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 G.J.C. 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], and Ronald Boellaard[2]

[1]King's College London and Guy's and St. Thomas' PET Center, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom; [2]Department of Radiology and Nuclear Medicine, Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Netherlands; [3]Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Netherlands; [4]Department of Clinical Oncology, Guy's and St. Thomas' NHS Foundation Trust and School of Cancer and Pharmaceutical Sciences, King's College London, London, United Kingdom; and [5]Department of Hematology, Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Metabolic tumor volume (MTV) is a promising biomarker of pretreatment risk in diffuse large B-cell lymphoma (DLBCL). Different segmentation methods can be used that predict prognosis equally well but give different optimal cutoffs for risk stratification. Segmentation can be cumbersome; a fast, easy, and robust method is needed. Our aims were to evaluate the best automated MTV workflow in DLBCL; determine whether uptake time, compliance or noncompliance with standardized recommendations for ^{18}F-FDG scanning, and subsequent disease progression influence the success of segmentation; and assess differences in MTVs and discriminatory power of segmentation methods. **Methods:** One hundred forty baseline ^{18}F-FDG PET/CT scans were selected from U.K. and Dutch studies on DLBCL to provide a balance between scans at 60 and 90 min of uptake, parameters compliant and noncompliant with standardized recommendations for scanning, and patients with and without progression. An automated tool was applied for segmentation using an SUV of 2.5 (SUV2.5), an SUV of 4.0 (SUV4.0), adaptive thresholding (A50P), 41% of SUV_{max} (41%), a majority vote including voxels detected by at least 2 methods (MV2), and a majority vote including voxels detected by at least 3 methods (MV3). Two independent observers rated the success of the tool to delineate MTV. Scans that required minimal interaction were rated as a success; scans that missed more than 50% of the tumor or required more than 2 editing steps were rated as a failure. **Results:** One hundred thirty-eight scans were evaluable, with significant differences in success and failure ratings among methods. The best performing was SUV4.0, with higher success and lower failure rates than any other method 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. The 41% segmentation method performed slightly worse, with longer uptake times; otherwise, scanning conditions and patient outcome did not influence the tool's performance. The discriminative power was similar among methods, but MTVs were significantly greater using SUV4.0 and MV2 than using 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; PET; standardization

J Nucl Med 2021; 62:332–337
DOI: 10.2967/jnumed.119.238923

Metabolic tumor burden assessed with ^{18}F-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, which is the product of MTV and SUV_{mean} (*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., an SUV of 2.5 (*8,12,13*) or 4.0 (*14,15*)) or a percentage of the SUV_{max} 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 a situation in which an intense area, such as with a SUV_{max} of 20, causes exclusion of voxels with an SUV of 8.2 or less (41% of SUV_{max}) 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 physiologic uptake in the urinary tract, brain, and heart, because many software algorithms use a seed approach to group regions with similar uptake for rapid outlining. The editing stage can introduce variation in delineation between observers (*20*).

Received Oct. 30, 2019; revision accepted Jun. 17, 2020.
For correspondence or reprints contact: Sally F. Barrington, St. Thomas Hospital, Westminster Bridge Rd., London SE1 7EH, U.K.
E-mail: sally.barrington@kcl.ac.uk
Published online Jul. 17, 2020.

Quantitative measurements can be affected by different methods of patient preparation, image acquisition, and reconstruction (*21*). Significant efforts have been made to standardize ^{18}F-FDG scanning in clinical trials, including initiatives by the European Association for Nuclear Medicine Research Limited (EARL) (*22*) and the Society of Nuclear Medicine and Molecular Imaging (*23*). However, there still exist differences in clinical practice and clinical trials that affect quantitative estimates such as MTV. Despite these methodologic 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 cutoff that separates patients with high-risk disease from patients with low-risk disease are crucially dependent on the segmentation method, the patient population characteristics, and the efficacy of treatment (*26*). Measurement of MTV can be cumbersome using current software approaches (*11*), and there is no agreed-upon consensus about the best method. These issues have 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, in clinical practice (*20*). Given that all methods appear to predict prognosis with equal effectiveness (*9,11*), efforts should focus on developing a quick and easy method that has high success rates for outlining visible tumor, 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 evaluate the best method using an automated tool to measure MTV in DLBCL, assessed by the success of segmentation of visible tumor; to determine whether the success of the measurement method is influenced by uptake time and compliance or noncompliance with standardized recommendations for ^{18}F-FDG scanning (*21*) and the presence or absence of progression or death at 2 y; and to assess the differences in MTV and discriminatory power obtained by different segmentation methods.

MATERIALS AND METHODS

PET/CT scans were selected from patients with newly diagnosed DLBCL scanned in research studies in The Netherlands and the United Kingdom. The scans are part of the comprehensive Positron Emission Tomography ReAnalysis (PETRA) database to validate interim ^{18}F-FDG PET as a biomarker of response for non-Hodgkin lymphoma (https://petralymphoma.org/). The studies had approval by institutional review boards or ethics committees. The scans were chosen to provide a balance between, first, patients scanned using 60 min of uptake (Netherlands scans) and patients scanned using 90 min of uptake (U.K. scans); second, patients who were scanned using reconstruction parameters compliant with standardized recommendations and patients who were not (*21*); and third, patients who had died or experienced disease progression at 2 y and patients who had not.

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. Physiologic uptake can be removed and lesions added, if required, using a single click on maximum-intensity-projection and volume images. Two independent readers, without knowledge of the patient outcome, performed measurements and rated the success or failure of methods and workflows to automatically delineate visible tumor. PET and CT datasets were displayed alongside each another, with an option to fuse the datasets if necessary. The consensus ratings of the 2 readers were used in analyses.

The following segmentation methods were applied: an SUV of 2.5 (SUV2.5), an SUV of 4.0 (SUV4.0), adaptive thresholding using 50% of peak voxel value adapted for local background (A50P) (*28*), 41% of SUV_{max} (41%), a majority vote segmenting voxels detected by at least 2 methods (MV2), and a majority vote segmenting voxels detected by at least 3 methods (MV3) (*29*). Majority-vote approaches were included because 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 on whether it succeeded or failed in the task of automatic tumor delineation or required some additional but limited user interaction to edit the MTV (Table 1). Use of more than 2 additional manual editing steps was considered not feasible for clinical practice and was rated as a failure of the method.

In statistical analyses, a sample size of 140 allowed for 70 scans in each of the 3 subgroups (uptake time, EARL compliance, and progression or death) or 35 scans 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; power, 0.80).

Descriptive statistical tests were performed for all segmentation methods. Differences in success rates among the 6 segmentation methods were assessed using χ^2 tests. The influence of uptake time, reconstruction method, and progression on success and failure rates was also assessed by χ^2 tests. MTVs were analyzed using raw and natural logarithmic transformed data because of their nonnormal distribution. To assess agreement in MTVs among segmentation methods, Pearson correlation coefficients were determined. The influence of uptake time, reconstruction method, and progression on MTVs obtained by the different methods was evaluated by *t* tests. The discriminative power regarding progression and nonprogression of the segmentation methods was assessed by comparing the mean volumes using *t* tests and receiver-operating-characteristic curves. All analyses were performed with IBM SPSS, version 22.

TABLE 1
Definition of Success, Failure, and Editing-Required Ratings

Rating	Findings
Success	No or minimal interaction was needed by observer; for example, removing brain or bladder uptake or adding single region with single mouse click
Failure	Automatic segmentation missed more than half the visible tumor on scan, or tumor flooded into (also included) uptake in adjacent physiologic structures that required complex slice-by-slice editing of 3 or more regions
Editing required	One or 2 additional manual steps were required; for example, adding missed regions or deleting slice-by-slice up to 2 regions of physiologic uptake adjacent to tumor using eraser tool (typically bladder or kidneys)

RESULTS

One hundred forty baseline PET/CT scans were assessed. Two patients without ^{18}F-FDG–avid disease were excluded, leaving 138 scans. Agreement between readers was excellent, at 91% for the 41% method and over 95% for all other methods.

Performance of Different Segmentation Methods

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

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

When different uptake times were compared, the 41% ($P < 0.05$) and MV3 ($P < 0.05$) methods were more likely to fail when scans were acquired at 90 min (Supplemental Table 1A; supplemental materials are available at http://jnm.snmjournals.org). 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 between scans that complied with EARL recommendations and scans that did not (Supplemental Table 1B). All methods performed as well in patients who died or progressed as in patients who did not (Supplemental Table 1C).

Comparison of MTVs Among Segmentation Methods

MTVs were log-transformed to obtain a normal distribution and were greater using SUV4.0 and MV2 than using the 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 the 41% and A50P methods showed only a moderate correlation (Fig. 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 the 41% method ($r = 0.94$).

For all segmentation methods, the means of MTVs did not significantly differ between Netherlands and U.K. patients (i.e., 60 vs. 90 min of uptake) (Supplemental Table 2A) or between patients who were scanned using EARL recommendations and patients who were not (Supplemental Table 2B). For all methods, MTVs were significantly higher in patients who progressed or died than in patients who did not (Supplemental Table 2C). The discriminative power of all methods was similar (Fig. 4).

TABLE 2
Pairwise Tests of Segmentation Methods Using SUV4.0 as Reference

Segmentation method	Success	Failure	Editing required
SUV4.0	105	6	27
MV2	102	10	26
MV3	90	40*	8
41%	82*	45*	11
A50P	75*	57*	6
SUV2.5	51*	57*	30

*$P < 0.005$, compared with SUV4.0.

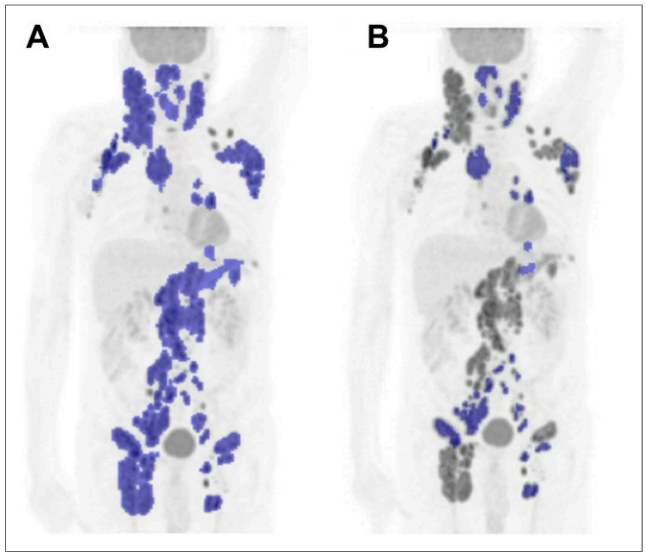

FIGURE 1. Case 1 was rated as successful using SUV4.0 (A) but as failure using the 41% method (B) because it missed more than half the visible tumor.

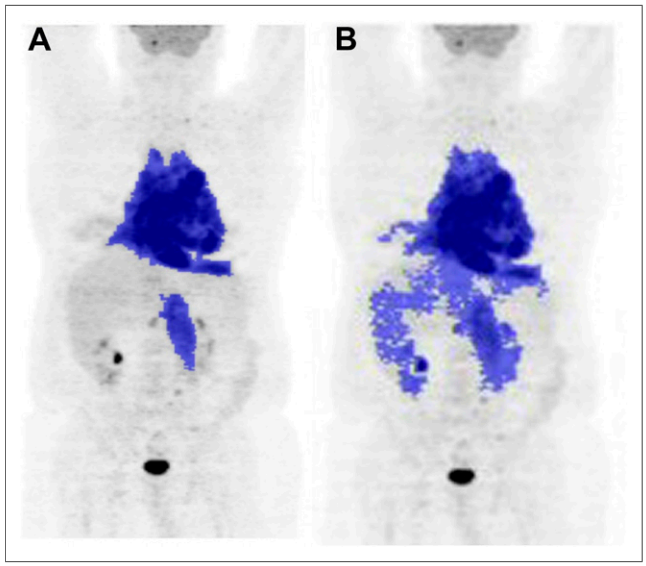

FIGURE 2. Case 2 was rated as successful using SUV4.0 (A) but as failure using SUV2.5 (B) because of inclusion of physiologic uptake, requiring complex editing.

TABLE 3
Untransformed and Log-Transformed MTVs (cm^3) by Segmentation Method

Segmentation method	Median volume	Interquartile range	Log-volume mean	Log-volume SD
SUV4.0	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

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 outcome in DLBCL irrespective of the delineation method, but the absolute values for MTV and the optimal cutoffs to divide good-prognosis groups from poor-prognosis groups differ (*11,26*). Measurement of MTV in patients with DLBCL takes around 3–6 min per scan, depending on the method, but complex cases can take 10–20 min (*11*). There is presently no agreement 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 for which true volumes are known (*16*), to try to overcome challenges in segmentation of PET images that have limited spatial resolution, causing partial-volume effects, when developing contouring algorithms (*31*). However, phantoms are not representative of the clinical situation, which includes variations in contrast, heterogeneity, and the shapes and sizes of lesions and patients. Moreover, recent studies suggest that the actual MTVs resulting from using different methods do not affect prognostic performance (*10*) and, thus, 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 tumors satisfactorily in patients with DLBCL according to the opinion of experienced observers—a method that represents how these tumors would be assessed in everyday practice (*32*). We devised a method to rate the success or failure of the Accurate tool a priori. To our knowledge, this is the first report to evaluate the success of an automated method in this way. Furthermore, we considered whether the choice of the best method was influenced by scanning conditions (i.e., uptake time and compliance or noncompliance with standardized recommendations (*21*)) and whether patients experienced later progression, or did not, using a case-control design.

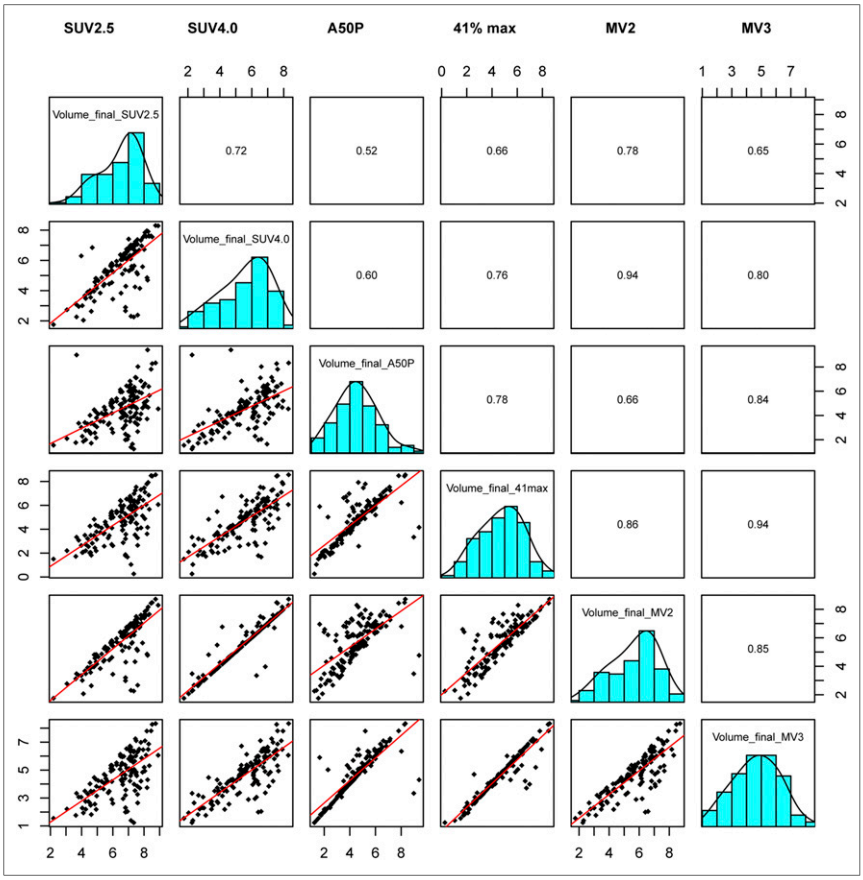

FIGURE 3. Distributions, scatterplots, and correlations of segmentation methods.

There were significant differences in performance for automated measurement of MTV with different segmentation methods. The best method for successful segmentation was SUV4.0; however, MV2 also performed well. A majority-vote method was included because 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 a majority vote and the Simultaneous Truth and Performance Level Estimation (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

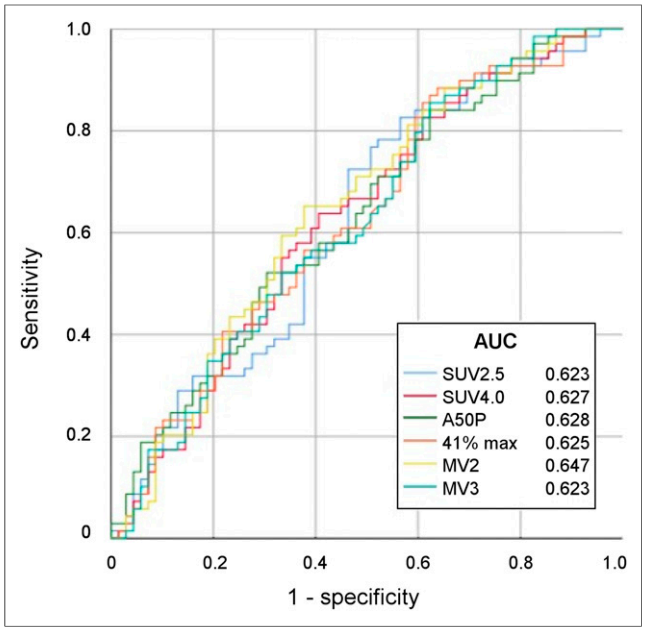

FIGURE 4. Receiver-operating-characteristic curves for MTVs using different methods.

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 any uptake present in the brain, urinary tract, and heart. In the remainder, an automated process was not completely reliable, generally because lesions were adjacent to areas with high physiologic uptake requiring user interaction. For some cases, the volume was altered substantially during editing, but for most, only 1 or 2 additional steps were needed to obtain a reasonable estimate of MTV. The other segmentation methods did not perform as well. Failure of SUV2.5 was mostly due to the complex editing needed to remove physiologic uptake, for a reasonable approximation of tumor burden. For SUV4.0, this situation was encountered in only 1 patient, because spillover of counts into other tissues was less common and, when it did occur, the overlap between tumor and physiologic uptake was less extensive and required 1 or 2 editing steps. Failure of the 41% and A50P methods was usually due to underestimation of tumor because of heterogeneity, with more than 50% of the tumor having uptake less than the chosen threshold. The 41% method performed slightly worse in patients scanned at 90 min, probably because uptake in tumor rises over time and, thus, for heterogeneous tumors with areas of high uptake, fewer voxels would be included in 41% of the SUV_{max}, 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 well methods performed.

The absolute values for MTV varied among methods, as previously reported (*20,26,33*). Across the whole range of volumes, a positive bias was seen for SUV2.5 and a negative bias for the other methods, in comparison with SUV4.0 and MV2, which performed similarly to each other. MV2 selects voxels included in at least 2 segmentation methods (SUV2.5, SUV4.0, A50, or 41%) and commonly included voxels delineated using SUV4.0 and the next method that came closest to delineating a similar volume, usually SUV2.5. MV3 selects voxels included in at least 3 of the segmentation methods, and it segmented volumes that were similar to those of the 41% method, which was most likely to delineate more of the same voxels as 2 or more of the other methods. MV2 performed similarly to SUV4.0 but required delineation using more than 1 segmentation method. This process is fully automated within Accurate but could be less easy to implement across software platforms than a single SUV4.0 threshold method. Clinically available software currently can measure MTV using SUV4.0, although Accurate has additional features to enable the user to quickly review the maximum-intensity-projection image and add missed lesions or remove physiologic uptake with a single click, speeding the segmentation process. Correlation among 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*). Because we used a case-control design, with an oversampling of patients with progression, 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 receiver-operating-characteristic curves demonstrated lower discriminative power than was found by Ilyas et al. (*11*), possibly because the latter used cases from a single institution and performed manual editing in most. Nonetheless, the fact that all methods predicted prognosis equally well suggests that selection of the best method can be based on success rate, ease of use, and time or user interaction to obtain total tumor burden.

Limitations are that the research software developed by our group is not yet widely available but the software has been designed as a tool that could be implemented across software platforms after discussions with manufacturers. Only classic segmentation methods published in lymphoma datasets were assessed, whereas more sophisticated methods may give more reliable estimates 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 that 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. MTV was assessed by 2 experienced observers with high concordance, mirroring the high reproducibility reported by others (*11,25*). However, the observers were aware of which method was being applied; masking was not possible because the delineation method was often obvious. We have assumed that the cases evaluated are representative, but the case-control design meant that 50% of patients progressed; the rate of progression would be lower in the clinic. This overrepresentation is likely to accentuate the challenges of measuring MTV, as patients who progress later would be expected to have a 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 larger, unselected, multicenter datasets representative of all patients with DLBCL. The results are also likely to be applicable to other lymphoma subtypes. Further work will explore the association of MTV with clinical outcome in a larger database within the PETRA consortium using the best methods evaluated in this study.

DISCLOSURE

Sally Barrington is supported by the National Institute for Health Research (NIHR) (RP-2-16-07-001). King's College London and the UCL Comprehensive Cancer Imaging Center are 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 authors 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). No other potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We thank the patients and collaborating investigators who kindly supplied their data.

KEY POINTS

QUESTION: What is the best automated workflow to measure MTV in DLBCL; is the choice of best workflow influenced by uptake time, compliance or noncompliance with standardized recommendations for ^{18}F-FDG scanning, and subsequent progression; and do segmentation methods give different MTVs or discriminate between patient outcomes equally?

PERTINENT FINDINGS: The best automated workflow (judged by segmentation of visible tumor by experienced observers) was SUV4.0, with significantly higher success and lower failure rates than other methods (SUV2.5, A50P, 41%, and MV3) except MV2, which also performed well. 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.

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 ^{18}FDG-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 ^{18}FDG-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 [^{18}F]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. Tridimensional 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. Schöder 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 [abstract]. *J Nucl Med.* 2018;59(suppl 1):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;22:1102–1110.
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.