour volume in diffuse large B cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2018;45:1142-1154.

7. Barrington SF, Zwezerijnen BG, de Vet HC, et al. Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful ? *J Nucl Med.* 2021;62:332-337.

8. Senjo H, Hirata K, Izumiyama K, et al. High metabolic heterogeneity on baseline 18FDG-PET/CT scan as a poor prognostic factor for newly diagnosed diffuse large B-cell lymphoma. *Blood Adv.* 2020;4:2286-2296.

9. Eertink JJ, van de Brug T, Wiegers SE, et al. 18F-FDG PET/CT baseline radiomics features are predictive of outcome in diffuse large B- cell lymphoma patients - European Association of Nuclear Medicine October 22 – 30, 2020 Virtual*. Eur J Nucl Med Mol Imaging* 2020;47:1-753.

10. Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J Nucl Med.* 2014;55:414-422.

11. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics.* 2011;27:1986-1994.

12. 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.

13. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging.* 2012;30:1234-1248.

14. 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.

15. Lugtenburg PJ, de Nully Brown P, van der Holt B, et al. Rituximab-CHOP with early rituximab intensification for diffuse large B-cell lymphoma: a randomized phase III trial of the HOVON and the Nordic Lymphoma Group (HOVON-84). *J Clin Oncol.* 2020;38(29):3377-3387.

16. Boellaard R. Quantitative oncology molecular analysis suite: ACCURATE. *J Nucl Med.* 2018;59:1753.

17. Frings V, van Velden FH, Velasquez LM, et al. Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study. *Radiology.* 2014;273:539-548.

18. Burggraaff CN, Rahman F, Kassner I, et al. Optimizing workflows for fast and reliable metabolic tumor volume measurements in diffuse large B cell lymphoma. *Mol Imaging Biol.* 2020;22:1102-1110.

19. 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.

20. Zwanenburg A, Vallieres M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* 2020;295:328-338.

21. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163.

22. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J.* 2012;24:69-71.

23. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, eds. *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York; 1998:199-213.

24. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 1998;10:1895-1923.

25. Pfaehler E, van Sluis J, Merema BBJ, et al. Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-dimensionally printed phantom inserts. *J Nucl Med.* 2020;61:469-476.

26. Hatt M, Majdoub M, Vallieres M, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med.* 2015;56:38-44.

27. Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. *J Nucl Med.* 2014;55:37-42.

28. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour (1)(8)F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging.* 2013;40:1662-1671.

29. Belli ML, Mori M, Broggi S, et al. Quantifying the robustness of [(18)F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. *Phys Med.* 2018;49:105-111.

30. Cysouw MCF, Jansen BHE, van de Brug T, et al. Machine learning-based analysis of [(18)F]DCFPyL PET radiomics for risk stratification in primary prostate cancer. *Eur J Nucl Med Mol Imaging.* 2021;48:340-349.

31. 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.

32. 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.

33. Orlhac F, Boughdad S, Philippe C, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med.* 2018;59:1321-1328.

34. Dissaux G, Visvikis D, Da-Ano R, et al. Pretreatment (18)F-FDG PET/CT Radiomics predict local recurrence in patients treated with stereotactic body radiotherapy for early-stage non-small cell lung cancer: a multicentric study. *J Nucl Med.* 2020;61:814-820.

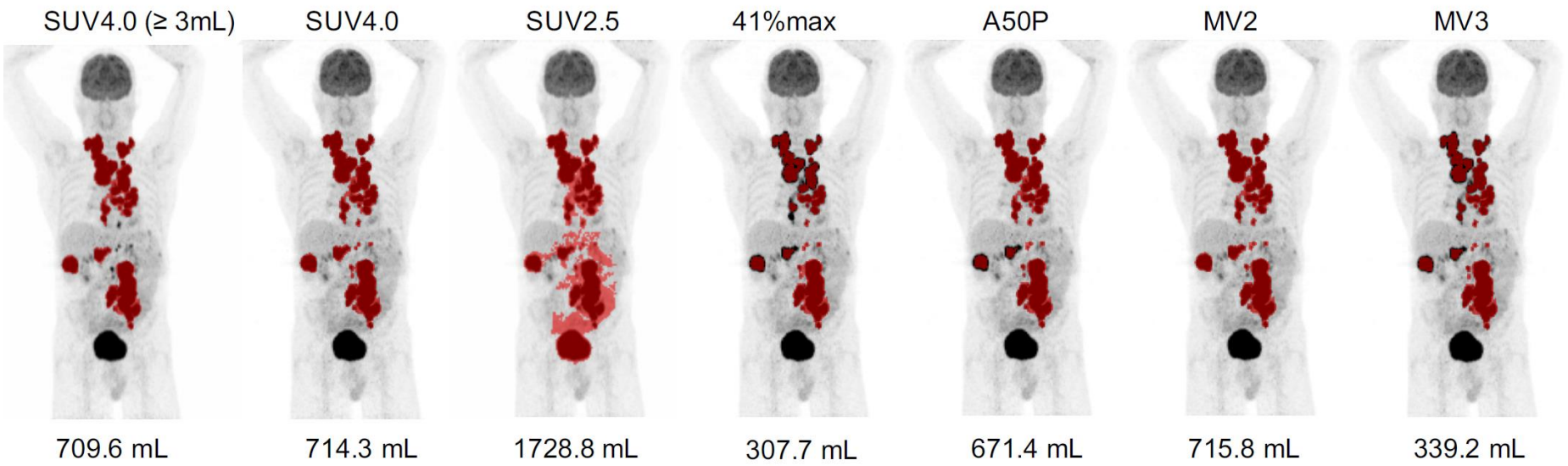


Figure 1. Maximum intensity projections of a patient with lesion segmentations indicated in red for al applied methods using a SUV0-10 scale.

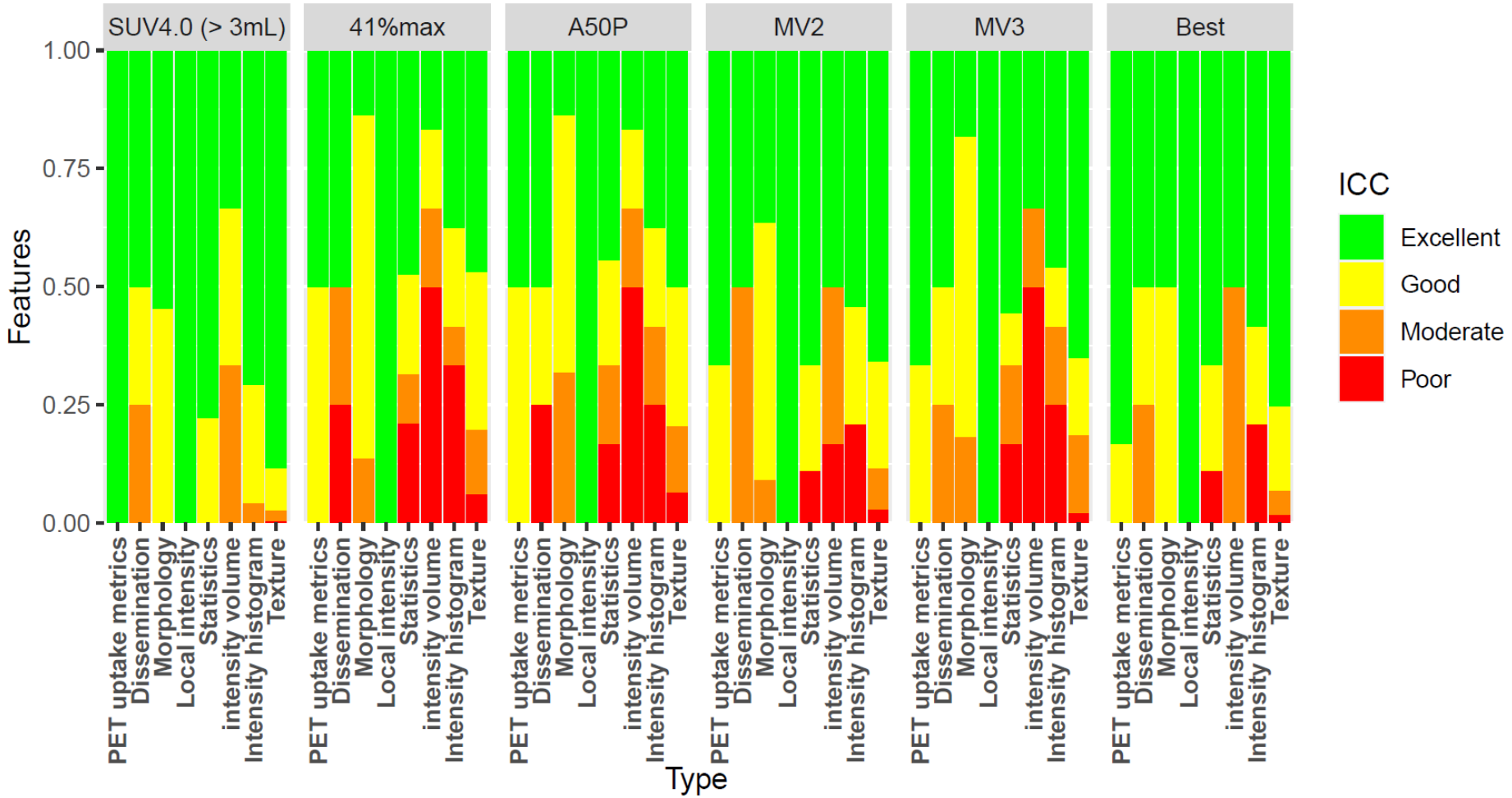


Figure 2. Percentage of radiomics features yielding excellent, good, moderate or poor intraclass correlation agreement between the SUV4.0 segmentation and other methods at patient level

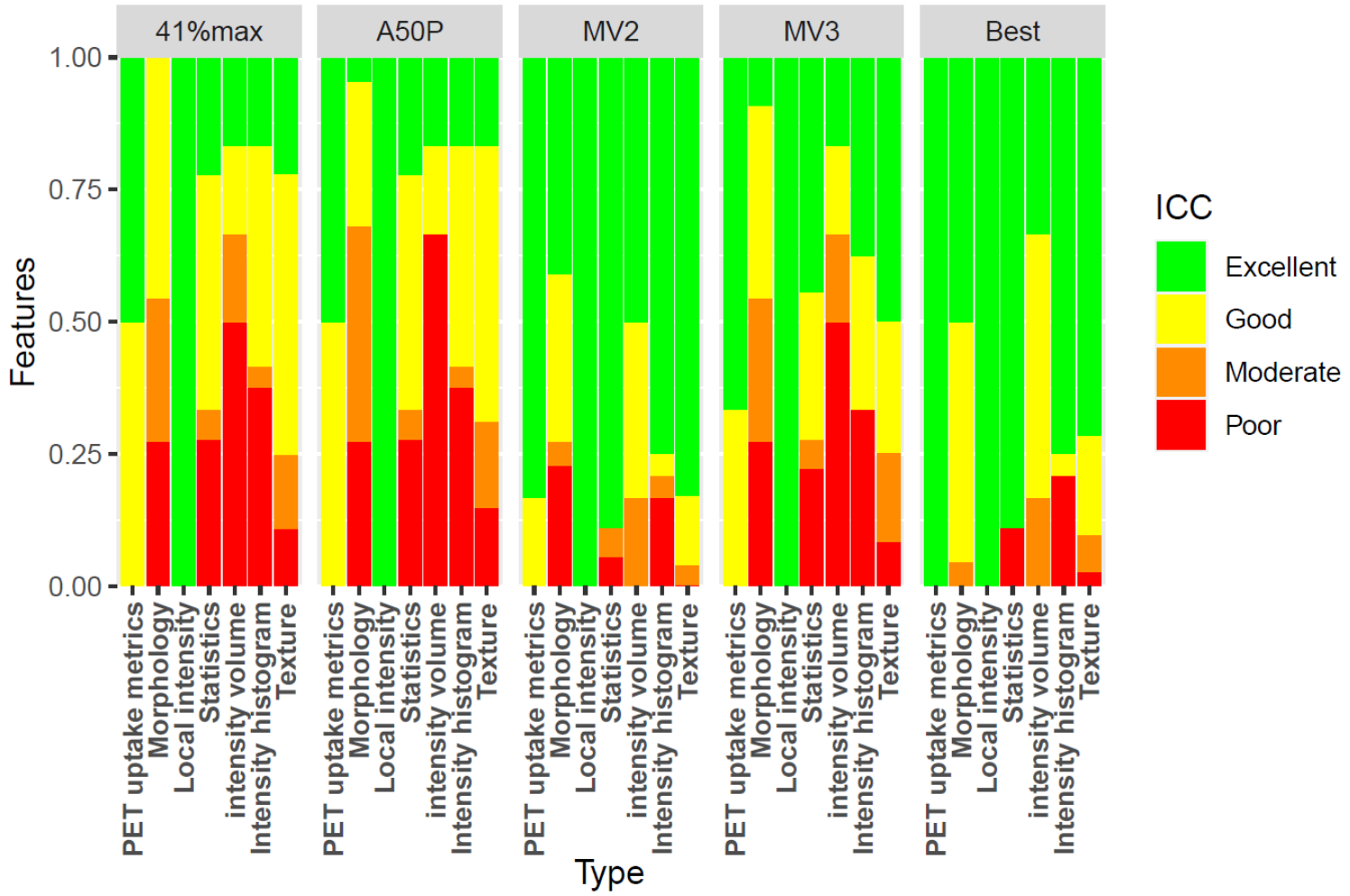


Figure 3. Percentage of radiomics features yielding excellent, good, moderate or poor intraclass correlation agreement between the SUV4.0 segmentation and other methods for the largest lesion.

Table 1. Characteristics of included patients

		Events	Non-events
Age	Median (IQR)	64 (61-71)	68 (63-74)
	≤60 years	11	11
	>60 years	39	39
Sex	Male	28	26
	Female	22	24
Ann Arbor Stage	2	3	6
	3	9	13
	4	38	31
Lactate dehydrogenase	normal	8	19
	>normal	42	31
Extranodal localisations	≤1	21	28
	>1	29	22
Performance status	0	16	29
	1	25	13
	2	9	8
International prognostic index	Low	3	5
	Low-intermediate	2	14
	High-intermediate	25	18
	High	20	13

Abbreviations: *IQR: interquartile range*

Table 2. SUV_{peak} and MTV per segmentation method

Abbreviations: SUV: standardized uptake value, IQR: interquartile range

	SUV_{peak} (median, IQR)	MTV patient level (median,IQR)	MTV largest lesion (median,IQR)
SUV4.0	17.1(12.8-22.0)	552.7(310.3-1117.2)	353.5(145.3-854.4)
SUV4.0(≥3mL)	17.2(12.8-22.3)	534.8(295.4-1116.4)	353.5(145.3-854.4)
A50P	16.8(12.5-22.0)	463.5(210.2-1164.0)	264.6(75.9-658.1)
41%max	16.8(12.5-22.0)	492.0(230.3-1203.5)	295.3(112.6-741.8)
MV2	16.8(12.8-22.0)	726.2(374.5-1299.9)	445.1(188.0-1041.6)
MV3	16.8(12.5-22.3)	502.5(235.5-1155.0)	280.2(98.9-693.9)
Best	16.6(12.4-21.9)	653.2(350.5-1283.8)	445.1(172.6-935.5)

Table 3. Number of independent features per segmentation method, number of included features and predictive value at patient level for all extracted features (n=488) and all reliable, repeatable and reproducible features (n=103)

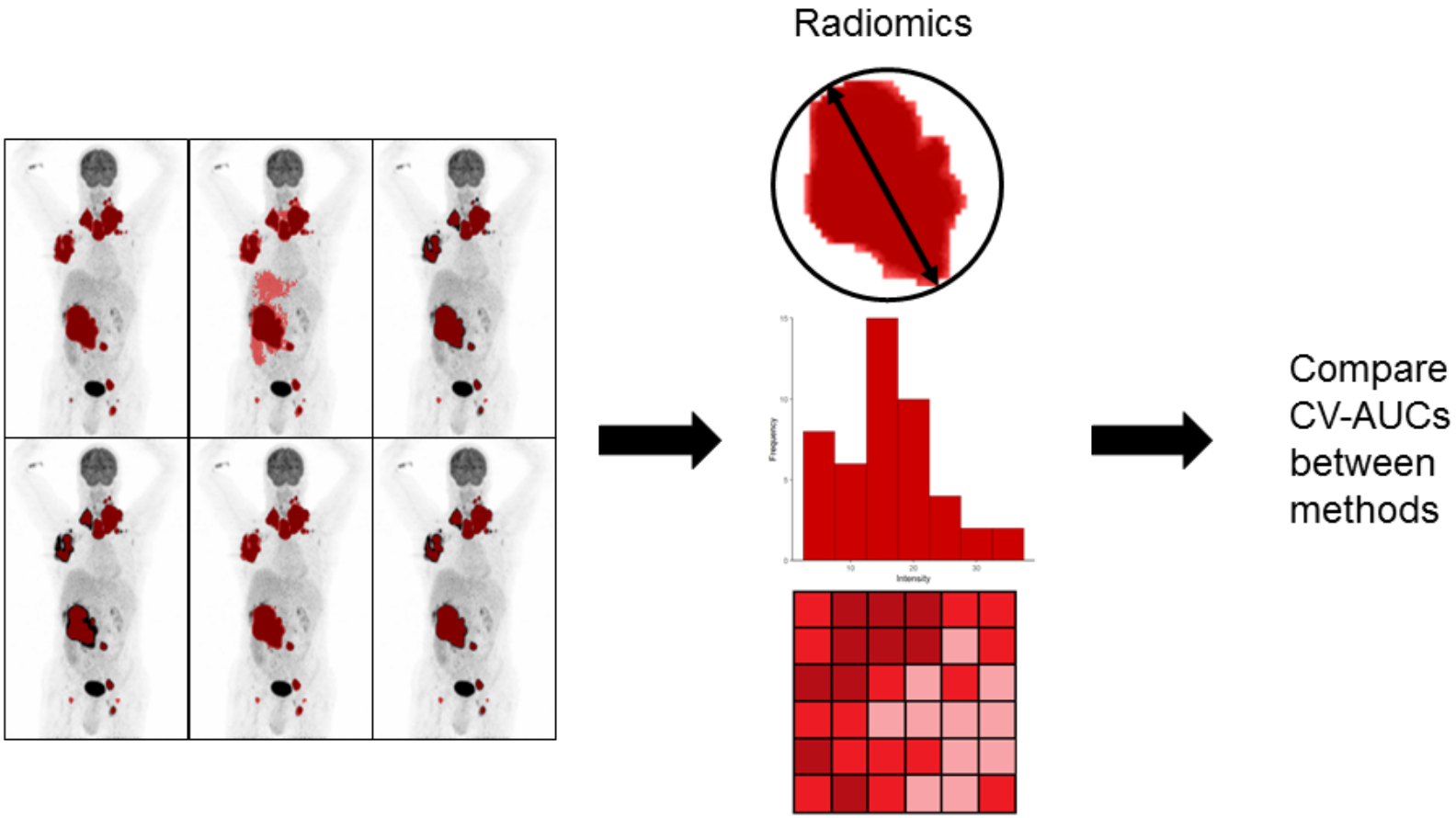
		Independent of MTV	Independent of SUV_{peak}	Independent of MTV and SUV_{peak}	Independent of MTV and SUV_{peak} and uncorrelated	Number of features in linear regression	CV-AUC (±SD)
n=488 features	SUV4.0	445	211	172	25	12	0.74±0.10
	SUV4.0(≥3mL)	443	212	170	25	4	0.74±0.10
	41%max	435	198	145	27	11	0.84±0.09
	A50P	441	204	157	32	20	0.78±0.10
	MV2	444	199	155	26	5	0.79±0.09
	MV3	441	203	156	29	18	0.80±0.09
	Best	445	190	147	25	12	0.69±0.11
n=103 features	SUV4.0	64	63	35	13	3	0.70±0.11
	SUV4.0(≥3mL)	61	63	32	12	6	0.70±0.11
	41%max	54	63	24	10	4	0.75±0.10
	A50P	58	65	30	10	3	0.63±0.11
	MV2	61	66	34	11	8	0.74±0.10
	MV3	58	65	30	9	4	0.73±0.10
	Best	62	67	36	11	7	0.69±0.11

Abbreviations: MTV: metabolic tumor volume, SUV: standardized uptake value, CV-AUC: cross-validated area under the curve, SD: standard deviation

Table 4. Number of independent features per segmentation method, number of included features and predictive value for the largest lesion for all extracted features (n=483) and all reliable, repeatable and reproducible features (n=99)

		Independent of MTV	Independent of SUV_{peak}	Independent of MTV and SUV_{peak}	Independent of MTV and SUV_{peak} and uncorrelated	Number of features in linear regression	CV-AUC (±SD)
n=483 features	SUV4.0	427	134	85	24	11	0.73±0.10
	41%max	409	158	84	19	10	0.71±0.11
	A50P	424	176	117	21	8	0.71±0.11
	MV2	435	141	93	21	3	0.71±0.10
	MV3	424	173	114	21	5	0.69±0.11
	Best	437	168	122	25	10	0.69±0.11
n=99 features	SUV4.0	57	46	13	10	5	0.73±0.10
	41%max	50	57	14	6	1	0.65±0.11
	A50P	54	59	20	9	4	0.63±0.11
	MV2	59	52	18	8	3	0.70±0.11
	MV3	54	52	18	7	3	0.67±0.11
	Best	59	55	21	10	3	0.69±0.11

Abbreviations: MTV: metabolic tumor volume, SUV: standardized uptake value, CV-AUC: cross-validated area under the curve, SD: standard deviation



Graphical Abstract

Supplemental data

Segmentation methods.

The following frequently used semi-automatic segmentation methods were applied to delineate lesions:

1. Fixed threshold of SUV2.5 (SUV2.5)
2. Fixed threshold of SUV4.0 (SUV4.0)
3. Adaptive threshold based on 50% of the SUV_{peak} adapted for local background (A50P) (*1*)
4. Adaptive threshold based on 41% of the SUV_{max} (41%max)
5. Majority vote segmenting voxels detected by $\geqslant$2 methods (MV2)
6. Majority vote segmenting voxels detected by $\geqslant$3 methods (MV3)

For the first 4 methods, all voxels above the fixed or adaptive threshold were added to the volume of interest (VOI). For the A50P method, background was defined by measuring the uptake around the lesion (1.5cm distance from 70% contour) (*1*). For the MV2 method all voxels that were selected by $\geqslant$2 methods (out of SUV4.0, SUV2.5, 41%max and A50P) were added to the VOI and for the MV3 method voxels detected by $\geqslant$ 3 methods were added to the VOI. All segmentations of individual lesions were included in one VOI to generate a patient level VOI by assigning all voxels within the individual lesions to one and all voxels outside any of the segmented individual lesions to zero.

Calculation of radiomics features

Radiomics features were extracted for the patient level VOI and largest lesion following descriptions of the Image Biomarker Standardization Initiative using RaCaT software (*2,3*). Definitions for calculation of individual radiomics features can be found in the Image Biomarker Standardization Initiative website (https://ibsi.readthedocs.io/). Radiomics features are sensitive to resolution, voxel size and image noise (*4*) therefore standardization of feature values is needed to reduce the variability of radiomics features across centers. By default, all images in RaCaT are resampled to 2x2x2 voxel size using tri-linear interpolation, as spatial resampling to cubic voxels led to better reproducibility of radiomics features in a multicenter setting (*4*). Intensity is discretized with a fixed bin size of 0.25 SUV before feature calculation to increase the percentage of consistent features, as image intensity discretization with a fixed bin width of 0.25 has shown to result in higher reliability of radiomics features in a multicenter setting compared to fixed bin number discretization. The exact same spatial rebinning was applied to the volumes of interest followed with a voxelwise 50% thresholding to generate a binary tumor map after rebinning and the latter was subsequently used to extract or calculate the radiomics features from the spatially rebinned PET images.

For both patient level and lesion level VOIs, 6 additional conventional PET uptake metrics were extracted before rebinning: MTV, SUV_{max}, local SUV_{peak}, global SUV_{peak}, SUV_{mean} and total lesion glycolysis (SUV_{mean}*MTV). Texture features were based on the grey level co-occurrence matrix (GLCM) , grey level run length matrix (GLRLM), grey level size zone matrix (GLSZM), grey level distance zone matrix (GDLZM), neighbourhood grey tone difference matrix (NGTDM) and neighbouring grey level dependence matrix (NGLDM) with up to 8 matrix calculation methods. For the patient level VOI, all voxels belonging to the different lesions were processed if they were part of one VOI and one matrix was used per patient to register all voxel pairs into the matrix. At patient level, 4 additional dissemination features were calculated: the distance between

the 2 lesions that were furthest apart ($Dmax_{patient}$), the distance between the largest lesion and the lesion furthest from that bulk ($Dmax_{bulk}$), the sum of the distances from the largest lesion to all other lesions ($spread_{bulk}$) and the sum of the distances from all lesions to all the other lesions ($spread_{patient}$) (*5*). Distances were calculated in millimeters based on the location of the SUVmax of individual lesions.

Intra-class correlation

The agreement between radiomics features extracted from different segmentation methods was quantified by calculating the intra-class correlation agreement ($ICC_{agreement}$) compared to the SUV4.0 segmentation method. The $ICC_{agreement}$ is a reliability index that reflects both degree of correlation and agreement between measurements. ICCs were categorized as poor (ICC: <0.5), moderate (ICC: 0.5-0.74), good (ICC: 0.75-0.89) or excellent reliability (ICC: ≥0.90) (*21*). Two features at patient level, and three features at lesion level did not show any variation and were therefore excluded before calculating the ICC.

Supplemental Table 1. Number of features with excellent, good, moderate or poor ICC values for different radiomics feature groups at patient level compared to the SUV4.0 segmentation method

		PET uptake metrics	Dissemination	Morphology	Local intensity	Statistics	Intensity volume	Intensity histogram	Texture
SUV4.0 (>3mL)	Excellent	6	2	12	2	14	2	17	359
	Good	0	1	10	0	4	2	6	36
	Moderate	0	1	0	0	0	2	1	9
	Poor	0	0	0	0	0	0	0	2
41%max	Excellent	3	2	3	2	8	1	9	190
	Good	3	0	16	0	4	1	5	136
	Moderate	0	1	3	0	2	1	2	55
	Poor	0	1	0	0	4	3	8	25
A50P	Excellent	3	2	3	2	8	1	9	203
	Good	3	1	12	0	4	1	5	120
	Moderate	0	0	7	0	3	1	4	57
	Poor	0	1	0	0	3	3	6	26
MV2	Excellent	4	2	8	2	12	3	13	267
	Good	2	0	12	0	4	0	6	92
	Moderate	0	2	2	0	0	2	0	35
	Poor	0	0	0	0	2	1	5	12
MV3	Excellent	4	2	4	2	10	2	11	264
	Good	2	1	14	0	2	0	3	66
	Moderate	0	1	4	0	3	1	4	67

	Poor	0	0	0	0	3	3	6	9
Best	Excellent	5	2	11	2	12	3	14	306
	Good	1	1	11	0	4	0	5	72
	Moderate	0	1	0	0	0	3	0	21
	Poor	0	0	0	0	2	0	5	7

Excellent: ICC≥0.90, good: ICC between 0.75-0.90, moderate: ICC between 0.5-0.75, poor: ICC <0.5

Supplemental Table 2. Radiomics features that were included in final logistic regression models of each segmentation method at patient level for all extracted features and all reliable, repeatable and reproducible features.

	All extracted features (n=488)	**Selection of reliable, repeatable and reproducible features (n=103)**
SUV4.0	*PET uptake metrics*: MTV *Dissemination*: DmaxBulk *Morphology*: center of mass shift, Gearys C, flatness *Intensity histogram*: skewness *Intensity volume*: volume at intensity fraction 10 *Texture*: First measure of information correlation (GLCM 2D), cluster shade (GLCM 2D), large distance low grey level emphasis (GLDZM 2D), large zone high grey level emphasis (GLDZM 2D), dependence count non uniformity normalized (NGLDM 2D)	*PET uptake metrics*: SUVpeak *Morphology*: elongation, maximum 3D diameter
SUV4.0(≥3mL)	*Morphology:* center of mass shift *Intensity histogram*: minimum *Texture*: First measure of information correlation (GLCM 2D) low dependence low grey level emphasis (NGLDM 2D)	*PET uptake metrics*: MTV, SUVpeak *Morphology*: elongation *Texture*: large distance emphasis (GLDZM 2D), Grey level non uniformity (GLDZM 2D), High dependence high grey level emphasis (NGLDM 3D)
41%max	*PET uptake metrics*: MTV *Morphology*: center of mass shift, elongation, volume density AEE *Statistics*: kurtosis *Intensity histogram*: minimum histogram gradient *Texture*: joint maximum (GLCM 2D), second measure of information correlation (GLCM 3D), small distance emphasis (GLDZM 3D), coarseness (NGTDM 2D), low dependence low grey level emphasis (NGLDM 3D)	*PET uptake metrics*: MTV *Morphology*: elongation *Intensity histogram*: skewness *Texture*: small distance emphasis (GLDZM 2D)
A50P	*PET uptake metrics*: MTV *Dissemination*: DmaxBulk, spreadPatient *Morphology:* center of mass shift, elongation *Statistics*: kurtosis *Intensity histogram*: skewness, quartile coefficient *Intensity volume*: difference volume at intensity fraction *Texture:* cluster shade (GLCM 2D), angular second moment (GLCM 2D), first measure of information correlation (GLCM 2D), zone percentage (GLSZM 2D), large distance low grey level emphasis (GLDZM 2D), small	*PET uptake metrics*: MTV *Statistics*: skewness *Texture*: small distance emphasis (GLDZM 2D)

	distance emphasis (GLDZM 2D), large zone high grey level emphasis (GLDZM 2D), grey level variance (GLDZM 2D), strength (NGTDM 2D), dependence count variance (NGLDM 3D), low dependence low grey level emphasis (NGLDM 3D)	
MV2	*Morphology*: center of mass shift, spherical disproportion *Statistics*: skewness *Texture*: first measure of information correlation (GLCM 2D), low dependence low grey level emphasis (NGLDM 3D)	*Morphology*: maximum 3D diameter, morans I, elongation *Statistics*: skewness *Texture*: small zone emphasis (GLSZM 2D), large distance low grey level emphasis (GLDZM 2D), small distance emphasis (GLDZM 2D), small distance emphasis (GLDZM 3D)
MV3	*PET uptake metrics*: MTV *Dissemination*: DmaxBulk *Morphology:* center of mass shift, elongation *Statistics*: kurtosis *Intensity histogram*: skewness, quartile coefficient *Texture*: sum average (GLCM 2D), first measure of information correlation (GLCM 2D), Cluster shade (GLCM 2D), first measure of information correlation (GLCM 3D), second measure of information correlation (GLCM 3D), large zone emphasis (GLSZM 2D), large zone high grey level emphasis (GLDZM 2D), zone distance non uniformity normalized (GLDZM 2D), grey level variance (GLDZM 2D), high dependence high grey level emphasis (NGLDM 3D), contrast (NGTDM 3D)	*PET uptake metrics*: MTV *Morphology*: elongation *Intensity histogram:* skewness *Texture*: zone distance non uniformity normalized (GLDZM 2D)
Best	*PET uptake metrics*: MTV, SUVpeak *Dissemination*: DmaxBulk *Morphology*: center of mass shift, elongation *Texture*: first measure of information correlation (GLCM 2D), angular second movement (GLCM 2D), zone size non uniformity normalized (GLSZM 2D), grey level variance (GLDZM 2D), small distance emphasis (GLDZM 2D), small distance emphasis (GLDZM 3D), low dependence low grey level emphasis (NGLDM 2D)	*PET uptake metrics*: SUVpeak *Morphology*: elongation, maximum 3D diameter *Intensity histogram*: minimum histogram gradient *Texture*: large distance low grey level emphasis (GLDZM 2D), small distance emphasis (GLDZM 2D), small zone emphasis (GLSZM 2D)

Supplemental Table 3. Number of features with excellent, good, moderate or poor ICC values for different radiomics feature groups at patient level compared to the SUV4.0 segmentation method

		PET uptake metrics	Morphology	Local intensity	Statistics	Intensity volume	Intensity histogram	Texture
41%max	Excellent	3	0	2	4	1	4	89
	Good	3	10	0	8	1	10	215
	Moderate	0	6	0	1	1	1	57
	Poor	0	6	0	5	3	9	44
A50P	Excellent	3	1	2	4	1	4	68
	Good	3	6	0	8	1	10	211
	Moderate	0	9	0	1	0	1	66
	Poor	0	6	0	5	4	9	60
MV2	Excellent	5	9	2	16	3	18	336
	Good	1	7	0	0	2	1	53
	Moderate	0	1	0	1	1	1	15
	Poor	0	5	0	1	0	4	1
MV3	Excellent	4	2	2	8	1	9	202
	Good	2	8	0	5	1	7	101
	Moderate	0	6	0	1	1	0	68
	Poor	0	6	0	4	3	8	34
Best	Excellent	6	11	2	16	2	18	290
	Good	0	10	0	0	3	1	76
	Moderate	0	1	0	0	1	0	28
	Poor	0	0	0	2	0	5	11

Excellent: ICC≥0.90, good: ICC between 0.75-0.90, moderate: ICC between 0.5-0.75, poor: ICC <0.5

Supplemental Table 4. Radiomics features that were included in final logistic regression models of each segmentation method for the largest lesion for all extracted features and all reliable, repeatable and reproducible features.

	All extracted features (n=483)	**Selection of reliable, repeatable and reproducible features (n=99)**
SUV4.0	*PET uptake metrics*: MTV, SUVpeak *Morphology*: volume density AEE, flatness, major axis length *Statistics*: skewness *Intensity histogram*: minimum, minimum histogram gradient grey level *Texture*: first measure of information correlation (GLCM 2D), large zone high grey level emphasis (GLSZM 2D), dependence count variance (NGLDM 2D)	*PET uptake metrics*: MTV, SUVpeak *Morphology*: surface to volume ratio *Intensity histogram*: minimum hisogram gradient grey level *Texture*: high dependence high grey level emphasis (NGLDM 2D)
41%max	*PET uptake metrics*: MTV *Morphology*: gearys C, volume density AABB, area density AABB *Intensity volume:* difference volume at intensity fraction *intensity histogram*: minimum histogram gradient *Texture*: first measure of information correlation (GLCM 2D), cluster shade (GLCM 2D), Second measure of information correlation (GLCM 3D), small distance emphasis (GLDZM 2D)	*Intensity histogram*: minimum histogram gradient
A50P	*PET uptake metrics*: SUVpeak *Intensity histogram*: skewness, minimum histogram gradient *Texture*: cluster shade (GLCM 2D), small distance emphasis (GLDZM 2D), contrast (NGTDM 2D), high dependence high grey level emphasis (NGLDM 2D), dependence count non uniformity normalized (NGLDM 2D)	*PET uptake metrics*: SUVpeak *Intensity histogram*: minimum histogram gradient *Texture*: small distance emphasis (GLDZM 2D), high dependence high grey level emphasis (NGLDM 2D)
MV2	*Intensity histogram*: minimum histogram gradient *Texture*: first measure of information correlation (GLCM 2D), low dependence low grey level emphasis (NGLDM 2D)	*Morphology*: minor axis length *Intensity histogram*: minimum histogram gradient grey level, minimum histogram gradient
MV3	*Intensity histogram*: minimum histogram gradient *Texture*: first measure of information correlation (GLCM 2D), cluster shade (GLCM 2D), short run emphasis (GLRM 2D), large zone high grey level emphasis (GLSZM 2D)	*PET uptake metrics*: SUVpeak *Intensity histogram*: minimum histogram gradient *Texture*: high dependence high grey level emphasis (NGLDM 2D)
Best	*PET uptake metrics*: MTV *Morphology*: elongation, flatness, volume density AEE, major axis length	*Morphology*: elongation *Intensity histogram*: minimum histogram gradient grey level, minimum histogram gradient

	Intensity histogram: minimum histogram gradient, minimum histogram gradient grey level *Texture*: First measure of information correlation (GLCM 2D), run percentage (GLRLM 2D), large zone emphasis (GLSZM 2D)	

References

1. Frings V, van Velden FH, Velasquez LM, et al. Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study. *Radiology.* 2014;273:539-548.

2. 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.

3. Zwanenburg A, Vallieres M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* 2020;295:328-338.

4. Pfaehler E, van Sluis J, Merema BBJ, et al. Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-dimensionally printed phantom inserts. *J Nucl Med.* 2020;61:469-476.

5. Cottereau AS, Nioche C, Dirand AS, et al. (18)F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. *J Nucl Med.* 2020;61:40-45.