DOI:
10.2967/jnumed.119.238923

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Barrington, S. F., Zwezerijnen, B. G., de Vet, H. C., Heymans, M. W., Mikhaeel, N. G., Burggraaff, C. N., Eertink, J. J., Pike, L. C., Hoekstra, O. S., Zijlstra, J. M., & Boellaard, R. (2020). Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful ? *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*. Advance online publication. https://doi.org/10.2967/jnumed.119.238923

Journal of Nuclear Medicine, published on July 17, 2020 as doi:10.2967/jnumed.119.238923

Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful ?

A study on behalf of the PETRA consortium.

Sally F Barrington[1], Ben GJC Zwezerijnen[2], Henrica C W de Vet[3], Martijn W Heymans[3], N George Mikhaeel[4], Coreline N Burggraaff[5], Jakoba J Eertink[5], Lucy C Pike[1], Otto S Hoekstra[2], Josée M Zijlstra[5], Ronald Boellaard[2].

[1] King's College London and Guy's and St Thomas' PET Center, School of Biomedical Engineering and Imaging Sciences, Kings College London, UK, [2] Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, de Boelelaan 1117, Amsterdam, Netherlands [3] Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, de Boelelaan 1089A, Amsterdam, Netherlands [4]Department of Clinical Oncology, Guy's and St Thomas' NHS Foundation Trust and School of Cancer & Pharmaceutical Sciences, Kings College London, [5] Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Hematology, Cancer Center Amsterdam, de Boelelaan 1117, Amsterdam, Netherlands

Corresponding author: Sally Barrington, St Thomas Hospital, London SE1 7EH UK

00 44 207 188 8364 (phone) 00 44 207 620 0790 (fax) ORCID ID 0000-0002-2516-5288

sally.barrington@kcl.ac.uk

Running title MTV in DLBCL: which method?

Word count: 4993

ABSTRACT

Introduction: Metabolic tumor volume (MTV) is a promising biomarker of pretreatment risk in diffuse large B-cell lymphoma (DLBCL). Different segmentation methods can be used which predict prognosis equally well but give different optimal cut-offs for risk stratification. Segmentation can be cumbersome meaning a fast, easy and robust method is needed. **Aims** were to i) evaluate the best automated MTV workflow in DLBCL ii) determine if uptake time, (non)compliance with standardized recommendations for FDG scanning and subsequent disease progression influenced the success of segmentation iii) assess differences in MTV values and discriminatory power of segmentation methods. **Methods**: 140 baseline FDG-PET/CT scans were selected from UK and Dutch studies in DLBCL to provide a balance between scans at 60- or 90-minutes uptake, parameters compliant or non-compliant with standardized recommendations for scanning and patients with or without progression. An automated tool was used for segmentation using i) standardized uptake value (SUV) 2.5 ii) SUV 4.0 iii) adaptive thresholding [A50P] iv) 41% of maximum SUV [41%] v) majority vote including voxels detected by ≥2 methods [MV2] and vi) detected by ≥3 methods [MV3]. Two independent observers rated the success of the tool to delineate MTV. Scans that required minimal interaction were rated "success"; scans where > 50% of tumor was missed or required more than 2 editing steps were rated as "failure". **Results**: 138 scans were evaluable, with significant differences in success and failure ratings between methods. The best performing was SUV4.0, with higher success and lower failure rates than all other methods except MV2 which also performed well. SUV4.0 gave a good approximation of MTV in 105 (76%) scans, with simple editing for a satisfactory result in additionally 20% of cases. MTV was significantly different for all methods between patients with and without progression. SUV41% performed slightly worse with longer uptake times, otherwise scanning conditions and patient outcome did not influence the tool's performance. The discriminative power of methods was similar, but MTV values were significantly greater using SUV4.0 and MV2 than other thresholds except for SUV2.5.

Conclusion: SUV4.0 and MV2 are recommended for further evaluation. Automated estimation of MTV is feasible.

Key words: lymphoma, metabolic tumor volume, positron emission tomography, standardization

INTRODUCTION

Metabolic tumor burden assessed with 18F-fluorodeoxyglucose positron emission tomography (FDG-PET) is a promising biomarker for pretreatment risk in lymphoma *(1-8)*. Published reports have used different methods to measure metabolic tumor volume (MTV) and tumor lesion glycolysis (TLG), which is the product of MTV and the mean standardized uptake value (SUV) *(9-11)*.

Measurement of MTV requires the observer to delineate tumor with uptake above a chosen threshold, which may be based on absolute SUV e.g. SUV 2.5 *(8,12,13)* or SUV 4.0 *(14,15)* or a percentage of the maximum SUV in each tumor region e.g. 25% *(7)* or 41% *(1,2,6)* which are summed together. For percentage methods, if counts vary by more than 10% within a heterogeneous tumor mass, the observer should subdivide it into parts *(16)* to avoid the situation where an intense area e.g. with a maximum SUV of 20, results in voxels with SUV ≤ 8.2 (41% of maximum SUV) being left out so that MTV is underestimated. Adaptive thresholding and other techniques that do not rely on fixed thresholds have been used in solid tumors *(17-19)* but not much in lymphoma *(9)*.

Tumor delineation can be time-consuming, especially in patients with lymphoma, who often have multiple and heterogeneous nodal and extranodal masses *(11)*. Sometimes the observer needs to edit tumor outlines to remove adjacent physiological uptake in urinary tract, brain and heart, because many software algorithms use a seed approach to group regions with similar uptake together for rapid outlining. The editing stage can introduce variation in delineation between observers *(20)*.

Quantitative measurements can be affected by patient preparation, image acquisition and reconstruction *(21)*. Significant efforts have been made to standardize FDG scanning in clinical trials including initiatives by the European Association for Nuclear Medicine Research Limited (EARL) *(22)* and the US Society of Nuclear Medicine *(23)*. Differences in clinical practice and clinical trials however still exist that affect quantitative estimates such as MTV. Despite these methodological issues, MTV is a robust predictor of progression-free-survival (PFS) and in some reports of overall survival in diffuse large B-cell lymphoma (DLBCL) *(8,12,24)* and other subtypes *(2,6,7,25)*. However, the median value and optimum cut-off that separates patients with high from low-risk disease is crucially dependent on the segmentation method, the patient population characteristics and efficacy of treatment *(26)*. Measurement of MTV can be cumbersome using current software approaches *(11)* and there is no agreed consensus on the best method. This has precluded assessment of MTV for risk stratification to date in multicenter trials *(20)*.

There is a clear unmet need to develop a standard method for MTV measurement in multicenter trials and ultimately for clinical practice *(20)*. Given that all methods appear to predict prognosis with equal effectiveness *(9,11)*, efforts should focus on developing a method with high success rates for outlining visible tumor that is quick and easy, gives consistent results and can be implemented in multiple software platforms. An automated approach to reduce user interaction and interobserver variation is desirable to achieve these goals. DLBCL is the most common lymphoma subtype and possibly the most challenging for MTV measurement as tumor is frequently disseminated and extranodal *(20)*.

The aims of this study were to a) evaluate the best method using an automated tool to measure MTV in DLBCL assessed by the success of segmentation of visible tumor b) determine if the success of the measurement method was influenced by the uptake time, (non)compliance with standardized recommendations for FDG scanning *(21)* and the presence/absence of progression or death at 2 years and c) to assess the differences in MTV values and discriminatory power obtained by different segmentation methods.

METHODS

PET-CT scans were selected from patients with newly diagnosed DLBCL scanned in research studies in the Netherlands (NL) and United Kingdom (UK). The scans are part of a comprehensive database to validate interim FDG-PET as a biomarker of response for non-Hodgkin lymphoma (https://petralymphoma.org/). Studies had approval by institutional review board and/or ethics committees. Scans were chosen to provide a balance between patients scanned

- using a protocol that specified 60-minute uptake (NL scans) or 90-minute uptake (UK scans)
- using reconstruction parameters that were compliant or non-compliant with standardized recommendations*(21)*
- with or without disease progression or death at 2 years

Software called 'ACCURATE' was used to automatically measure MTV on baseline scans *(27)*. It minimizes user interaction by automatically outlining tumor regions and allows multiple segmentation methods to be applied. Physiological uptake can be removed and lesions added, if required, using single clicks on MIP and volume images. Two independent readers performed measurements and rated the success or failure of methods/workflows to automatically delineate visible tumor, blinded to patient outcome. PET and CT datasets were displayed alongside one another with options to fuse datasets as required. The consensus ratings of the two readers were used in analyses.

The following segmentation methods were applied: SUV of 2.5 [SUV2.5], SUV of 4.0 [SUV4.0] , adaptive thresholding, using 50% of peak voxel value adapted for local background [A50P]*(28)*, 41% of maximum SUV [41%] majority vote segmenting voxels detected by ≥2 methods [MV2] and detected by ≥3 methods [MV3]*(29)* . Majority vote approaches were included as previous studies showed they may outperform single underlying standard methods *(30)* which may not necessarily be best for all lesions and patients. Each method was rated as to whether it was successful or failed in the task of automatic tumor delineation or required some additional but limited user interaction to edit the MTV (Table 1). More than 2 additional manual editing steps were considered not feasible for clinical practice and rated as a failure of the method.

Statistical analysis: A sample size of 140 allowed for 70 scans in each of the three subgroups (uptake time, EARL compliance, progression or death) and 35 if divided further as all subgroups were balanced, to allow for robust identification of differences larger than 20% in success and failure rates (significance level 0.05 and power 0.80).

Descriptive statistics were performed for all segmentation methods. Differences between success rates among the 6 segmentation methods were assessed using Chi square tests. Influence of uptake time, reconstruction method and progression on success and failure rates were also assessed by Chi square tests. Analyses on MTV values were performed on raw and natural logarithmic transformed data due to their non-normal distribution. To assess the agreement in MTV values between segmentation methods, Pearson correlation coefficients were determined. Influence of uptake time, reconstruction method and progression on MTV values obtained by the different methods were evaluated by t-tests. The discriminative power regarding progression and non-progression of the segmentation methods was assessed by comparing the mean volumes using t-tests and Receiver Operating Characteristic (ROC) curves. All analyses were performed with IBM SPSS version 22.

RESULTS

140 baseline PET-CT scans were assessed. Two patients without FDG-avid disease were excluded leaving 138 scans. The agreement between readers was excellent, with 91% agreement for SUV41% and over 95% for all other methods.

Performance of Different Segmentation Methods

There were significant differences in rating between methods. The best performing method was SUV4.0, with significantly higher success and lower failure rates than all other methods (P<0.005) except MV2 (Table 2). The SUV4.0 method gave good visual approximation of tumor burden in 105 (76%) scans, with minimal user interaction (Table 2). Editing was required to achieve satisfactory estimation of visible tumor in an additional 20% (27/138) comprising a single editing step in 21 patients and two steps in 6 patients. After editing, the volume was altered by less than 10% in 12 cases, between 10-25% in 9 cases and in 5 cases by more than 25%. The commonest reason for failure of the 41% and A50P methods was because more than half the visible tumor was not outlined (Figure 1)and for failure of the SUV2.5 method was because the automatic segmentation included physiological uptake that would require complex editing to remove (Figure 2).

Influence of Uptake Time, Reconstruction Method and Patient Outcome on Success and Failure Rates

When comparing different uptake times, the 41% (P<0.05) and MV3 (P<0.05) methods were more likely to fail with scans acquired at 90 minutes (Supplemental Table 1A). The uptake time had no influence on the success and failure rates of other methods to delineate MTV. There were no statistically significant differences in the performance of the methods whether scans complied with EARL recommendations or not (Supplemental Table 1B). All methods performed equally well in patients who died or progressed as patients who did not. (Supplemental Table 1C).

Comparison of MTV Values between Different Segmentation Methods.

MTV values were log-transformed to obtain a normal distribution. MTV values were greater using SUV4.0 and MV2 than other thresholds (Table 3) except for SUV2.5 which gave the largest volumes. These differences were statistically significant. A high correlation (r=0.72) was observed between values obtained using SUV4.0 and SUV2.5, but volumes obtained by methods 41% and A50P showed only moderate correlation (Figure 3) hence recalculating the volume obtained using one method by applying simple linear transformation to give the volume that would be obtained using another method is not possible. MV2 showed the highest correlation (r=0.94) with SUV4.0 and MV3 showed the highest correlation with 41% (r=0.94).

The means of MTV values were not statistically different for NL and UK patients (i.e. 60 vs. 90 min uptake) (Supplemental Table 2A) nor when comparing patients scanned using EARL recommendations or not for all segmentation methods (Supplemental table 2B). Patients who progressed or died had significantly higher MTV than patients who did not using all methods (Supplemental Table 2C). The discriminative power of all methods was similar (Figure 4).

DISCUSSION

The principle aim of this study was to determine the best segmentation method to measure MTV in DLBCL at baseline using automated software. MTV is a robust predictor of patient outcomes in DLBCL irrespective of the delineation method, but the absolute values for MTV and optimal cut-offs that divide good and poor prognosis groups differ *(11,26)*. Measurement of MTV in patients with DLBCL takes around 3-6 minutes per scan depending on the method, but complex cases can take 10-20 minutes *(11)*. There is no agreement presently about which method to use, however there is a consensus that reproducible and rapid automated measurements are needed to explore MTV for prognostic stratification in prospective trials and ultimately for clinical application *(20)*. MTV measurement methods can be assessed using simulated and phantom data where true volumes are known *(16)* to try and overcome challenges in segmentation of PET images with limited spatial resolution causing partial volume effects when developing contouring algorithms *(31)*. However phantoms are not representative of the clinical situation with varying contrast, heterogeneity, shapes and sizes of lesions and patients. Moreover, recent studies suggest that the actual MTV values resulting from using different methods do not affect prognostic performance *(10)*. The latter suggests that bias in observed MTV data is clinically less relevant than good reproducibility. Therefore we chose to rate the success of an automated tool with a fixed color table and SUV scale to delineate visible tumor satisfactorily in patients with DLBCL according to the opinion of experienced observers, which represents how it would be assessed in everyday practice *(32)*. We devised a method to rate the success or failure of the ACCURATE tool *a priori*. This is the first report to evaluate the success of an automated method in this way to our knowledge. Furthermore, we considered whether the choice of 'best' method was influenced by scanning conditions i.e. the uptake time and (non)compliance with standardized recommendations *(21)* and whether patients experienced later progression or not using a case-control design.

There were significant differences in performance for automated measurement of MTV with different segmentation methods. The best method for successful segmentation used a SUV threshold of 4.0 however the MV2 method also performed well. A majority vote method was included as no single method can be expected to perform optimally for every patient or every lesion, but a majority method is likely to provide a good approximation of tumor delineation that will be close to the best performing method in most patients. Consensus approaches using majority vote and the STAPLE algorithm are being explored in radiotherapy planning *(30)* and performed better than segmentations based on a single algorithm in an imaging analysis challenge to contour a large dataset comprising simulated, phantom and clinical images of solitary 'tumors' *(31)*.

Automatic delineation using SUV4.0 was successful in over three-quarters of patients using single clicks to remove uptake in brain, urinary tract and heart, if present. In the remainder an automated process was not completely reliable, generally because lesions were adjacent to areas with high physiological uptake requiring user interaction. For some cases, the volume was altered substantially during editing but for the vast majority only one or two additional steps were needed to obtain a reasonable estimate of MTV. The other segmentation methods did not perform as well. The

SUV2.5 method failed mostly because physiological uptake required complex editing to remove for a reasonable approximation of tumor burden. This was encountered in only one patient using the SUV4.0 method because spillover of counts into other tissues was less common and where it did occur, the overlap between tumor and physiological uptake was less extensive and required one or two editing steps. The 41% and A50P methods failed usually because of under-estimation of tumor due to heterogeneity with more than 50% of the tumor having uptake less than the chosen threshold. The 41% threshold performed slightly worse in patients scanned at 90 minutes, probably because uptake in tumor rises over time, meaning for heterogeneous tumors with areas of high uptake, fewer voxels would be included in 41% of the maximum SUV compounding the problem of underestimation. The influence of uptake time could possibly explain the preference of different groups for particular methods in reported studies *(20)*. Other scanning conditions did not influence how methods performed.

The absolute values for MTV varied between methods as previously reported *(20,26,33)* with a positive bias for SUV2.5 and a negative bias for the other methods compared to SUV4.0 and MV2, seen across the whole range of volumes. The SUV4.0 and MV2 methods performed in a similar way. The MV2 method selects voxels included in at least two segmentation methods (from SUV2.5, SUV4.0, A50, 41) and commonly included voxels delineated using the SUV4.0 threshold and the next method that came closest to delineating a similar volume, usually SUV2.5. The MV3 method selects voxels included in at least three of the segmentation methods and segmented similar volumes to the 41% method, which was most likely to delineate more of the same voxels as two or more of the other methods. The MV2 had similar performance to the SUV4.0 method but requires delineation using more than one segmentation method. This is fully automated within ACCURATE but could be less easy to implement across software platforms than a single SUV4.0 threshold method. Clinically available softwares currently can measure MTV using SUV4.0 although ACCURATE has additional features to enable the user to quickly review the

MIP image, add missed lesions or remove physiological uptake with single clicks, speeding up the segmentation process. Correlation between all the other thresholds was moderate or good but not sufficient to allow the MTV from different segmentation methods to be used interchangeably by a simple linear transformation.

All methods had similar discriminative power, as previously reported *(11)*. We used a case-control design, with an oversampling of patients with progression, so we cannot express results as positive and negative predictive values to decide which method is best for clinical use. Yet, our results seem to confirm previous findings *(9,11)* although the ROC curves demonstrated lower discriminative power than Ilyas et al *(11)*, possibly because the latter used cases from a single institution and manual editing was performed in the majority. Nonetheless the fact that all methods predicted prognosis equally suggests that selection of the best method can be based on success rate, ease of use and time/user interaction to obtain total tumor burden.

Limitations are that research software was used which is not yet widely available, but which has been designed as a tool that could be implemented across software platforms in discussion with manufacturers. Only 'classical' segmentation methods published in lymphoma datasets were assessed whereas more sophisticated methods may give more reliable estimations of tumor volume *(31)*. However, as we realized that a single method may not be able to reliably delineate all lesions for all patients, we included majority vote based approaches which have been shown to outperform single method segmentations. Yet, we observed that at baseline MTV measurements in DLBCL patients were equally feasible using MV2 and SUV4 methods. The assessment of MTV was made by two experienced observers with high concordance, mirroring high reproducibility reported by others *(11,25)*. The observers were aware which method was being applied however, blinding was not possible as the delineation method was often obvious. We have assumed that the cases evaluated are representative, however the case-control design meant that 50% of patients progressed whereas this would be lower in the clinic. This overrepresentation is likely however to accentuate the challenges of measuring MTV as patients who progress later would be expected to have higher disease burden and more extranodal disease than the average clinical population.

CONCLUSION

Automated estimation of MTV is feasible. SUV4.0 and possibly MV2 are recommended for further evaluation of baseline MTV in DLBCL in larger, unselected, multicenter datasets representative of all patients with DLBCL . The results are also likely to be applicable in other lymphoma subtypes. Further work will explore the association of MTV with clinical outcomes in a larger database within the PETRA consortium using the best methods evaluated in this study.

KEY POINTS

Questions i) what is the best automated workflow to measure MTV in DLBCL? ii) is the choice influenced by uptake time, non(compliance) with standardized recommendations for FDG scanning and subsequent progression? iii) do segmentation methods give different MTV values and/or discriminate between patient outcomes equally?

Findings i) The best automated workflow (judged by segmentation of visible tumor by experienced observers) was SUV 4.0 with significantly higher success and lower failure rates than other methods (SUV2.5, A50P, 41% maximum SUV and MV3) except MV2 which also performed well. ii) The choice of the best workflow was not influenced by use of standardized scanning recommendations or subsequent patient progression although the 41% method performed slightly worse with longer uptake times.

Implications for patient care Automated estimation of MTV is feasible in clinical practice using SUV4.0 and possibly also MV2.

ACKNOWLEDGEMENT

SFB acknowledges support from the National Institute for Health Research (NIHR) [RP-2-16-07-001]. King's College London and UCL Comprehensive Cancer Imaging Center is funded by the CRUK and EPSRC in association with the MRC and Department of Health and Social Care (England).The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The PETRA project is supported by the Alpe d'HuZes/KWF fund, provided by the Dutch Cancer Society (# VU 2012-5848).The authors thank patients and collaborating investigators who kindly supplied their data. No other potential conflicts of interest relevant to this article exist.

REFERENCES

1. Sasanelli M, Meignan M, Haioun C, et al. Pretherapy metabolic tumour volume is an independent predictor of outcome in patients with diffuse large B-cell lymphoma. *Eur J Nucl Med Mol Imaging*. 2014;41:2017-2022.

2. Meignan M, Cottereau AS, Versari A, et al. Baseline metabolic tumor volume predicts outcome in high-tumor-burden follicular lymphoma: A pooled analysis of three multicenter studies. *J Clin Oncol*. 2016;34:3618-3626.

3. Ceriani L, Martelli M, Zinzani PL, et al. Utility of baseline 18FDG-PET/CT functional parameters in defining prognosis of primary mediastinal (thymic) large B-cell lymphoma. *Blood*. 2015;126:950-956.

4. Pike LC, Kirkwood AA, Patrick P, et al. Can baseline PET-CT features predict outcomes in advanced hodgkin lymphoma? A prospective evaluation of UK patients in the RATHL trial (CRUK/07/033). *Hematol Oncol*. 2017;35:37-38.

5. Moskowitz AJ, Schoder H, Gavane S, et al. Prognostic significance of baseline metabolic tumor volume in relapsed and refractory hodgkin lymphoma. *Blood*. 2017;130:2196-2203.

6. Cottereau AS, El-Galaly TC, Becker S, et al. Predictive value of PET response combined with baseline metabolic tumor volume in peripheral T-cell lymphoma patients. *J Nucl Med*. 2018;59:589-595.

7. Ceriani L, Milan L, Martelli M, et al. Metabolic heterogeneity on baseline 18FDG-PET/CT scan is a predictor of outcome in primary mediastinal B-cell lymphoma. *Blood*. 2018;132:179-186.

8. Song MK, Yang DH, Lee GW, et al. High total metabolic tumor volume in PET/CT predicts worse prognosis in diffuse large B cell lymphoma patients with bone marrow involvement in rituximab era. *Leuk Res*. 2016;42:1-6.

9. Cottereau AS, Hapdey S, Chartier L, et al. Baseline total metabolic tumor volume measured with fixed or different adaptive thresholding methods equally predicts outcome in peripheral T cell lymphoma. *J Nucl Med*. 2017;58:276-281.

10. 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:e0140830.

11. 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.

12. 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.

13. Akhtari M, Milgrom SA, Pinnix CC, et al. Reclassifying patients with early-stage hodgkin lymphoma based on functional radiographic markers at presentation. *Blood*. 2018;131:84-94.

14. Guezennec C, Kirkwood AA, Pike LC, et al. Baseline PET features as predictors of outcome in advanced HL: A prospective evaluation of UK patients in the RATHL trial (CRUK/07/033). *Hemasphere*. 2018;2:T020(0067).

15. Kurtz DM, Green MR, Bratman SV, et al. Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood*. 2015;125:3679-3687.

16. Meignan M, Sasanelli M, Casasnovas RO, et al. Metabolic tumour volumes measured at staging in lymphoma: Methodological evaluation on phantom experiments and patients. *Eur J Nucl Med Mol Imaging*. 2014;41:1113-1122.

17. Daisne JF, Sibomana M, Bol A, Doumont T, Lonneux M, Gregoire V. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: Influence of reconstruction algorithms. *Radiother Oncol*. 2003;69:247-250.

18. Geets X, Lee JA, Bol A, Lonneux M, Grégoire V. A gradient-based method for segmenting FDG-PET images: Methodology and validation. *Eur J Nucl Med Mol Imaging.* 2007;34:1427-1438.

19. Hatt M, Laurent B, Fayad H, Jaouen V, Visvikis D, Le Rest CC. Tumour functional sphericity from PET images: Prognostic value in NSCLC and impact of delineation method. *Eur J Nucl Med Mol Imaging*. 2018;45:630-641.

20. 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.

21. 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.

22. Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: From daily practice to multicentre oncological studies. *Eur J Nucl Med Mol Imaging*. 2017;44:17-31.

23. Sunderland JJ, Christian PE. Quantitative PET/CT scanner performance characterization based upon the society of nuclear medicine and molecular imaging clinical trials network oncology clinical simulator phantom. *J Nucl Med*. 2015;56:145-152.

24. Tout M, Casasnovas O, Meignan M, et al. Rituximab exposure is influenced by baseline metabolic tumor volume and predicts outcome of DLBCL patients: A lymphoma study association report. *Blood*. 2017;129:2616-2623.

25. Cottereau AS, Versari A, Loft A, et al. Prognostic value of baseline metabolic tumor volume in early-stage hodgkin lymphoma in the standard arm of the H10 trial. *Blood*. 2018;131:1456-1463.

26. Schoder H, Moskowitz C. Metabolic tumor volume in lymphoma: Hype or hope? *J Clin Oncol*. 2016;34:3591-3594.

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

28. Cysouw MCF, Kramer GM, Hoekstra OS, et al. Accuracy and precision of partial-volume correction in oncological PET/CT studies. *J Nucl Med*. 2016;57:1642-1649.

29. 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.

30. 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.

31. Hatt M, Laurent B, Ouahabi A, et al. The first MICCAI challenge on PET tumor segmentation. *Med Image Anal*. 2018;44:177-195.

32. 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.

33. 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.

FIGURES

Figure 1 Case 1 was rated as successful using SUV4.0 (A) but as a failure with SUV41% (B) because it missed more than half the visible tumor.

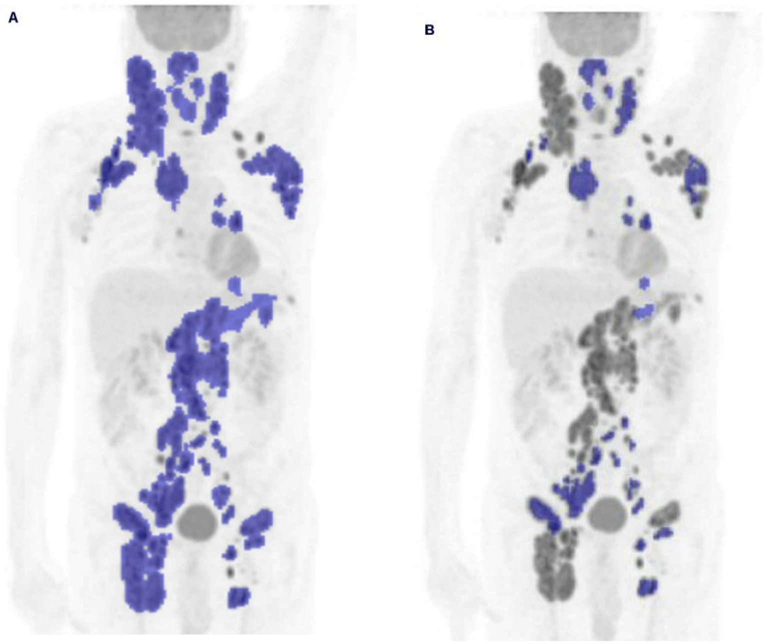


Figure 2 Case 2 shows a scan rated as successful with SUV4.0 (A) but as a failure with SUV2.5 (B) due to inclusion of physiological uptake requiring complex editing.

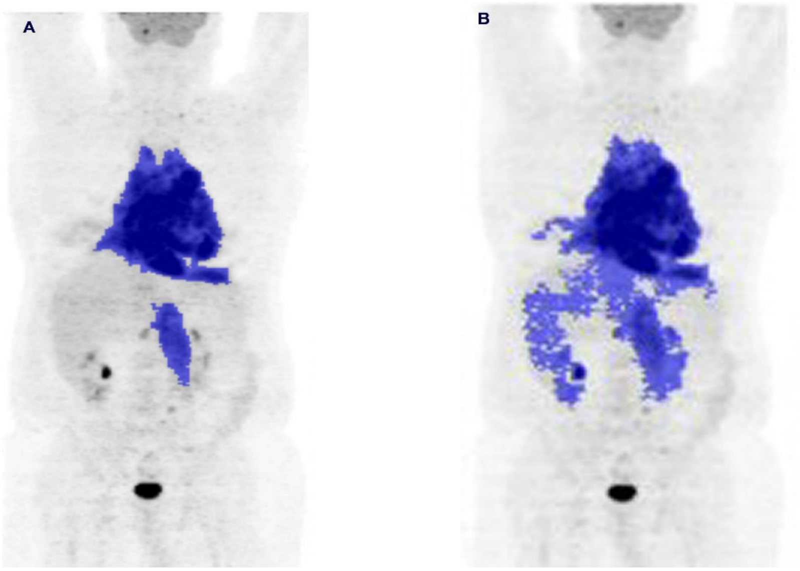


Figure 3: Distributions, scatterplots and correlations of segmentation methods.

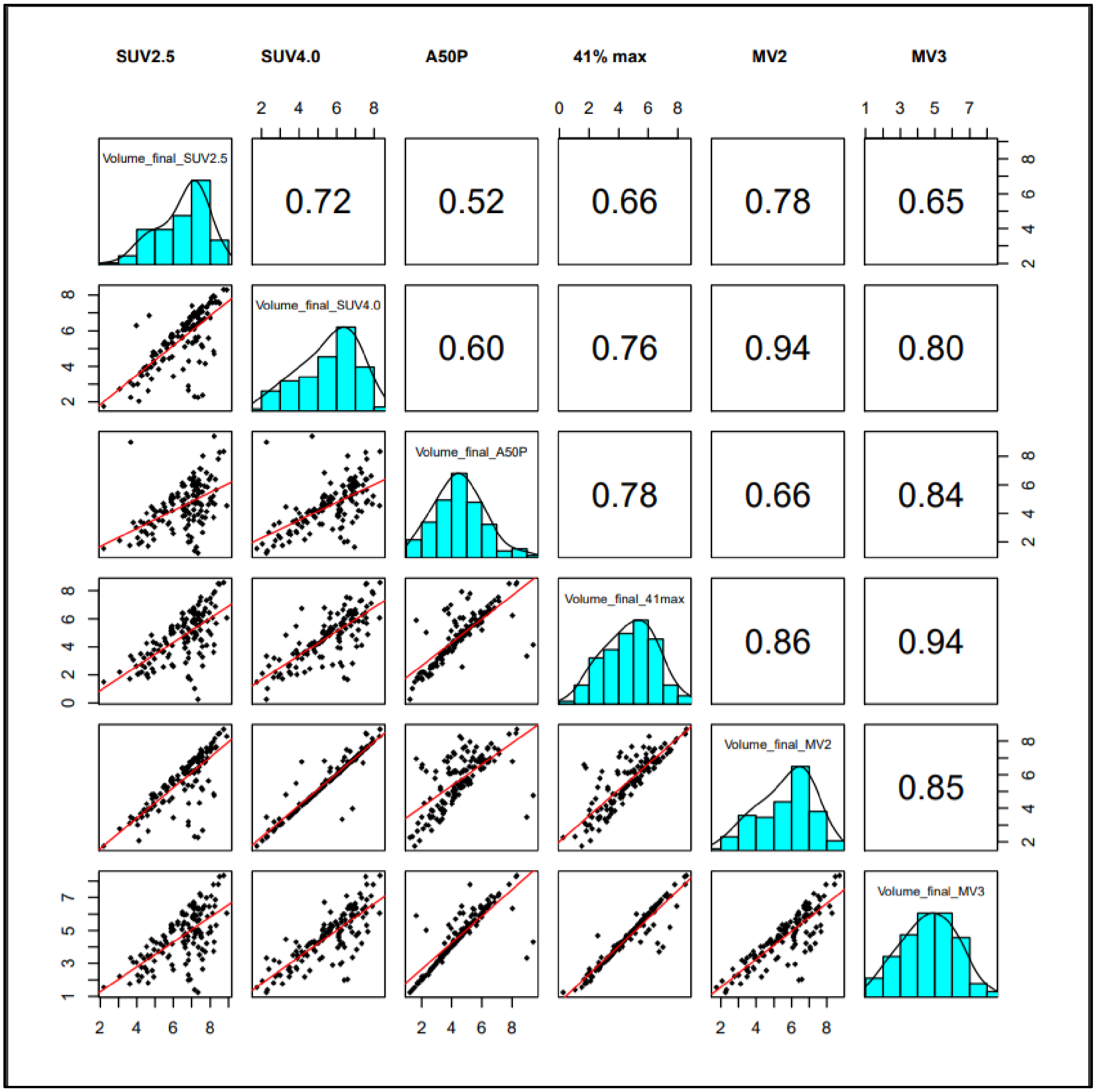
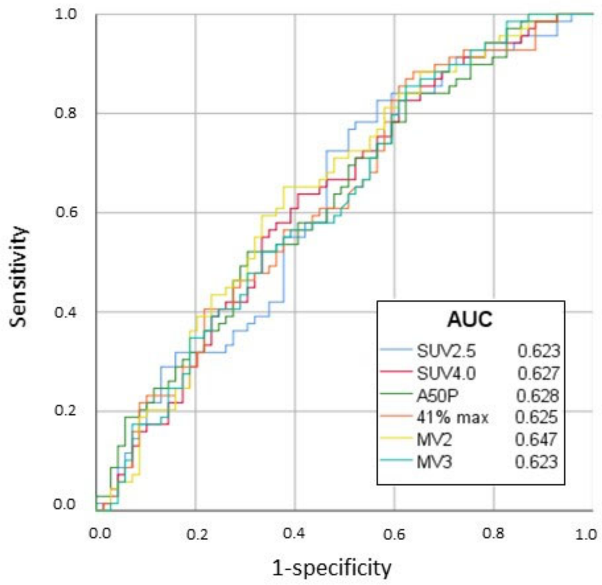


Figure 4: ROC curves of MTV values using different methods



Tables

Table 1: Definition of Success, Failure and Editing required

Rating	Findings
SUCCESS	No or minimal interaction by observer e.g. removing brain, bladder uptake or adding single region with single mouse clicks
FAILURE	Automatic segmentation missed more than half the visible tumor on scan or tumor 'flooded' into (also included) uptake in adjacent physiological structures that required complex slice-by-slice editing of 3 or more regions
EDITING required	Required 1 or 2 additional manual steps e.g. adding 'missed' regions or deleting slice-by-slice up to two regions of physiological uptake adjacent to tumor using an eraser tool (typically bladder and/or kidneys)

Table 2: Pairwise tests of segmentation methods using the SUV 4.0 method as reference.

Segmentation method	Success		Failure		Editing required
SUV 4.0	105		6		27
MV 2	102		10		26
MV 3	90		40	*	8
41%	82	*	45	*	11
A50P	75	*	57	*	6
SUV 2.5	51	*	57	*	30

P-value compared with SUV 4.0 method: * P<0.005

Table 3: Untransformed and log-transformed MTV values in cm^3 by segmentation method.

Segmentation method	Median volume	IQR	Mean log-volume	SD log-volume
SUV 4.0	311	75 ; 888	5.56	1.54
SUV2.5	906	255 ; 1616	6.45	1.34
41%	125	31 ; 398	4.73	1.73
A50P	87	29 ; 246	4.50	1.62
MV2	329	82 ; 921	5.66	1.55
MV3	109	32 ; 356	4.66	1.56

Supplemental Tables 1A –C: Success and failure rates for subgroups (for online publication only)

Table 1A; Success and failure rates for subgroups uptake time 60 versus 90 minutes

Uptake time 60 min (HOVON)

	PASS	FAIL	EDIT	TOTAL
41%	46	15	8	69
A50	41	26	2	69
SUV2.5	23	34	12	69
SUV4.0	53	2	14	69
MV2	53	2	14	69
MV3	51	13	5	69

Uptake time 90 min (UK scans)

	PASS	FAIL	EDIT	TOTAL
41%	36	30	3	69
A50	34	31	4	69
SUV2.5	28	23	18	69
SUV4.0	52	4	13	69
MV2	49	8	12	69
MV3	39	27	3	69

	X^2	df	P-value
41%	8.49	2	0.014
A50P	1.76	2	0.415
SUV2.5	3.81	2	0.149
SUV4.0	0.71	2	0.701
MV2	3.91	2	0.147
MV3	7.00	2	0.030

For 41%max and MV3 there were less PASS and more FAIL for the uptake time 90 min (UK scans)

Table 1B: Success and failure rates for subgroups EARL compatible or not.

Not EARL compatible

	PASS	FAIL	EDITING	TOTAL
41%	42	24	5	71
A50	37	30	4	71
SUV2.5	23	34	14	71
SUV4.0	58	2	11	71
MV2	55	5	11	71
MV3	49	19	3	71

EARL compatible

	PASS	FAIL	EDITING	TOTAL
41%	40	21	6	67
A50	38	27	2	67
SUV2.5	28	23	16	67
SUV4.0	47	4	16	67
MV2	47	5	15	67
MV3	41	21	5	67

	X^2	df	P-value
41%	0.22	2	0.896
A50	0.72	2	0.698
SUV2.5	2.63	2	0.269
SUV4.0	2.72	2	0.295
MV2	1.13	2	0.568
MV3	1.20	2	0.549

No statistically significant differences between non-EARL and EARL compatible.

Table 1C: Success and failure rates for subgroups progression or not

No progression

	PASS	FAIL	EDITING	TOTAL
41%	45	18	6	69
A50	43	23	3	69
SUV2.5	26	28	15	69
SUV4.0	51	4	14	69
MV2	51	5	13	69
MV3	50	16	3	69

Progression

	PASS	FAIL	EDITING	TOTAL
41%	37	27	5	69
A50	32	34	3	69
SUV2.5	25	29	15	69
SUV4.0	54	2	13	69
MV2	51	5	13	69
MV3	40	24	5	69

	X^2	df	P-value
41%	2.67	2	0.263
A50	3.74	2	0.154
SUV2.5	0.04	2	0.980
SUV4.0	0.79	2	0.674
MV2	0.00	2	1.000
MV3	3.21	2	0.200

No statistically significant differences in the ‘progression’ group.

Supplementary Tables 2A -C (for publication on line only)

Table 2A: Comparison of volumes (using transformed data) between scans performed at 60 min and 90 min

	Uptake time	N*	Mean	Std.Dev.	P-value
SUV2.5	60 min	69	6.47	1.33	0.847
	90 min	69	6.43	1.38	
SUV4.0	60 min	68	5.54	1.42	0.848
	90 min	68	5.59	1.66	
A50P	60 min	68	4.61	1.57	0.449
	90 min	68	4.40	1.67	
41%	60 min	68	4.73	1.47	0.999
	90 min	68	4.73	1.97	
MV2	60 min	68	5.56	1.40	0.481
	90 min	68	5.75	1.69	
MV3	60 min	68	4.69	1.28	0.813
	90 min	68	4.62	1.80	

*N= 136 for all methods except SUV2.5. In two scans SUV2.5 method identified a small MTV value where the other methods had a MTV value of 0. A value of 0 cannot be natural log-transformed.

Table 2B: Comparison of volumes (using transformed data) between scans that were compliant or non-compliant with standardized scanning recommendations

	EARL reconstruction	N*	Mean	Std.Dev.	P-value
SUV2.5	Non-compliant	71	6.63	1.26	0.110
	Compliant	67	6.26	1.41	
SUV4.0	Non-compliant	70	5.54	1.60	0.891
	Compliant	66	5.58	1.48	
A50P	Non-compliant	70	4.53	1.65	0.816
	Compliant	66	4.47	1.59	
41%	Non-compliant	70	4.64	1.76	0.529
	Compliant	66	4.82	1.71	
MV2	Non-compliant	70	5.65	1.56	0.982
	Compliant	66	5.66	1.56	
MV3	Non-compliant	70	4.61	1.56	0.733
	Compliant	66	4.71	1.57	

*N= 136 for all methods except SUV2.5. In two scans SUV2.5 method identified a small MTV value where the other methods had a MTV value of 0. A value of 0 cannot be natural log-transformed

Table 2C: Comparison of volumes (using transformed data) between patient outcome: with and without progression

	Outcome	N*	Mean	Std.Dev.	P-value
SUV2.5	No progression	69	6.11	1.46	0.003
	Progression	69	6.79	1.13	
SUV4.0	No progression	67	5.23	1.64	0.013
	Progression	69	5.88	1.36	
A50P	No progression	67	4.16	1.57	0.014
	Progression	69	4.84	1.60	
41%	No progression	67	4.37	1.83	0.018
	Progression	69	5.07	1.56	
MV2	No progression	67	5.26	1.67	0.003
	Progression	69	6.04	1.34	
MV3	No progression	67	4.35	1.68	0.021
	Progression	69	4.96	1.37	

*N= 136 for all methods except SUV2.5. In two scans SUV2.5 method identified a small MTV value where the other methods had a MTV value of 0. A value of 0 cannot be natural log-transformed.

JNM The Journal of NUCLEAR MEDICINE

Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful ?

Sally F Barrington, Ben GJC Zwezerijnen, Henrica CW de Vet, Martijn W Heymans, N George Mikhaeel, Coreline N Burggraaff, Jakoba J Eertink, Lucy C Pike, Otto S Hoekstra, Josee M Zijlstra and Ronald Boellaard

J Nucl Med.
Published online: July 17, 2020.
Doi: 10.2967/jnumed.119.238923

The Journal of Nuclear Medicine is published monthly.
SNMMI | Society of Nuclear Medicine and Molecular Imaging
1850 Samuel Morse Drive, Reston, VA 20190.
(Print ISSN: 0161-5505, Online ISSN: 2159-662X)

SNMMI SOCIETY OF NUCLEAR MEDICINE AND MOLECULAR IMAGING