Journal of Nuclear Medicine, published on March 5, 2021 as doi:10.2967/jnumed.120.258673

Interobserver agreement in automated metabolic tumor volume measurements of Deauville score 4 and 5 lesions at interim 18F-FDG PET in DLBCL

Gerben JC Zwezerijnen[1], Jakoba J Eertink[2], Coreline N Burggraaff[2], Sanne E Wiegers[2], Ekhlas AIN Shaban[3], Simone Pieplenbosch[1], Daniela E Oprea-Lager[1], Pieternella J Lugtenburg[4], Otto S Hoekstra[1], Henrica CW de Vet[5], Josee M Zijlstra[2], Ronald Boellaard[1]

[1]Amsterdam UMC, Vrije Universiteit Amsterdam, department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, De Boelelaan 1117, Amsterdam, Netherlands
[2]Amsterdam UMC, Vrije Universiteit Amsterdam, department of Hematology, Cancer Center Amsterdam, De Boelelaan 1117, Amsterdam, Netherlands
[3]Radiodiagnosis and medical imaging department, Faculty of Medicine, Tanta University, Egypt
[4]Erasmus MC Cancer Institute, University Medical Center, department of Hematology, Wytemaweg 80, 3015 CN, Rotterdam, The Netherlands
[5]Amsterdam UMC, Vrije Universiteit Amsterdam, department of Epidemiology and Data Science, Amsterdam Public Health research institute, De Boelelaan 1117, Amsterdam, Netherlands

Corresponding author:	**First author:**
Ronald Boellaard, MD, PhD	Gerben JC Zwezerijnen, MD
r.boellaard@amsterdamumc.nl	g.zwezerijnen@amsterdamumc.nl
ORCID: 0000-0002-0313-5686	ORCID: 0000-0002-9571-9362

Address: Department of Radiology and Nuclear Medicine
Amsterdam UMC, Vrije Universiteit Amsterdam,
Cancer Center Amsterdam
De Boelelaan 1117
1081HV Amsterdam, Netherlands
Phone: +31(0)2044449638

Word count:
4999

Financial support:
PETRA and RADIOMICS studies are supported by Alpe d'HuZes/KWF (Dutch Cancer Society; #VU2012-5848 and #VU2018-11648).
Running title:
MTV measurements on interim-PET DLBCL

ABSTRACT

Metabolic tumor volume (MTV) on interim-PET (I-PET) is a potential prognostic biomarker for diffuse large B-cell lymphoma (DLBCL). Implementation of MTV on I-PET requires consensus which semi-automated segmentation method delineates lesions most successfully with least user interaction. Methods used for baseline PET are not necessarily optimal for I-PET due to lower lesional standardized uptake values (SUV) at I-PET. Therefore, we aimed to evaluate which method provides the best delineation quality of Deauville-score (DS) 4-5 DLBCL lesions on I-PET at best interobserver agreement on delineation quality and, secondly, to assess the effect of lesional SUVmax on delineation quality and performance agreements. **Methods:** DS4-5 lesions from 45 I-PET scans were delineated using six semi-automated methods i) SUV 2.5, ii) SUV 4.0, iii) adaptive threshold [A50%peak], iv) 41% of maximum SUV [41%max], v) majority vote including voxels detected by ≥2 methods [MV2] and vi) detected by ≥3 methods [MV3]. Delineation quality per MTV was rated by three independent observers as acceptable or non-acceptable. For each method, observer scores on delineation quality, specific agreements and MTV were assessed for all lesions, and per category of lesional SUVmax (<5, 5-10, >10). **Results:** In 60 DS4-5 lesions on I-PET, MV3 performed best, with acceptable delineation in 90% of lesions, with a positive agreement (PA) of 93%. Delineation quality scores and agreements per method strongly depended on lesional SUV: the best delineation quality scores were obtained using MV3 in lesions with SUVmax<10 and SUV4.0 in more FDG-avid lesions. Consequently, overall delineation quality and PA improved by applying the most preferred method per SUV category instead of using MV3 as single best method. MV3- and SUV4.0-derived MTVs of lesions with SUVmax>10, were comparable after excluding visually failed MV3 contouring. For lesions with SUVmax<10, MTVs using different methods correlated poorly. **Conclusion:** On I-PET, MV3 performed best and provided the highest interobserver agreement regarding acceptable delineations of DS4-5 DLBCL lesions. However, delineation method preference strongly depended on lesional SUV. Therefore, we suggest to explore an approach that identifies the optimal delineation method per lesion as function of tumor FDG uptake characteristics, i.e. SUVmax.

Keywords: lymphoma, metabolic tumor volume, positron emission tomography, standardization

INTRODUCTION

18F-fluoro-2-deoxyglucose (FDG) positron emission tomography (PET) is a cornerstone in staging and response evaluation of malignant lymphoma (*1,2*). In Hodgkin's lymphoma, the utility of interim-PET (I-PET) in early response-adapted therapy has been demonstrated (*3*). In diffuse large B-cell lymphoma (DLBCL) the role of I-PET-adapted strategies is still controversial, mainly because of its insufficient positive predictive value (*4*).

To date, I-PET scans are classified using the Deauville five-point scale (DS) as described in the International Conference on Malignant Lymphoma (ICML) guidelines (*2*). However, DS4-5 scores which usually are considered as treatment failures suffer from a poor positive predictive value (*4,5*). Discrimination between true non-responders and responders might improve by quantification (e.g. the relative change of standardized tracer uptake) (*5–11*). Quantification will reduce observer variability, and this is essential for successful clinical implementation.

Metabolically active tumor volume (MTV) before treatment seems to have prognostic value (*12–14*). It has been suggested that MTV at I-PET might add prognostic value as well (*7–9,15–17*). Measuring MTV on I-PET is, however, challenging since lesional contrast in I-PET is often limited. Moreover, FDG uptake can be heterogeneous within and between lesions. Low, heterogeneous uptake results in poor delineation reproducibility (*18*). In addition, manual tumor segmentation is extremely time-consuming. Semi- or fully automated segmentation methods may partially eliminate these drawbacks (*19–21*), such as the so-called threshold-based methods, in which the delineation threshold is based on a fixed SUV (e.g. SUV ≥4.0 or 2.5), a fixed percentage of tumor SUV (e.g. ≥41% of SUVmax), or by a contrast oriented algorithm (adaptive thresholds, e.g. 50%peak) (*22*).

The prognostic relevance of baseline MTV is relatively independent of delineation methodology (albeit with different cut-off values) (*12–14*). Here, SUV4.0 seems to be most successful based on interobserver reliability and ease of use (*13,23*). However, this method is not necessarily optimal at I-PET because, at that time point, lesional tracer uptake and target-to-background contrast are lower, which may affect delineation quality. Consequently, more user interaction is needed to obtain proper delineations resulting in a potentially higher interobserver variability. Initial studies evaluating MTV at I-PET showed prognostic value but each applied a different threshold method (37%, 40-42% SUVmax, SUV2.5, SUV4.0, gradient-based method). Moreover, observer variation was not reported .

Delineation performance of these methods may depend on tumor FDG uptake characteristics (*24,25*). Therefore, selection of the most optimal method based on lesional imaging characteristics, as suggested by the Automatic decision Tree-based Learning Algorithm (ATLAAS) method selection approach, might improve delineation quality instead of using a single method for all lesions (*26*).

Successful validation and implementation of I-PET MTV in clinical trials and practice require reliable, reproducible MTV measurements at minimal operator interaction. Therefore, the aim of this study was to evaluate which method provides qualitative acceptable delineations of DS(4-5) DLBCL lesions on I-PET most often, with high interobserver agreement, and to study whether lesional SUVmax affects delineation performance agreements, and finally to assess to which extent overall delineation performance improves by selecting the best delineation method based on lesional SUVmax.

MATERIAL AND METHODS

Patient and PET Imaging Selection

Newly diagnosed DLBCL patients of the HOVON-84 (Haemato Oncology Foundation for Adults in the Netherlands) study, an international randomized clinical trial approved by institutional review board and/or ethics committees, conducted between November 2007 and April 2012 (EudraCT 2006-005174-42, NTR10140) with available I-PET data, were included for this study (*27,28*).

HOVON-84 was designed to evaluate early intensification of rituximab in the first four cycles combined with cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP14) on the response rate and time to reach response in previously untreated eligible patients with CD-20+ DLBCL. I-PET was performed after four cycles, and were centrally reviewed by two independent, experienced nuclear medicine physicians using the DS system, and a third reviewer when adjudication was required (*27*).

For the present study, we randomly selected 45 I-PET/CT scans of patients with an incomplete metabolic response (DS4-5).

Automated PET Delineation Methods

Six automated PET delineation methods were applied using in-house developed software (ACCURATE tool) four threshold-based methods using a fixed SUV threshold of 2.5 g/mL (SUV2.5) and 4.0 g/mL (SUV4.0), a threshold at 41% of the SUVmax per lesion (41%max) and an adaptive threshold corrected for source-to-local background activity contrast at 50% of the SUVpeak (A50%peak) (*19,29*). A50%peak segments lesions when lesional uptake is less than twice the local background, defined as the mean uptake of a single-voxel shell 2.5 cm around the edge of a 70% of SUVmax value isocontour excluding voxels with SUVmax>2.5 (*30*). SUVpeak was defined as the highest average SUV of a 1 mL sphere volume of interest (VOI) across all positions within the target lesion (*31*).

The two remaining delineation methods are based on a majority vote (MV) approach by which contours are determined by the intersection of the four abovementioned threshold-based delineations. For these MVs, a voxel was included in the consensus delineation, according to the results of the majority of the threshold-based methods. This implies agreement between at least 2 or 3 of the threshold-based methods defined as MV2 and MV3, respectively.

These six methods semi-automatically segmented MTVs based on the voxel with the highest detected SUV(max/peak) within the manually selected lymphoma target lesion (prepared by CE, SW, SP). Semi-automated derived delineations were not manually adapted.

Observer Evaluation

VOIs from these six methods were visualized on all I-PET images to allow assessment of delineation quality separately by three observers (nuclear medicine physician (GZ, 5 y experience), radiologist (ES, 5 y experience) and hematologist (JZ, 15 y experience)). Observers were blinded to the delineation method and clinical outcomes, but not to baseline PET/CT. Each observer evaluated the quality of the MTV segmentation at a lesional basis.

The delineation quality per method was scored as (Supplemental Figure 1) (*23*):

- Acceptable: MTV required no (“good”) or minimal manual adaption (“moderate”) to obtain a visually accurate lymphoma segmentation.
- Non-acceptable: MTV with a lot of physiological background activity or visually not selected the complete lymphoma-lesion requiring extensive manual adaption (“poor”), or MTV so poorly defined that even (extensive) manual adaption was no longer considered feasible (“failed”).

Statistical Analysis

The median and interquartile range of the observed MTVs were calculated for each delineation method.

For all lesions, the dichotomous delineation quality scores, as rated by the observers, were summarized as frequencies for each of the six delineation methods. Additionally, observer scores were calculated on lesions categorized by their SUVmax (<5, 5-10, >10, respectively) to evaluate the effect of lesional SUVmax on these quality scores.

To evaluate which method obtained the best agreement among observers on providing "acceptable" delineations (i.e. without need for extensive manual correction) of DLBCL I-PET lesions, we used percentage agreement, specified for rating "acceptable" (for a positive test result: positive agreement, PA) (*32,33*). PA implies the percentage probability that observer B scores a method's MTV as "acceptable" identical to observer A. The negative agreement measures (NA) reflects the probability that observers agreed that the delineation performance rating was not "acceptable". We primarily focused on the agreement on best performing method; therefore, NA measures are only reported to provide a complete overview of the results.

Additional specific agreement analyses were performed per SUVmax category. Based on the highest acceptable delineation quality scores and its PA, we explored which method was preferred per SUVmax category. Next, we evaluated the extent to which the overall interobserver scoring performance improved when per SUVmax category, the most preferred method was used.

We explored if MTV calculated with the (multi)method of preference approach can be converted by applying a mathematical transformation into values that are comparable when using the single method of preference. Therefore, we tested whether these MTVs per SUVmax category were (log)normally distributed, and assessed their correlation with a Pearson correlation coefficient.

RESULTS

The 45 I-PET scans showed 60 DS4-5 lesions (range: 1-4 per scan), with a median SUVmax of 6.9 (Interquartile Range; 5.0-10.2). The smallest median MTV was obtained with SUV4.0 (2.3 mL) and MV3 (4.9 mL) and the largest with methods SUV2.5 (29.9 mL) and 41%max (27.6 mL) (Table 1).

Regarding delineation quality scores, MV3 derived MTVs were most frequently considered visually acceptable, with an average acceptable observer score frequency of 90% of the lesions (Figure 1, Table 2, Supplemental Table 1). A50%peak and MV2 showed lower delineation performance (in 77.2% and 72.8% of lesions, respectively), whereas SUV4.0 least frequently provided acceptable delineations (52.8%). The observer scores per method differed for lesions between the SUVmax categories: acceptable observer score frequency for method SUV2.5 and A50%peak was higher for lesions with SUVmax<5 than for lesions with SUVmax>5. An opposite trend was observed for SUV4.0 and 41%max. The acceptable observer score frequency for MV3 showed a difference of 6.6 percentage points among the SUV categories and was, therefore, least affected by lesional SUVmax compared to the other methods.

Specific agreement analysis for the delineation rating "acceptable" calculated over all lesions revealed the highest PA for methods MV3 (93.2%), A50%peak (92.1%) and MV2 (90.8%), whereas the PA for the other methods was lower with 79.6% for SUV2.5, 84.2% for SUV4.0 and 87.2% for 41%max.

Observers agreed in only 38.9% of lesions that MV3 provided non-acceptable delineations (NA). The highest NA for rating "acceptable" was reached for methods 41%max (80.3%) and SUV4.0 (82.4%).

The specific agreements per method depended on lesional SUVmax (Table 2, Figure 2). The absolute threshold methods showed opposite trends in which SUV2.5 mainly performed well in lesions with low SUVs and SUV4.0 in lesions with a high SUVmax. Of the relative threshold methods, method 41%max performed suboptimally in lesions with SUVmax<5 compared to lesions with higher SUVs (PA 77.8% versus 91.8 and 85.7%, respectively), while A50%peak performed best in low avid lesions (SUVmax<5, PA 94.9%). Both MV methods showed a high PA in lesions with SUVmax<10, but MV3 performed best with a lower NA in these low and medium avid (SUVmax 5-10) lesions.

MV3 was the method of preference for lesions with SUVmax<10, based on the highest acceptable observer score frequency combined with one of highest PA and lowest NA values (Table 2, Supplemental Table 1). SUV4.0 was considered as most preferred method for lesions with SUVmax>10. Comparing the method of preference approach with MV3, as overall best performing method, resulted in an overall increased acceptable score frequency from 90% to 92.8%, an increased PA from 93.2% to 95.2% and a decreased NA from 38.9% to 30.5%.

Log transformed MTVs obtained using MV3 and SUV4.0 for lesions with SUVmax>10, both normally distributed (Shapiro-Wilk, p>0.05), showed a strong positive linear relationship (R2 = 0.87, p<0.001). All SUV4.0 and MV3 derived MTVs for lesions with SUVmax>10 were nearly equal and within 1.96 standard deviation and/or 10 mL from the line of identity, except for two outliers (*A and *B in Figure 3, Supplemental Figure 2). The MV3 method for both of these outlier MTVs was rated as “non-acceptable” by the observers, each suggesting an underestimation of lesion volume while MV3 was considered “acceptable” for the other MTVs in this SUV category.

Since MV3 was considered as the overall single method of preference and the preferred method for lesions with SUVmax<10, no additional transformation analyses were required. Thereby none of the log transformed method derived volumes of lesions with a SUVmax<10, including MV3 and SUV4.0, were normally distributed and/or showed high Spearman correlation coefficients (Supplemental Figure 3).

DISCUSSION

To the best of our knowledge, this is the first multi-observer study that evaluated the delineation performance of several semi-automated segmentations methods for DS4-5 DLBCL lesions at I-PET. Overall, MV3-derived MTVs were most frequently scored as acceptable, with the highest positive and lowest negative agreement.

Previous studies suggested that a low MTV on I-PET or a major decrease of MTV vs. baseline PET in DLBCL, predicts response at end-of-treatment PET and progression-free-survival (*7–9,15–17*). However, MTV cut-offs were different, and it is unclear whether and/or to which extent this relates to the use of different semi-automated delineation algorithms; Oñate-Ocaña used a 40%SUVmax threshold, Zhang a 41%, Wu a 42% and Malek a 37%SUVmax threshold and a gradient-based segmentation method (Gradient), while Islam applied a SUV4.0 threshold and Yang and Mikhaeel a SUV2.5 threshold (*7–9,15–17,34*). This precludes any meaningful meta-analysis to build the case of evidence for MTV as a predictor for clinical outcome additional to (delta)SUVmax and the five-point DS (*10,11*). The latter particularly for I-PET DLBCL studies, since correlation between MTVs obtained by different segmentation methods is generally low for lesions with SUVmax<10 (Supplemental Figure 3) frequently prevailing at I-PET in contrary to baseline PET.

Delineation quality and reproducibility of a fixed SUV threshold-based method may be most sensitive to lesion uptake and local tumor-to-background contrast (*24,25*). Evaluating this hypothesis on our DLBCL I-PET cohort showed indeed that besides MTV correlations, also quality scores and interobserver agreement strongly depended on the lesional SUVmax (Figure 2, Table 2, Supplemental Figure 3). The latter explains, at least partly, the discordance in method preference at baseline PET/CT where the SUV4.0 is preferred versus the preference for MV3 at I-PET, as the lesional tracer uptake is much lower at I-PET than at baseline (*13,23*). The delineation performance of SUV4.0 was still successful at I-PET for high avid lesions (SUVmax>10), i.e. for lesions with uptake levels comparable to baseline ones.

Delineation performance of the 41%max method was also considered less successful in low avid lesions i.e. with a low tumor-to-background ratio (Table 2; 40% “acceptable” delineation quality score frequency, PA 78%). This is in line with the European Association of Nuclear Medicine (EANM) guidelines for tumor imaging (*35*). Our results suggest that delineating low avid lesions is most successful using SUV2.5, A50%peak, MV2 and MV3 (Figure 1, 2, Table 2). Overall, the observer score frequency and PA for successful delineation was best for MV3 and was least affected by lesional SUV. Therefore, MV3 might be considered as the (single) method of choice for assessing MTV in patients or PET studies showing a large variation in lesional tracer uptake.

However, no single semi-automated delineation method, including MV3 performs optimally for different types of lymphoma at different therapeutic stages without the need for manual correction (*25*). Therefore a workflow where observers select the visually best performing method per lesion might improve overall delineation success while minimizing interobserver variability compared to manual segmentation (*36*). Translating this workflow at I-PET DBLCL might imply that lesions with SUVmax<10 should be delineated using method MV3 and that lesions with SUVmax>10 should be delineated using method SUV4.0, or only in case observers consider MV3 contouring as failed.

Berthon et al. introduced a delineation method selection approach using an Automatic decision Tree-based Learning Algorithm (ATLAAS) to further improve accurate and reproducible lesion segmentation (*26*). This concept is based on selecting the best method from several predefined methods using lesion characteristics as input, which outperformed the PET segmentation accuracy of each single method. We also found that using the method with the highest acceptable score frequency per SUV category, resulted in more successful delineation performance compared to the performance of each delineation method separately (Table 2). The overall good performance of MV3 is to some extent in line with the ATLAAS approach, i.e. the MV3 method is based on majority vote selection of voxels to be included in the final MV3 VOI using 4 segmentation methods as input. Therefore, other consensus approaches like simultaneous truth and performance level estimation (STAPLE) might demonstrate an overall good delineation performance in DLBCL I-PET as well (*37*). However, identifying a single MTV delineation algorithm which is accurate, easy to use, reliable when applied in multicenter/observer setting and with good prognostic performance may need to be reconsidered against an approach based on selecting the most preferred method on a lesional basis, in particular for I-PET. Adding tumor volume, tumor SUVpeak to background ratio and other PET metrics for selecting the best delineation method per lesion might further improve the delineation performance (*26*). Nevertheless, development of such an approach requires a much larger dataset. Other advanced semi-automated segmentation methods, e.g. based on artificial intelligence, might also increase the delineation success performance but are not yet available and presently hampers implementation in a multicenter setting (*38*). Our proposed approach can be applied simply by first determining SUVmax and subsequently apply the MV3 or SUV4.0 method without the need for developing complicated new tools and are thus readily available.

Overall, we agree that the current literature has not made a convincing case that MTV outperforms deltaSUVmax at I-PET. However, its potential added value can only be demonstrated if MTV methodology is optimized and harmonized. Finally, it is unclear whether negative I-PET guided trials are caused by inappropriate patient selection (relying solely on the far from perfect accuracy of its positivity criteria used so far). Therefore, attempts to improve and standardize the I-PET response criteria, possibly including MTV, are urgently needed.

CONCLUSION

To delineate DS4-5 DLBCL lesions on I-PET, the semi-automatic delineation approach MV3 was most often successful at the highest interobserver agreement. However, delineation quality and interobserver agreement strongly depended on SUVmax. Therefore, a delineation method selection strategy using lesional tracer uptake metrics as input may provide better segmentations. Since MV3 already showed a very high success rate of 90% across all lesions, we propose to use this method for measuring MTV of DS4-5 lesions at I-PET in a supervised manner, i.e. by visually inspecting the delineation and optionally choose the SUV4.0 method for very high avid lesion (SUVmax>10) when deemed necessary.

Disclosure

No potential conflict of interest by the authors.

Key points

Questions i) to evaluate which method provides the best delineation quality of Deauville-score (DS) 4-5 DLBCL lesions on I-PET at best interobserver agreement on delineation quality. ii) to assess the effect of lesional SUVmax on delineation quality and performance agreements.

Findings i) MV3 performed best and at the highest interobserver agreement regarding acceptable delineations of DS4-5 DLBCL lesions on I-PET. ii) delineation method preference strongly depended on lesional SUV.

Implications for patient care Automated estimation of MTV of DS4-5 DLBCL lesions at I-PET is feasible in clinical practice in a supervised manner by using MV3 and optionally SUV4.0 for very high avid lesions.

REFERENCES

1. Barrington SF, Mikhaeel NG, Kostakoglu L, et al. Role of imaging in the staging and response assessment of lymphoma: Consensus of the international conference on malignant lymphomas imaging working group. *J Clin Oncol*. 2014;32:3048-3058.
2. Cheson BD, Fisher RI, Barrington SF, et al. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: The lugano classification. *J Clin Oncol*. 2014;32:3059-3067.
3. Barrington SF, Johnson PWM. 18 F-FDG PET/CT in lymphoma: has imaging-directed personalized medicine become a reality? *J Nucl Med*. 2017;58:1539-1544.
4. Burggraaff CN, de Jong A, Hoekstra OS, et al. Predictive value of interim positron emission tomography in diffuse large B-cell lymphoma: a systematic review and meta-analysis. *Eur J Nucl Med Mol Imaging*. 2019;46:65-79.
5. Györke T, Carr R, Cerci JJ, et al. Combined visual and semiquantitative evaluation improves outcome prediction by early midtreatment (18)F-FDG PET in diffuse Large B-dell lymphoma. *J Nucl Med*. 2020;61:999-1005.
6. Dührsen U, Müller S, Hertenstein B, et al. Positron emission tomography-guided therapy of aggressive non-hodgkin lymphomas (PETAL): A multicenter, randomized phase III trial. *J Clin Oncol*. 2018;36:2024-2034.
7. Malek E, Sendilnathan A, Yellu M, Petersen A, Fernandez-Ulloa M, Driscoll JJ. Metabolic tumor volume on interim PET is a better predictor of outcome in diffuse large B-cell lymphoma than semiquantitative methods. *Blood Cancer J*. 2015;5:e326.
8. Islam P, Goldstein J, Flowers CR. PET-derived tumor metrics predict DLBCL response and progression-free survival. *Leuk Lymphoma*. 2019;60:1965-1971.
9. Oñate-Ocaña LF, Cortés V, Castillo-Llanos R, et al. Metabolic tumor volume changes assessed by interval18fluorodeoxyglucose positron emission tomography-computed tomography for the prediction of complete response and survival in patients with diffuse large b-cell lymphoma. *Oncol Lett*. 2018;16:1411-1418.
10. Rekowski J, Hüttmann A, Schmitz C, et al. Interim PET evaluation in diffuse large B-cell lymphoma employing published recommendations: Comparison of the Deauville 5-point scale and the ΔSUV(max) method. *J Nucl Med*. 2020. Epub ahead of print.
11. Casasnovas R-O, Ysebaert L, Thieblemont C, et al. FDG-PET-driven consolidation strategy in diffuse large B-cell lymphoma: final results of a randomized phase 2 study. *Blood*. 2017;130:1315-

1326.

12. Schmitz C, Hüttmann A, Müller SP, et al. Dynamic risk assessment based on positron emission tomography scanning in diffuse large B-cell lymphoma: Post-hoc analysis from the PETAL trial. *Eur J Cancer*. 2020;124:25-36.

13. Barrington S, 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*. 2020. Epub ahead of print.

14. Prieto Prieto JC, Vallejo Casas JA, Hatzimichael E, Fotopoulos A, Kiortsis D-N, Sioka C. The contribution of metabolic parameters of FDG PET/CT prior and during therapy of adult patients with lymphomas. *Ann Nucl Med*. 2020;34:707-717.

15. Yang DH, Ahn JS, Byun BH, et al. Interim PET/CT-based prognostic model for the treatment of diffuse large B cell lymphoma in the post-rituximab era. *Ann Hematol*. 2013;92:471-479.

16. Mikhaeel NG, Smith D, Dunn JT, et al. Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL. *Eur J Nucl Med Mol Imaging*. 2016;43:1209-1219.

17. Zhang Y-Y, Song L, Zhao M-X, Hu K. A better prediction of progression-free survival in diffuse large B-cell lymphoma by a prognostic model consisting of baseline TLG and %ΔSUV(max). *Cancer Med*. 2019;8:5137-5147.

18. Hofheinz F, Pötzsch C, Oehme L, et al. Automatic volume delineation in oncological PET evaluation of a dedicated software tool and comparison with manual delineation in clinical data sets. *NuklearMedizin*. 2012;51:9-16.

19. Schaefer A, Vermandel M, Baillet C, et al. Impact of consensus contours from multiple PET segmentation methods on the accuracy of functional volume delineation. *Eur J Nucl Med Mol Imaging*. 2016;43:911-924.

20. Kanoun S, Tal I, Berriolo-Riedinger A, et al. Influence of software tool and methodological aspects of total metabolic tumor volume calculation on baseline [18F]FDG PET to predict survival in Hodgkin lymphoma. *PLoS One*. 2015;10:1-15.

21. Hatt M, Cheze Le Rest C, Albarghach N, Pradier O, Visvikis D. PET functional volume delineation: A robustness and repeatability study. *Eur J Nucl Med Mol Imaging*. 2011;38:663-672.

22. Frings V, Van Velden FHP, 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.

23. 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.

24. Im HJ, Bradshaw T, Solaiyappan M, Cho SY. Current methods to define metabolic tumor volume in positron emission tomography: Which one is better? *Nucl Med Mol Imaging*. 2018;52:5-15.

25. Barrington SF, Meignan M. Time to prepare for risk adaptation in lymphoma by standardizing measurement of metabolic tumor burden. *J Nucl Med*. 2019;60:1096-1102.

26. Berthon B, Marshall C, Evans M, Spezi E. ATLAAS: An automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. *Phys Med Biol*. 2016;61:4855-4869.

27. Burggraaff CN, Cornelisse AC, Hoekstra OS, et al. Interobserver agreement of interim and end-of-treatment 18F-FDG PET/CT in diffuse large B-cell lymphoma: Impact on clinical practice and trials. *J Nucl Med*. 2018;59:1831-1836.

28. 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;38:3377-3387.

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

30. Cheebsumon P, Yaqub M, Van Velden FHP, Hoekstra OS, Lammertsma AA, Boellaard R. Impact of [18F]FDG PET imaging parameters on automatic tumour delineation: Need for improved tumour delineation methodology. *Eur J Nucl Med Mol Imaging*. 2011;38:2136-2144.

31. Frings V, De Langen AJ, Smit EF, et al. Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. *J Nucl Med*. 2010;51:1870-1877.

32. de Vet HCW, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's κ. *Bmj.* 2013;346:f2125.

33. de Vet HCW, Mullender MG, Eekhout I. Specific agreement on ordinal and multiple nominal outcomes can be calculated for more than two raters. *J Clin Epidemiol*. 2018;96:47-53.

34. Wu X, Pertovaara H, Korkola P, et al. Early interim PET/CT predicts post-treatment response in diffuse large B-cell lymphoma. *Acta Oncol*. 2014;53:1093-1099.

35. Boellaard R, Delgado-Bolton R, Oyen WJG, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42:328-354.

36. Pfaehler E, Burggraaff C, Kramer G, et al. PET segmentation of bulky tumors: Strategies and

workflows to improve inter-observer variability. *PLoS One*. 2020;15:e0230901.

37. Dewalle-Vignion AS, Betrouni N, Baillet C, Vermandel M. Is STAPLE algorithm confident to assess segmentation methods in PET imaging? *Phys Med Biol*. 2015;60:9473-9491.

38. Weisman AJ, Kieler M, Perlman S, et al. Comparison of 11 automated PET segmentation methods in lymphoma. *Phys Med Biol*. 2020. Epub ahead of print.

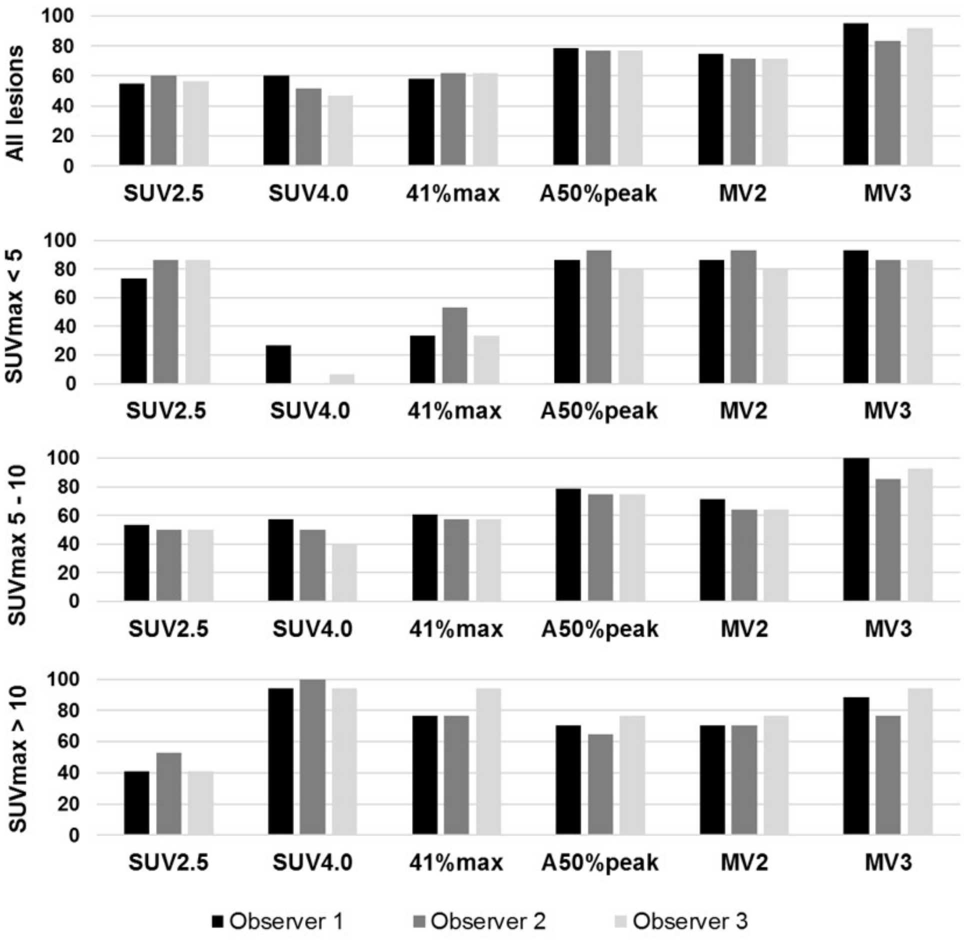


FIGURE 1. Frequency of delineation quality scores ("acceptable") per delineation method as rated by the observers.

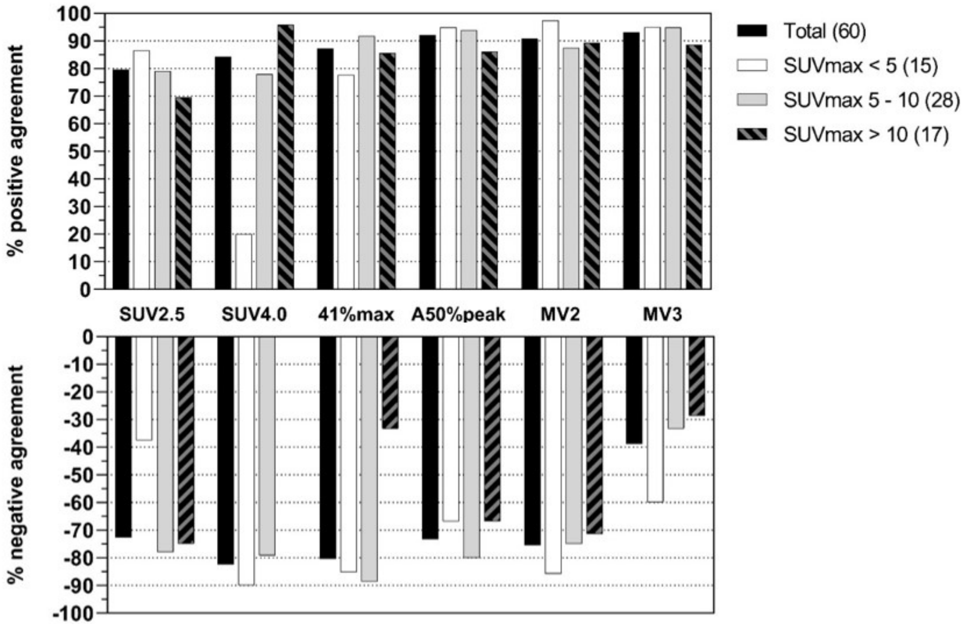


FIGURE 2. Positive and negative agreement on delineation quality score "acceptable" as function of lesional SUVmax per method.

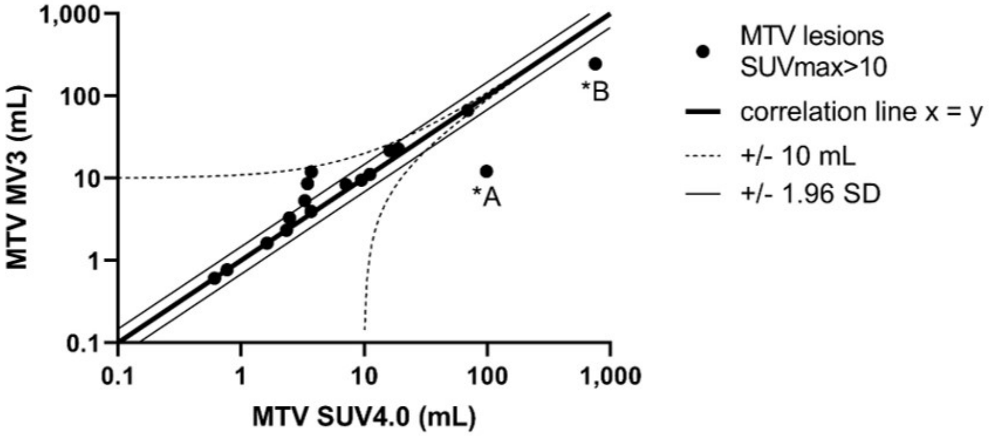
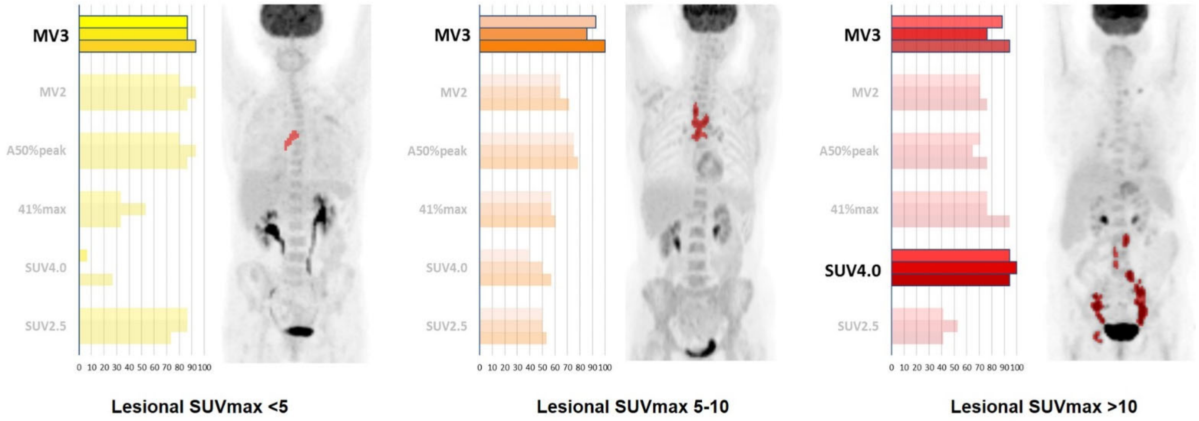


FIGURE 3. Scatterplot of MTVs derived from MV3 versus SUV4.0 for lesions with SUVmax>10. MTVs *A and *B applied by MV3 and SUV4.0 differed more than 10 mL and 1.96 standard deviation (SD) from the line of identity (see *A, *B demonstrated in Supplemental Figure 2).

GRAPHICAL ABSTRACT

Delineation quality of metabolic tumor volume measurements at interim-PET in DLBCL
as scored by 3 observers



Implications Automated estimation of MTV of DS4-5 DLBCL lesions at interim-PET is feasible in clinical practice in a supervised manner by using MV3 and optionally SUV4.0 for very high avid lesions.

TABLE 1. Descriptive of observed MTVs per method

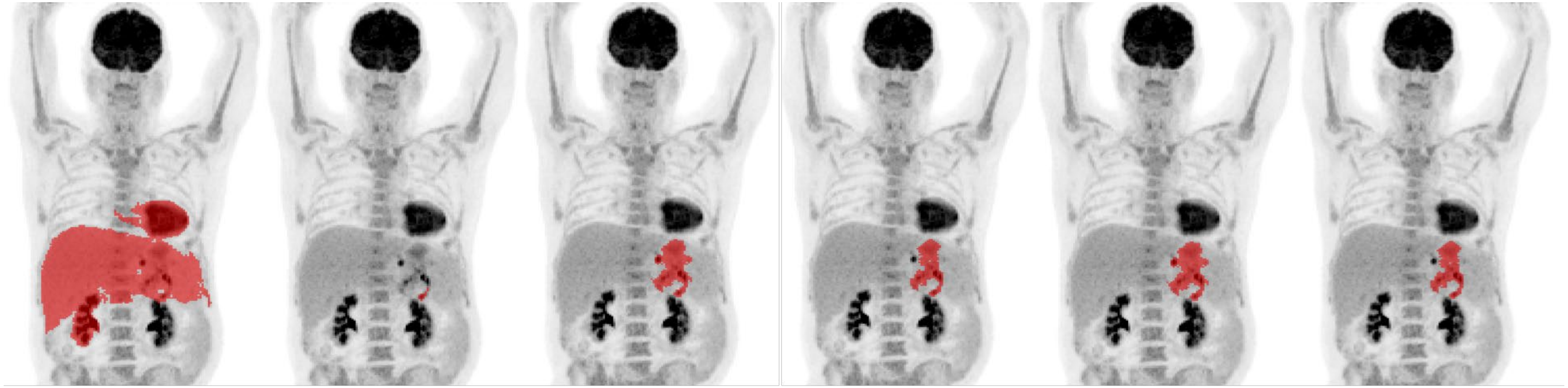
MTV per method (mL)	Median	Interquartile Range
SUV2.5	29.9	4.8 - 181.4
SUV4.0	2.3	0.6 - 9.7
41%max	27.6	3.5 - 214.6
A50%peak	14.7	3.7 - 37.3
MV2	19.4	5.7 - 65.0
MV3	4.9	3.1 - 19.8

TABLE 2. Frequency of delineation quality scores and specific agreements

Frequency of scores ("acceptable") per delineation method		SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3	*Method of preference approach*
Average percentage	Total (60 lesions)	57.2	52.8	60.6	77.2	72.8	90.0	**92.8**
	SUVmax <5 (15)	82.2	11.1	40.0	86.7	86.7	88.9	88.9
	SUVmax 5-10 (28)	51.2	48.8	58.3	76.2	66.7	92.9	92.9
	SUVmax >10 (17)	45.1	96.1	82.4	70.6	72.5	86.3	96.1

Specific Agreement; acceptable vs non-acceptable		SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3	*Method of preference approach*
Percentage PA	Total (60 lesions)	79.6	84.2	87.2	92.1	90.8	93.2	**95.2**
	SUVmax <5 (15)	86.5	20.0	77.8	94.9	97.4	95.0	95.0
	SUVmax 5-10 (28)	79.1	78.0	91.8	93.8	87.5	94.9	94.9
	SUVmax >10 (17)	69.6	95.9	85.7	86.1	89.2	88.6	95.9
Percentage NA	Total (60)	72.7	82.4	80.3	73.2	75.5	38.9	**30.5**
	SUVmax <5 (15)	37.5	90.0	85.2	66.7	85.7	60.0	60.0
	SUVmax 5-10 (28)	78.0	79.1	88.6	80.0	75.0	33.3	33.3
	SUVmax >10 (17)	75.0	0.0	33.3	66.7	71.4	28.6	0.0

Abbreviations: PA: positive agreement, NA: negative agreement



SUV 2.5 (3427 ml)
“failed”

SUV 4.0 (2.13 ml)
“poor”

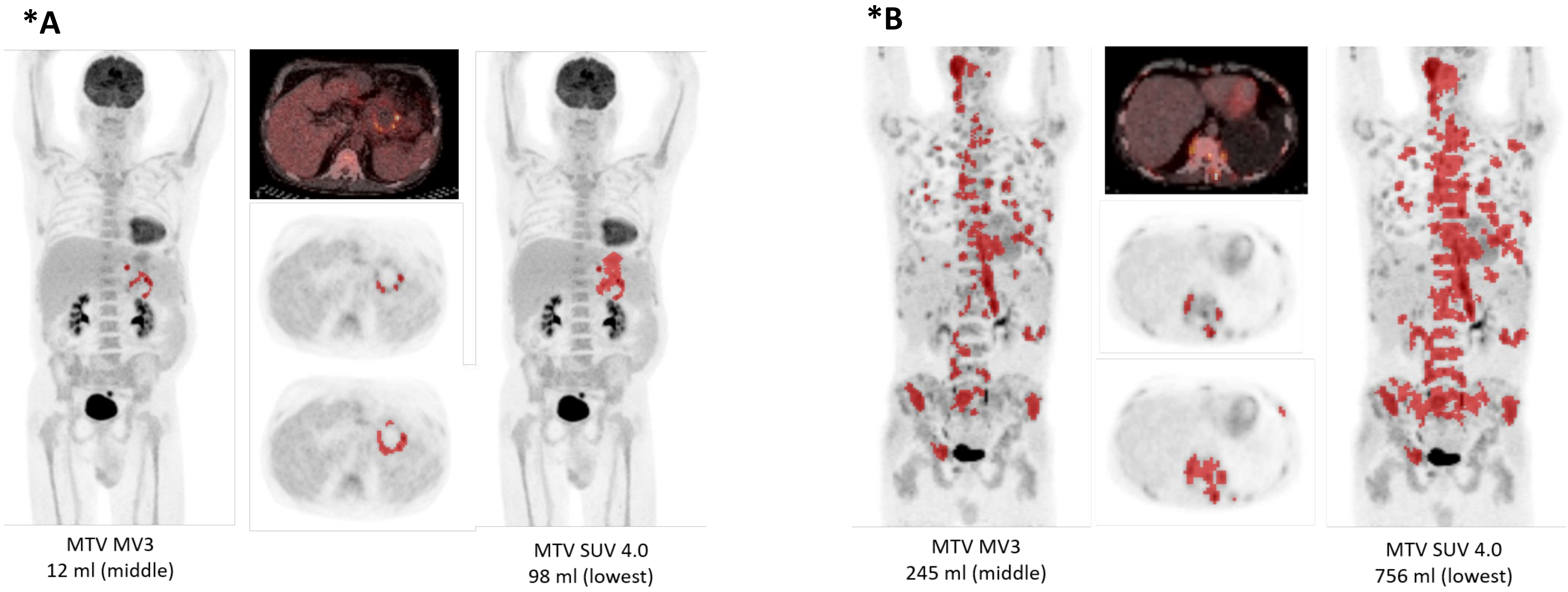
41%max (143 ml)
“good”

A50%peak (96 ml)
“moderate”

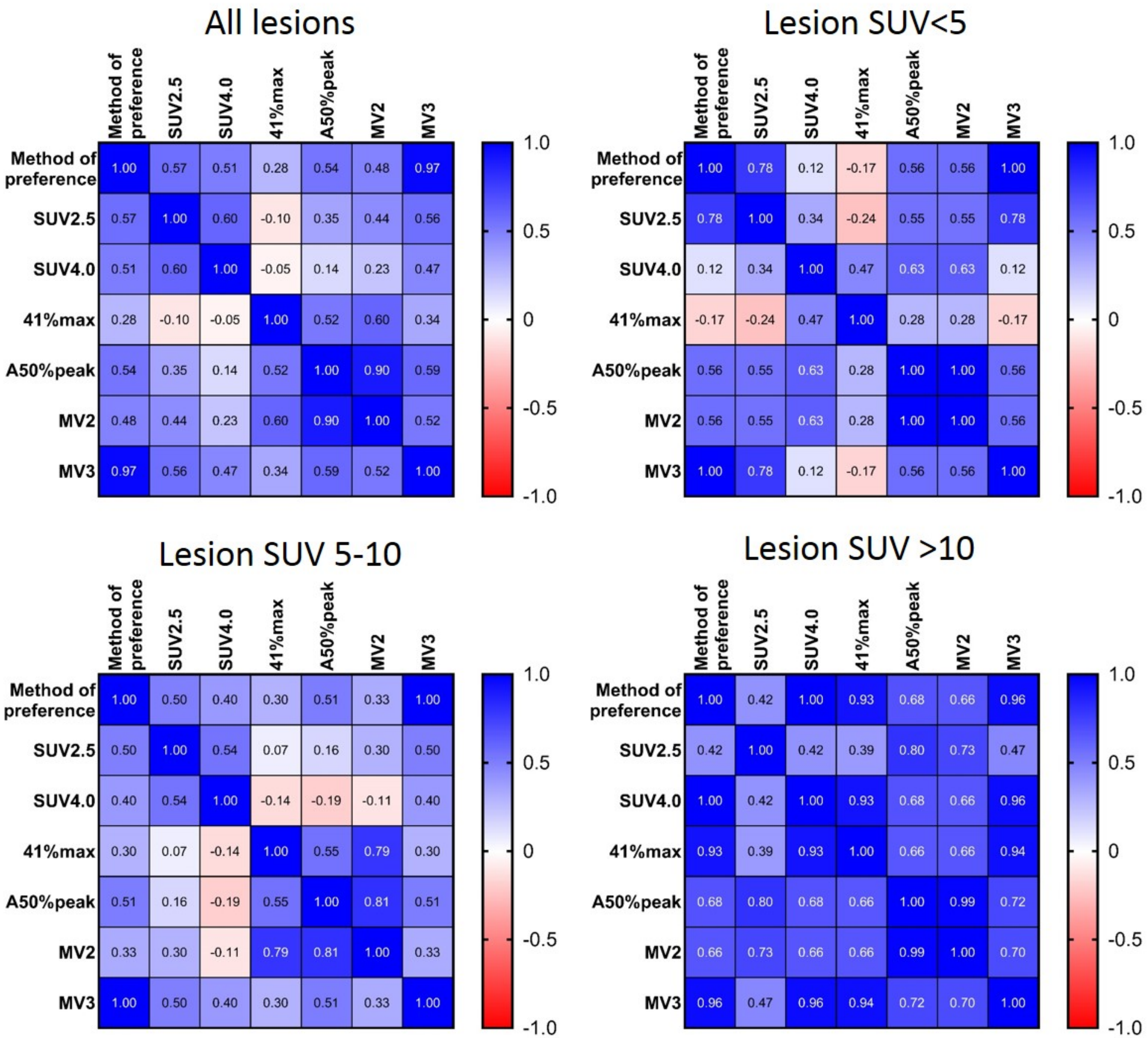
MV2 (143 ml)
“good”

MV3 (96 ml)
“moderate”

Supplemental FIGURE 1. Example scores per method derived MTV in a casus of DLBCL located in stomach, DS5. Scores “good” and “moderate” were considered as acceptable delineations whereas “poor” and “failed” were not.



Supplemental FIGURE 2. The delineation quality of MV3 derived MTVs *A and *B were rated as “ non-acceptable” and were considered as providing underestimation of tumor volume by all observers. All observers scored SUV4.0 delineated MTV *A as “acceptable” while one observer scored MTV *B as “non-acceptable” based on missing tumor lesions and delineating physiological activity in left kidney.


All lesions

	Method of preference	SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Method of preference	1.00	0.57	0.51	0.28	0.54	0.48	0.97
SUV2.5	0.57	1.00	0.60	-0.10	0.35	0.44	0.56
SUV4.0	0.51	0.60	1.00	-0.05	0.14	0.23	0.47
41%max	0.28	-0.10	-0.05	1.00	0.52	0.60	0.34
A50%peak	0.54	0.35	0.14	0.52	1.00	0.90	0.59
MV2	0.48	0.44	0.23	0.60	0.90	1.00	0.52
MV3	0.97	0.56	0.47	0.34	0.59	0.52	1.00

Lesion SUV<5

	Method of preference	SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Method of preference	1.00	0.78	0.12	-0.17	0.56	0.56	1.00
SUV2.5	0.78	1.00	0.34	-0.24	0.55	0.55	0.78
SUV4.0	0.12	0.34	1.00	0.47	0.63	0.63	0.12
41%max	-0.17	-0.24	0.47	1.00	0.28	0.28	-0.17
A50%peak	0.56	0.55	0.63	0.28	1.00	1.00	0.56
MV2	0.56	0.55	0.63	0.28	1.00	1.00	0.56
MV3	1.00	0.78	0.12	-0.17	0.56	0.56	1.00

Lesion SUV 5-10

	Method of preference	SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Method of preference	1.00	0.50	0.40	0.30	0.51	0.33	1.00
SUV2.5	0.50	1.00	0.54	0.07	0.16	0.30	0.50
SUV4.0	0.40	0.54	1.00	-0.14	-0.19	-0.11	0.40
41%max	0.30	0.07	-0.14	1.00	0.55	0.79	0.30
A50%peak	0.51	0.16	-0.19	0.55	1.00	0.81	0.51
MV2	0.33	0.30	-0.11	0.79	0.81	1.00	0.33
MV3	1.00	0.50	0.40	0.30	0.51	0.33	1.00

Lesion SUV >10

	Method of preference	SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Method of preference	1.00	0.42	1.00	0.93	0.68	0.66	0.96
SUV2.5	0.42	1.00	0.42	0.39	0.80	0.73	0.47
SUV4.0	1.00	0.42	1.00	0.93	0.68	0.66	0.96
41%max	0.93	0.39	0.93	1.00	0.66	0.66	0.94
A50%peak	0.68	0.80	0.68	0.66	1.00	0.99	0.72
MV2	0.66	0.73	0.66	0.66	0.99	1.00	0.70
MV3	0.96	0.47	0.96	0.94	0.72	0.70	1.00

Supplemental FIGURE 3. Spearman correlation coefficients between the method derived MTVs per category of lesional SUVmax. None of the log transformed method derived volumes of lesions with a SUVmax <10 were normally distributed and/or showed high Spearman correlation coefficients. For lesions with SUVmax>10, only 50%peak and MV2 derived MTVs were not normally distributed after log transformation.

SUPPLEMENTAL TABLE 1

Acceptable score frequency per method		SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Percentag	Total (60 lesions)	57.2	52.8	60.6	77.2	72.8	90.0
	SUVmax > 5 (45)	48.9	66.7	67.4	74.1	68.9	90.4
	SUVmax < 5 (15)	82.2	11.1	40.0	86.7	86.7	88.9
	SUVmax 5 - 10 (28)	51.2	48.8	58.3	76.2	66.7	92.9
	SUVmax > 10 (17)	45.1	96.1	82.4	70.6	72.5	86.3
	SUVmax < 10 (43)	62.0	35.7	51.9	79.8	72.9	91.5
	SUVmax 10 - 15 (11)	42.4	93.9	81.8	84.8	84.8	87.9
	SUVmax > 15 (6)	50.0	100.0	83.3	44.4	50.0	83.3
	SUVmax < 15 (54)	58.0	47.5	58.0	80.9	75.3	90.7

Good score frequency per method (of original four scoring variables)		SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Percentag	Total (60 lesions)	21.7	20.6	26.7	38.3	30.0	48.9
	SUVmax > 5 (45)	13.3	25.9	31.9	37.8	26.7	45.9
	SUVmax < 5 (15)	46.7	4.4	11.1	40.0	40.0	57.8
	SUVmax 5 - 10 (28)	15.5	19.0	25.0	46.4	23.8	54.8
	SUVmax > 10 (17)	9.8	37.3	43.1	23.5	31.4	31.4
	SUVmax < 10 (43)	26.4	14.0	20.2	44.2	29.5	55.8
	SUVmax 10 - 15 (11)	12.1	36.4	45.5	30.3	39.4	33.3
	SUVmax > 15 (6)	5.6	38.9	38.9	11.1	16.7	27.8
	SUVmax < 15 (54)	23.5	18.5	25.3	41.4	31.5	51.2

Specific Agreement; acceptable vs non-acceptable		SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Percentag	Total (60 lesions)	79.6	84.2	87.2	92.1	90.8	93.2
	SUVmax > 5 (45)	75.8	87.8	89.0	91.0	88.2	92.6
	SUVmax < 5 (15)	**86.5**	**20.0**	**77.8**	**94.9**	**97.4**	**95.0**
	SUVmax 5 - 10 (28)	**79.1**	**78.0**	**91.8**	**93.8**	**87.5**	**94.9**
	SUVmax > 10 (17)	**69.6**	**95.9**	**85.7**	**86.1**	**89.2**	**88.6**
	SUVmax < 10 (43)	82.5	71.7	88.1	94.2	91.5	94.9
	SUVmax 10 - 15 (11)	71.4	93.5	85.2	85.7	85.7	89.7
	SUVmax > 15 (6)	66.7	100.0	86.7	87.5	100.0	86.7
	SUVmax < 15 (54)	80.9	80.5	87.2	92.4	90.2	93.9

Percentag	Total (60)	72.7	82.4	80.3	73.2	75.5	38.9
	SUVmax > 5 (45)	76.8	75.6	77.3	74.3	73.8	30.8
	SUVmax < 5 (15)	37.5	90.0	85.2	66.7	85.7	60.0
	SUVmax 5 - 10 (28)	78.0	79.1	88.6	80.0	75.0	33.3
	SUVmax > 10 (17)	75.0	0.0	33.3	66.7	71.4	28.6
	SUVmax < 10 (43)	71.4	84.3	87.1	76.9	77.1	45.5
	SUVmax 10 - 15 (11)	78.9	0.0	33.3	20.0	20.0	25.0
	SUVmax > 15 (6)	66.7	0.0	33.3	90.0	100.0	33.3
	SUVmax < 15 (54)	73.5	82.4	82.4	67.7	70.0	40.0

Specific Agreement; good vs others		SUV2.5	SUV4.0	41%max	A50%peak	MV2	MV3
Percentag	Total (60 lesions)	33.3	32.4	47.9	49.3	37.0	58.0
	SUVmax > 5 (45)	11.1	34.3	51.2	49.0	27.8	54.8
	SUVmax < 5 (15)	52.4	0.0	20.0	50.0	55.6	65.4
	SUV 5 - 10 (28)	15.4	37.5	57.1	59.0	30.0	65.2
	SUVmax > 10 (17)	0.0	31.6	45.5	16.7	25.0	25.0
	SUVmax < 10 (43)	38.2	33.3	50.0	56.1	42.1	65.3
	SUVmax 10 - 15 (11)	0.0	25.0	53.3	20.0	23.1	27.3
	SUVmax > 15 (6)	0.0	42.9	28.6	0.0	33.3	20.0
	SUVmax < 15 (54)	34.2	30.0	51.2	50.7	37.3	60.2
Percentag	Total (60)	81.6	82.5	81.1	68.5	73.0	59.8
	SUVmax > 5 (45)	86.3	77.0	77.2	69.0	73.7	61.6
	SUVmax < 5 (15)	58.3	95.3	90.0	66.7	70.4	52.6
	SUVmax 5 - 10 (28)	84.5	85.3	85.7	64.4	78.1	57.9
	SUVmax > 10 (17)	89.1	59.4	58.6	74.4	65.7	65.7
	SUVmax < 10 (43)	77.9	89.2	87.4	65.3	75.8	56.1
	SUVmax 10 - 15 (11)	86.2	57.1	61.1	65.2	50.0	63.6
	SUVmax > 15 (6)	94.1	63.6	54.5	87.5	86.7	69.2
	SUVmax < 15 (54)	79.8	84.1	83.5	65.3	71.2	58.2

Abbreviations: PA: positive agreement, NA: negative agreement