Journal of Nuclear Medicine, published on June 12, 2020 as doi:10.2967/jnumed.120.242412

Deep learning FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-cell lymphoma

Nicolò Capobianco,[1] Michel Meignan,[2] Anne-Ségolène Cottereau,[3] Laetitia Vercellino,[4] Ludovic Sibille,[5] Bruce Spottiswoode,[5] Sven Zuehlsdorff,[5] Olivier Casasnovas,[6] Catherine Thieblemont,[7] and Irène Buvat[8]

[1]Siemens Healthcare GmbH, Erlangen, Germany; [2]Lysa Imaging, Henri Mondor University Hospitals, APHP, University Paris East, Créteil, France; [3]Department of Nuclear Medicine, Cochin Hospital, AP-HP, Paris, France; [4]Department of Nuclear Medicine, Saint-Louis Hospital, AP-HP, Paris, France; [5]Siemens Medical Solutions USA, Inc., Knoxville, TN, USA; [6]Department of Hematology, University Hospital of Dijon, Dijon, France; [7]Department of Hematology, Saint Louis Hospital, APHP, Paris, France; [8]Laboratoire d'Imagerie Translationnelle en Oncologie, Inserm, Institut Curie, Université Paris Saclay, France.

Corresponding author:

Nicolò Capobianco
Siemens Healthcare GmbH
Hartmannstr. 16
91052, Erlangen, Germany
E-mail: nicolo.capobianco@siemens-healthineers.com.
Mobile: +49 (1522) 6388791
Fax: +49 9131 84 4189

Word count**:** 5398

Short running title: TMTV estimation using deep learning

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska–Curie grant agreement (no. 764458).

ABSTRACT

Total metabolic tumor volume (TMTV), calculated from ^{18}F-labeled fluoro-2-deoxyglucose (^{18}F-FDG) positron-emission tomography–computed tomography (PET/CT) baseline studies, is a prognostic factor in diffuse large B-cell lymphoma (DLBCL) whose measurement requires the segmentation of all malignant foci throughout the body. No consensus currently exists regarding the most accurate approach for such segmentation. Further, all methods still require extensive manual input from an experienced reader. We examined whether an artificial intelligence (AI)-based method could estimate TMTV with a comparable prognostic value to TMTV measured by experts.

Methods: Baseline ^{18}F-FDG PET/CT scans of 301 DLBCL patients from the REMARC trial (NCT01122472) were retrospectively analyzed. An automated whole-body high-uptake segmentation algorithm identified all three-dimensional regions of interest (ROI) with increased tracer uptake. The resulting ROIs were processed using a convolutional neural network trained on an independent cohort and classified as nonsuspicious or suspicious uptake. The AI-based TMTV was estimated as the sum of the volumes of ROIs classified as suspicious uptake. The reference TMTV was measured by two experienced readers using independent semiautomatic software. The AI-based TMTV was compared to the reference TMTV in terms of prognostic value for progression-free survival (PFS) and overall survival (OS).

Results: The AI-based TMTV was significantly correlated with the reference TMTV (ρ=0.76; $p<0.001$). Using the AI-based approach, an average of 24 regions per subject with increased tracer uptake were identified, and an average of 20 regions per subject were correctly identified as nonsuspicious or suspicious, yielding 85% classification accuracy, 80% sensitivity, 88% specificity, compared to the reference TMTV region. Both TMTV results were predictive of PFS

(hazard ratio: 2.4 and 2.6 for AI-based and reference TMTVs, respectively; $p<0.001$) and OS (hazard ratio: 2.8 and 3.7 for AI-based and reference TMTVs, respectively; $p<0.001$).

Conclusion: TMTV estimated fully automatically using an AI-based approach was consistent with that obtained by experts and displayed a significant prognostic value for PFS and OS in DLBCL patients. Classification of high uptake regions using deep learning for rapidly discarding physiological uptake may considerably simplify TMTV estimation, reduce observer variability and facilitate the use of TMTV as a predictive factor in DLBCL patients.

Keywords: Metabolic Tumor Volume, Lymphoma, Deep Learning, FDG, PET/CT

INTRODUCTION

Total metabolic tumor volume (TMTV) derived from ^{18}F-labeled fluoro-2-deoxyglucose (^{18}F-FDG) positron-emission tomography–computed tomography (PET/CT) baseline studies is a promising prognostic factor in diffuse large B-cell lymphoma (DLBCL) (*1*,*2*) and other types of lymphoma (*3*-*5*). DLBCL is the most frequent non-Hodgkin's lymphoma present in about 30% to 40% of non-Hodgkin's lymphoma cases worldwide. Although the prognosis of DLBCL can be improved with immunochemotherapy, more than 30% of patients are refractory or relapse following first-line treatment, with a poor outcome (*6*,*7*). Therefore, there is a need to identify high-risk patients who could benefit from intensive or novel therapies early. Unfortunately, the role of current prognostic factors such as the International Prognostic Index (IPI) (*8*), Revised IPI (*9*), and National Comprehensive Cancer Network IPI (*10*), based on tumor burden surrogates is limited. Thus, baseline TMTV, which estimates the total metabolic tumor burden at diagnosis, has been proposed as an alternative prognostic tool for early risk stratification.

To date, TMTV is not yet routinely used in clinical lymphoma patient management in part because of a lack of consensus throughout the literature. Several methods have been proposed to calculate TMTV (*11*-*13*), and the cutoff values reported to detect high-risk patients differed among methods and studies. However, recent studies have suggested that, despite these differences, the majority of the methods yielded similar accuracy in predicting patient prognosis when applied in similar patient groups (*11*,*12*), emphasizing the strong prognostic power of baseline TMTV.

Regardless of the criteria used for delineating tumor regions, all methods for deriving TMTV require extensive and time-consuming manual input from an experienced reader. The reader either manually segments the tumor regions or, more commonly, uses an automated

method to detect all regions with increased uptake and then manually eliminates the regions of physiological uptake and adds in undetected tumor regions (*13*). Recently, a machine learning algorithm using a convolutional neural network (CNN) was trained to differentiate physiological from non-physiological uptake regions in whole-body ^{18}F-FDG PET scans acquired from an unselected population of more than 600 patients, including half who were lymphoma patients with different subtypes of diseases (*14*,*15*). This CNN achieved a high degree of accuracy in characterizing increased tracer uptake in the whole body as physiological or non-physiological. Such automated identification of non-physiological regions would facilitate TMTV measurement and clinical adoption. This study therefore sought to assess the ability of this CNN to identify regions from which TMTV could be automatically calculated and to evaluate the ability of the resulting TMTV in predicting patient outcome among a large group of DLBCL patients included in an international phase III trial wherein TMTV has already been demonstrated to be a strong predictor of four-year progression-free survival (PFS) and overall survival (OS). To evaluate the CNN performance, regions with elevated tracer uptake automatically identified as physiological or suspicious are compared to regions attributed to suspicious uptake by an expert reader using a semi-automatic method.

MATERIALS AND METHODS

Patients

Patients from an ancillary study (*16*,*17*) of the REMARC trial (NCT01122472) were retrospectively analyzed. This trial is a phase III study that was designed to assess the efficacy of lenalidomide versus placebo in responding elderly DLBCL patients (60–80 years old) treated with the standard first-line rituximab, cyclophosphamide, doxorubicin hydrochloride (hydroxydaunorubicin), vincristine sulfate, and prednisone (R-CHOP) therapy approach (*18*).

The institutional review board approval and the informed consent of the REMARC trial included all the ancillary studies. The ancillary study was conducted by involving 301 patients who underwent baseline PET/CT scans before R-CHOP and showed that TMTV was a strong prognosticator of outcome in patients responding to first-line chemotherapy combined with monoclonal antibody treatment.

Image Acquisition and Analysis

All baseline ^{18}F-FDG PET/CT images from the ancillary study were collected in an anonymized Digital Imaging and Communications in Medicine (DICOM) format. Patients with PET or CT DICOM series with incomplete axial slices or irregular slice intervals were excluded. PET images were expressed in standardized uptake value (SUV) units, accounting for injected dose and patient body weight.

PET/CT images were analyzed using an investigational software prototype [PET Assisted Reporting System (PARS), Siemens Medical Solutions USA, Inc., Knoxville, TN, USA]. The prototype first automatically located a cylindrical reference region at the center of the proximal descending aorta by applying a landmarking algorithm to the CT image (*19*). This region was used to determine the mean blood pool uptake (SUV_{BP}) and standard deviation (std_{SUVBP}), following PET Response Criteria in Solid Tumors (PERCIST) recommendations (*20*). The three-dimensional regions of the PET image with increased tracer uptake were identified for each subject using an automated whole-body high-uptake segmentation algorithm (multi-foci segmentation, MFS) (*21*). In line with the PERCIST recommendations, only the regions with $SUV_{peak} > 2\ SUV_{BP} + 2\ std_{SUVBP}$ were included. Those regions were then further segmented according to 42% of the SUV_{max} threshold, and the ones with volumes below 2 mL were discarded. The resulting regions, called regions of interest $(ROI)_{PARS}$ thereafter, were then

automatically processed by a CNN. Details of the training and validation of this CNN were previously reported (*15*). The input of the CNN was the PET/CT data together with the set of ROI_{PARS}. For each ROI_{PARS}, the output of the CNN was the anatomical localization among a set of possible anatomical sites relevant for staging and whether the ROI_{PARS} uptake was physiological (e.g. due to unspecific bowel uptake, muscle activation, inflammation/infection, bone degeneration) or suspicious (e.g. lymphoma) uptake. The volumes of all ROI_{PARS} classified as suspicious uptake were then summed to obtain the $TMTV_{PARS}$.

The CNN was also employed in combination with other settings of the initial high-uptake ROI segmentation: 1) using an initial threshold of 2.5 SUV instead of the blood-pool based threshold, followed by thresholding with 41% of SUV_{max}, 2) including also ROIs with a volume between 0.1mL and 2mL.

The TMTV obtained by two experienced nuclear medicine physicians in the context of a previous study (*16*,*17*) was used as a reference ($TMTV_{REF}$). The $TMTV_{REF}$ was obtained using the semi-automatic version of the Beth Israel FIJI (ImageJ) software plugin (*22*), which was previously used to demonstrate the prognostic value of TMTV in various lymphoma subtypes (*5*,*23*). To calculate $TMTV_{REF}$, the physician combined automated and manual steps as follows. First, volumes of interest with high uptake in the PET images were segmented using an automated method, which applied in sequence an algorithm based on component trees and shape priors (*24*), a region growing, and a final region delineation using 41% of the region SUV_{max} threshold (*25*). Second, the resulting ROIs were manually reviewed by the reader, who selected only the regions corresponding to lymphoma (ROI_{REF}), adding an ROI_{REF} wherever a lymphoma lesion had been missed by the algorithm by drawing a prism around that lesion and applying a

41% SUV_{max} threshold. The volumes of all lymphoma ROI_{REF} were summed to obtain the reference $TMTV_{REF}$.

Statistical Analysis

To evaluate the performance of the CNN classification, for each patient, each ROI_{PARS}, having been labeled as presenting suspicious or physiological uptake by the CNN, was compared to all ROI_{REF} regions of that patient taken together. The ROI_{PARS} was considered to "match" the ROI_{REF} regions if at least 50% of its volume overlapped with one or several ROI_{REF}. ROI_{PARS} classified as suspicious and matching one or several ROI_{REF} were considered as true positives, ROI_{PARS} classified as physiological and matching one or several ROI_{REF} were considered as false negatives, ROI_{PARS} classified as physiological that did not match any ROI_{REF} were considered as true negatives, and ROI_{PARS} classified as suspicious that did not match any ROI_{REF} were considered as false positives. The sensitivity, specificity, and accuracy of the uptake classification were calculated. The performance of the CNN classification was also assessed in case a minimum overlap of 25% and 75% was required to consider a ROI_{PARS} as matching the ROI_{REF} regions.

To evaluate differences between $TMTV_{PARS}$ and $TMTV_{REF}$, Bland–Altman analysis was performed. Since the Shapiro–Wilk test revealed significant non-normal distribution of the differences between $TMTV_{PARS}$ and $TMTV_{REF}$ ($p < 0.001$), the median bias and limits of agreement at 2.5 and 97.5 percentiles were reported in the Bland–Altman plot. To assess the correlation between ranked TMTV values, Spearman's rank correlation coefficient was used. For each patient, the agreement between the patient set of ROI_{PARS} classified as suspicious and the patient set of ROI_{REF} was characterized using the Dice score, precision (the fraction of voxels in the set of ROI_{PARS} classified as suspicious that were also present in the set of ROI_{REF}), and recall

(the fraction of voxels in the set of ROI_{REF} that were also present in the set of ROI_{PARS} classified as suspicious).

Survival analysis was performed for both $TMTV_{PARS}$ and $TMTV_{REF}$ with respect to PFS and OS. Receiver operating characteristic (ROC) curves were used to determine TMTV cutoff thresholds to predict the occurrence of events for both PFS and OS, by maximizing Youden's index (sensitivity + specificity − 1). Survival functions were computed by Kaplan–Meier analyses and used to estimate survival time statistics (such as four-year PFS rate and four-year OS rate) for "low" and "high" TMTV groups. A log-rank test was employed to assess whether differences between Kaplan–Meier survival curves were significant. Univariate Cox regression was used to calculate hazard ratios between survival groups. Statistical significance was set at $p < 0.05$. Statistical analysis was performed using R version 3.6.1 with pROC version 1.15.3 (*26*).

RESULTS

In total, 280 patients from 124 centers were included in the analysis. Patient characteristics are reported in Table 1. All received first-line treatment with R-CHOP and were responders at the time of inclusion in the trial, 142 received lenalidomide regimen afterward as maintenance, and 138 received placebo. After a median follow-up of five years, 86 patients presented a PFS event and 51 patients had an OS event; the four-year survival rates were 69% for PFS and 83% for OS. The four-year survival rates were comparable to those of the entire trial.

PET/CT images of the 280 included patients were acquired using different scanner models from different vendors as summarized in Supplemental Table 1. The delay between injection and acquisition time was 71.7 ± 14.1 min (mean ± std). The mean SUV in the proximal

descending aorta cylindrical region was 1.6 ± 0.5 (mean ± std across subjects), resulting in a SUV_{peak} threshold of 3.6 ± 1.2 for detecting ROIs with increased tracer uptake.

Results below are described for the PERCIST-based setting of the initial high-uptake ROI segmentation, while changes observed with other settings are reported in Supplemental Tables 2-4.

Uptake Classification

In total, 6,737 ROI_{PARS} exhibiting increased uptake were obtained from the 280 subjects by the MFS algorithm using PARS. There were 7,996 ROI_{REF} in the 280 subjects. Descriptive statistics for the number of ROI_{PARS} and ROI_{REF} per subject are summarized in Supplemental Table 5. Among the 6,737 ROI_{PARS} with increased uptake, 2,831 ROI_{PARS} (42%) were classified as having suspicious uptake by the CNN.

When compared with the ROI_{REF} obtained by the experienced reader, the identification of the ROI_{PARS} with suspicious uptake by the CNN yielded 3,317 true negatives, 2,399 true positives, 589 false negatives, and 432 false positives. Corresponding sensitivity was 80%, specificity was 88%, and accuracy was 85%.

Additionally, the mean per-subject ROI_{PARS} classification accuracy was 87% [median: 89%, Inter Quartile Range (IQR): 81%–96%]. There were an average of 20 correctly classified ROI_{PARS} per subject (median: 17 ROI_{PARS}, IQR: 11–27 ROI_{PARS}) and an average of four incorrectly classified ROI_{PARS} per subject (median: 2 ROI_{PARS}, IQR: 1–5 ROI_{PARS}), which were regions classified as suspicious by the CNN that did not overlap with the set of ROI_{REF} or regions classified as physiological by the CNN but which overlapped with the set of ROI_{REF}. Two examples of uptake classification using PARS with corresponding ROI_{REF} are shown in Fig.

1. Results with a minimum overlap of 25% and 75% required to consider a ROI_{PARS} as matching the ROI_{REF} regions are reported in Supplemental Table 6.

Total Metabolic Tumor Volume

After discarding the ROI_{PARS} classified as physiological uptake by the CNN, a median $TMTV_{PARS}$ of 110 mL was obtained (IQR: 33–281 mL). The median $TMTV_{REF}$ was 240 mL (IQR: 80–529 mL) (Table 2).

There was a significant correlation between ranked TMTV estimates ($\rho = 0.76$; $p < 0.001$). The median Dice score across all patients between the patient set of ROI_{PARS} labeled as suspicious and the patient set of ROI_{REF} was 0.73 (IQR: 0.33–0.86), the median recall of the patient set of ROI_{PARS} labeled as suspicious with respect to the patient set of ROI_{REF} was 0.62 (IQR: 0.20–0.81), and the median precision was 0.96 (IQR: 0.86–0.99). The Bland–Altman plot comparing $TMTV_{PARS}$ and $TMTV_{REF}$ (Fig. 2) showed wide limits of agreement.

Survival Analysis

The area under the ROC curve for predicting PFS was 0.61 for $TMTV_{PARS}$ and 0.64 for $TMTV_{REF}$ (Fig. 3). The optimal cutoffs for predicting PFS were 110 mL for $TMTV_{PARS}$ and 242 mL for $TMTV_{REF}$. Kaplan–Meier survival curves are shown in Fig. 4. The four-year PFS rates were 81% and 58% for the low- and high-$TMTV_{PARS}$ groups and 83% and 55% for the low- and high-$TMTV_{REF}$ groups, respectively. The log-rank test indicated a significantly longer PFS time in the low-TMTV patient group for both TMTV estimation methods ($p < 0.001$ for $TMTV_{PARS}$ and $TMTV_{REF}$). Cox regression for PFS resulted in hazard ratios (high-TMTV group vs. low-TMTV group) of 2.4 [95% confidence interval (CI): 1.5–3.7; $p < 0.001$ for Wald test] for

$TMTV_{PARS}$ and 2.6 (95% CI: 1.6–4.1; $p < 0.001$) for $TMTV_{REF}$. The survival results are summarized in Table 3.

For OS, the area under the ROC curve was 0.64 for $TMTV_{PARS}$ and 0.66 for $TMTV_{REF}$. The optimal TMTV cutoffs for predicting OS were 148 mL for $TMTV_{PARS}$ and 223 mL for $TMTV_{REF}$. The four-year OS rates were 90% and 74% for the low- and high-$TMTV_{PARS}$ groups and 93% and 74% for the low- and high-$TMTV_{REF}$ groups, respectively. The log-rank test revealed a significantly higher OS time in the low-TMTV patient group for both TMTV estimation methods ($p < 0.001$ for $TMTV_{PARS}$ and $TMTV_{REF}$). Cox regression for OS resulted in hazard ratios (high-TMTV group vs. low-TMTV group) of 2.8 (95% CI: 1.6–5.1; $p < 0.001$) for $TMTV_{PARS}$ and 3.7 (95% CI: 1.9–7.2; $p < 0.001$) for $TMTV_{REF}$.

The sensitivity, specificity, negative predictive value, positive predictive value, and accuracy for predicting the occurrence of survival events, determined at the optimal TMTV cutoff point for each method, are reported in Supplemental Table 7, and were similar for both PFS and OS.

DISCUSSION

Our main result is that a fully automated method combining a region delineation method based on PERCIST recommendations and a CNN-based algorithm to distinguish between regions with elevated physiological uptake and non-physiological regions is able to generate, in a uniform population of DLBCL patients, TMTV values predictive of four-year PFS and OS with an accuracy comparable to that obtained when TMTV is calculated by manual selection of the tumor regions by medical experts. Although the CNN-based algorithm was trained using PET/CT images from only two scanner models from the same vendor, it showed high accuracy

in classifying regions with increased uptake in a group of patients from an international multicenter trial involving 124 centers, with PET/CT images obtained from different scanner models from different vendors with variable reconstruction settings. This underlines the robustness of the CNN despite different image quality. Moreover, this algorithm was not originally trained for TMTV computation and outcome prediction and was developed with data from patients with different lymphoma subtypes and lung cancer who underwent PET at baseline and for response assessment. However, we showed that the algorithm was successful in a group of patients with a homogenous lymphoma subtype scanned at baseline, enabling the identification of a TMTV cutoff separating high-risk and low-risk patients and predicting prognosis with accuracy comparable to that of the reference method. No subject was excluded due to failures of the initial high uptake ROI segmentation, which identified at least one high uptake region for all subjects. Furthermore, when employing different settings of the initial high uptake ROI segmentation using a lower threshold of 2.5 SUV in comparison with PERCIST recommended blood-pool based threshold, comparable results were obtained (Supplemental Tables 2 and 3), suggesting the robustness of the algorithm to the initial segmentation results. Additionally, the high-uptake ROI classification accuracy was not substantially impacted when a different level of overlap was required to consider a ROI as matching the reference TMTV and when ROIs with volumes less than 2mL were included in the analysis (Supplemental Tables 4 and 6).

The median TMTV and the resulting cutoff observed with PARS were lower than those observed with the reference method. This could be due to multiple factors, including 1) the higher initial SUV threshold used for PARS relative to the one used for the reference TMTV, 2) the manual addition of suspicious regions with low uptake in the reference TMTV, 3) regions

being classified as physiological in PARS but considered suspicious for the reference TMTV, and 4) differences in the contouring of suspicious regions between PARS and the reference TMTV. However, the ability of the TMTV estimates to be predictive of PFS and OS despite involving a TMTV range different from that of the reference TMTV is consistent with what has already been reported (*11,12*) when comparing different TMTV estimation methods. This confirms both the validity of the CNN method and the value of TMTV as a prognostic indicator.

Our study has limitations. Since there is currently no gold-standard method for TMTV calculation from ^{18}F-FDG PET/CT images (*27*), the reported figures of merit supporting the uptake classification performance and accuracy of our TMTV segmentation are limited to the comparison with the reference method considered in the study. Moreover, a uniform cohort of lymphoma patients was evaluated in the current study and results may differ for different lymphoma subtypes or different cancer types.

In the present work, we evaluated a fully automated application of PARS. However, PARS was initially intended to be used in a supervised manner, giving the ability to the reader to correct for potentially misclassified regions where appropriate. In particular, PET/CT image quality pitfalls such as misalignment due to motion or image artifacts may influence the classification output of the CNN algorithm, and the results should be validated by an expert. This is especially true when the labeling results are used to derive a prognostic index such as TMTV that can be used to stratify the risk and to guide personalized therapy. Nevertheless, this approach could be employed by expert readers to efficiently estimate TMTV, as the deep-learning based method is able to automatically identify several relevant suspicious uptake sites and automatically discard physiological uptake sites, with the expert only having to correct the potential improper classification of a limited number of regions per subject, requiring limited

user interaction and potentially improving inter-reader variability. This approach may introduce bias in the TMTV estimation process by relying on pre-generated results. However, this risk should be marginal especially when a careful revision of the results is performed by an experienced reader.

To our knowledge, this is the first study showing that an AI method can generate a TMTV value prognostic of outcome in a large series of patients with DLBCL, with results comparable to other currently employed methodologies. Other machine learning–based approaches for TMTV estimates in lymphoma patients, including some involving CNN, are being developed and evaluated (*28*). The automated method for TMTV segmentation assessed in the present study combined a region-delineation method based on PERCIST recommendations and a deep learning–based classification scheme for rapidly discarding physiological uptake. Further efforts toward developing a stricter definition of TMTV, standardizing volume-segmentation methods, and establishing guidelines for the inclusion of tumor-bearing anatomical regions are ongoing, and these will constitute a prerequisite for the optimization of a complete automated method (*13*).

CONCLUSION

We showed that Total Metabolic Tumor Volume can be estimated fully automatically using a deep learning approach. The resulting TMTV was consistent with that obtained by independent experts and showed significant prognostic value for PFS and OS in a large cohort of DLBCL subjects.

DISCLOSURE

N.C. is a full-time employee at Siemens Healthcare GmbH. L.S., B.S., and S.Z. are full-time employees at Siemens Medical Solutions USA, Inc. All the other authors declare to have no conflict of interest.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska–Curie grant agreement (no. 764458).

KEY POINTS

Question: Can deep learning be used to obtain an automated estimation of Total Metabolic Tumor Volume in baseline ^{18}F-FDG PET/CT for risk stratification in DLBCL patients?

Pertinent findings: In a cohort of 280 DLBCL patients from the REMARC trial, a deep learning algorithm could classify volumes of interest with elevated uptake in ^{18}F-FDG PET/CT as physiological or suspicious in good agreement with expert human reader assessment. By aggregating the volumes of interest classified as suspicious uptake by the deep learning algorithm, the automated Total Metabolic Tumor Volume estimates were significant for PFS and OS prediction.

Implications for patient care: Total Metabolic Tumor Volume estimated with an automated method using deep learning may contribute to reproducible and accurate identification of high risk patients with DLBCL.

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. Song M-K, Chung J-S, Shin H-J, et al. Clinical significance of metabolic tumor volume by PET/CT in stages II and III of diffuse large B cell lymphoma without extranodal site involvement. *Ann Hematol*. 2012;91:697–703.

3. Kanoun S, Rossi C, Berriolo-Riedinger A, et al. Baseline metabolic tumour volume is an independent prognostic factor in Hodgkin lymphoma. *Eur J Nucl Med Mol Imaging*. 2014;41:1735–1743.

4. Cottereau AS, Becker S, Broussais F, et al. Prognostic value of baseline total metabolic tumor volume (TMTV0) measured on FDG-PET/CT in patients with peripheral T-cell lymphoma (PTCL)†. *Ann Oncol*. 2016;27:719-724.

5. 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.

6. Gisselbrecht C, Glass B, Mounier N, et al. Salvage regimens with autologous transplantation for relapsed large B-cell lymphoma in the rituximab era. *J Clin Oncol*. 2010;28:4184-4190.

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

8. The international non-Hodgkin's lymphoma prognostic factors project. A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med*. 1993;329:987-994.

9. Sehn LH, Berry B, Chhanabhai M, et al. The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. *Blood*. 2007;109:1857–1861.

10. Zhou Z, Sehn LH, Rademaker AW, et al. An enhanced International Prognostic Index (NCCN-IPI) for patients with diffuse large B-cell lymphoma treated in the rituximab era. *Blood*. 2014;123:837–842.

11. Cottereau A-S, 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.

12. 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.

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

14. Sibille L, Avramovic N, Spottiswoode B, Schaefers M, Zuehlsdorff S, Declerck J. PET uptake classification in lymphoma and lung cancer using deep learning [abstract]. *J Nucl Med*. 2018;59(suppl 1):325–325.

15. Sibille L, Seifert R, Avramovic N, et al. 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*. 2020; 294:445-452.

16. Cottereau A, Vercellino L, Casasnovas O, et al. High total metabolic tumor volume at baseline allows to discriminate for survival patients in response after R-CHOP: an ancillary analysis of the REMARC study [abstract]. *Hematol Oncol*. 2019;37(suppl 2):49-50.

17. Vercellino L, Cottereau A-S, Casasnovas R-O, et al. High total metabolic tumor volume at baseline allows discrimination of survival even in patients aged 60 to 80 years responding to R-CHOP. *Blood.* January 24, 2020 [Epub ahead of print].

18. Thieblemont C, Tilly H, Gomes da Silva M, et al. Lenalidomide maintenance compared with placebo in responding elderly patients with diffuse large B-cell lymphoma treated with first-line rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone. *J Clin Oncol*. 2017;35:2473-2481.

19. Tao Y, Peng Z, Krishnan A, Zhou XS. Robust learning-based parsing and annotation of medical radiographs. *IEEE Trans Med Imaging*. 2011;30:338-350.

20. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med*. 2009;50:122S-50S.

21. Brito A, Santos A, Mosci C, et al. Comparison of manual versus semi-automatic quantification of skeletal tumor burden on 18F-Fluoride PET/CT [abstract]. *J Nucl Med*. 2017;58(suppl 1):766–766.

22. 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.

23. Cottereau A-S, 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.

24. Grossiord E, Talbot H, Passat N, Meignan M, Tervé P, Najman L. Hierarchies and shape-space for PET image segmentation. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. New York, NY: IEEE; 2015:1118–1121.

25. 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.

26. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.

27. Cottereau, A., Buvat, I., Kanoun, S. et al. Is there an optimal method for measuring baseline metabolic tumor volume in diffuse large B cell lymphoma? *Eur J Nucl Med Mol Imaging*. 2018;45:1463-1464.

28. Jemaa S, Fredrickson J, Coimbra A, et al. A fully automated measurement of total metabolic tumor burden in diffuse large B-cell lymphoma and follicular lymphoma [abstract]. *Blood*. 2019;134(suppl 1):4666-4666.

FIGURES

Figure 1.

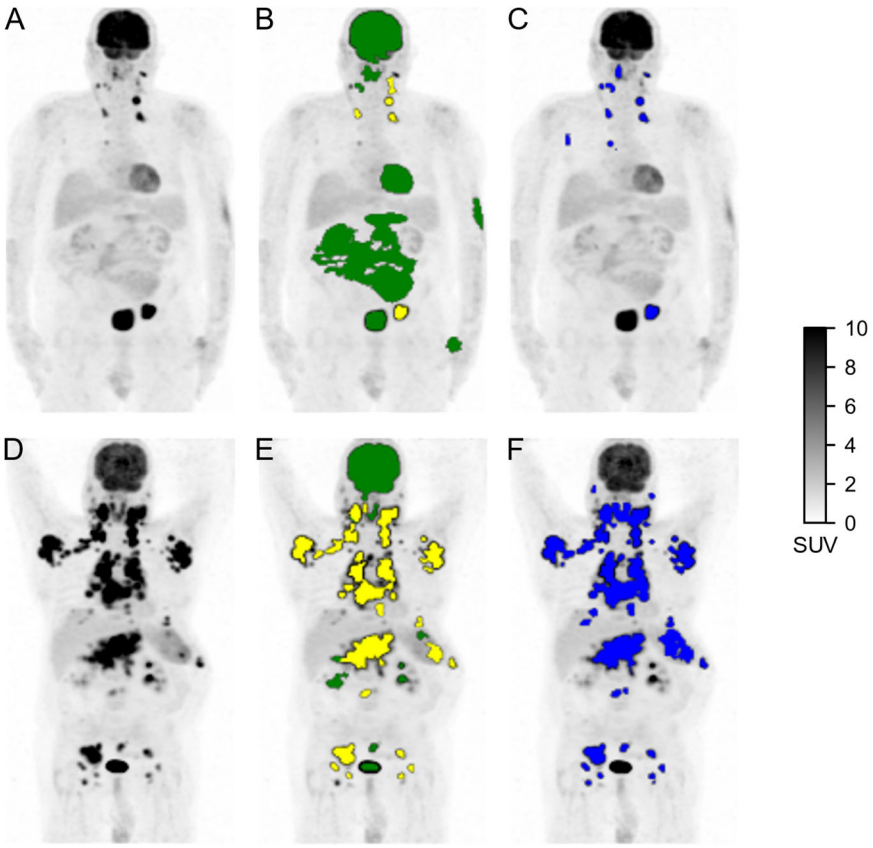
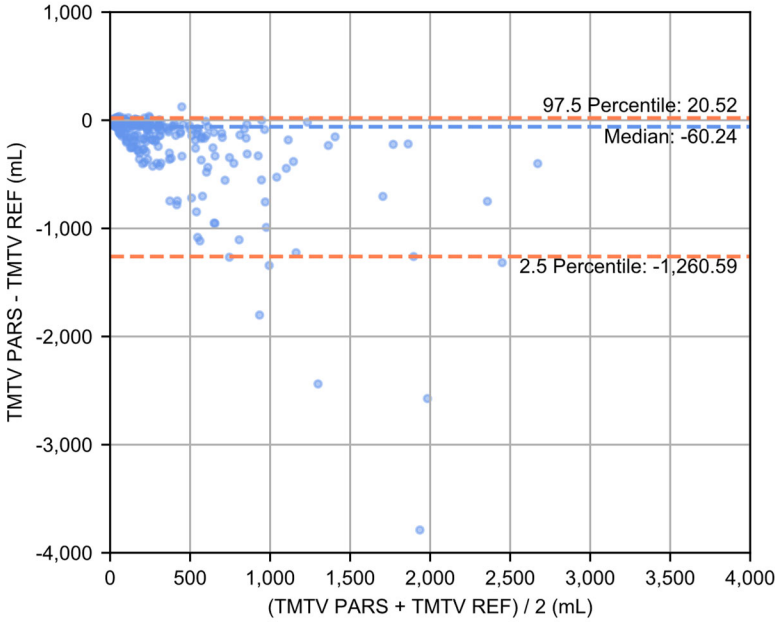


Figure 1. Detection and classification of high ^{18}F-FDG uptake regions as physiological or suspicious.

(A, D) Maximum-intensity projection (MIP) PET images of two subjects with low TMTV (A) and high TMTV (D). (B, E) ROI_{PARS} obtained automatically using the PARS software prototype. ROI_{PARS} detected by the MFS algorithm are overlaid on to the PET MIP. ROI_{PARS} classified by the deep learning algorithm as physiological are shown in green, and ROI_{PARS} classified as suspicious are shown in yellow. (C, F) ROI_{REF} regions obtained by an experienced nuclear medicine physician using a semiautomatic software.

Figure 2.



.

Figure 2. Bland–Altman plot comparing fully automated and reference TMTV estimations.

Bland–Altman plot comparing the TMTV obtained using the software prototype PARS ($TMTV_{PARS}$) and the reference TMTV ($TMTV_{REF}$) obtained by a nuclear medicine physician using a semiautomatic software.

Figure 3.

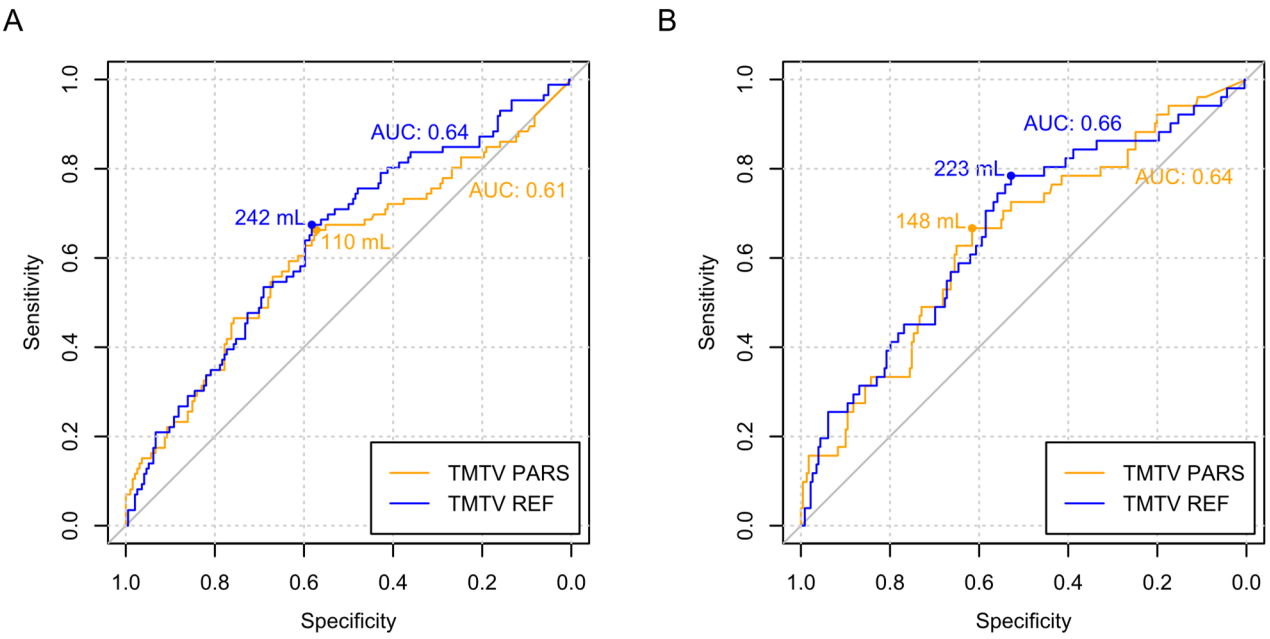


Figure 3. ROC curves for determining the occurrence of PFS or OS events using a TMTV threshold.

ROC curves for $TMTV_{PARS}$ and $TMTV_{REF}$ for (A) PFS and (B) OS. Areas under the ROC curves (AUC) and optimal TMTV cutoff thresholds are reported.

Figure 4.

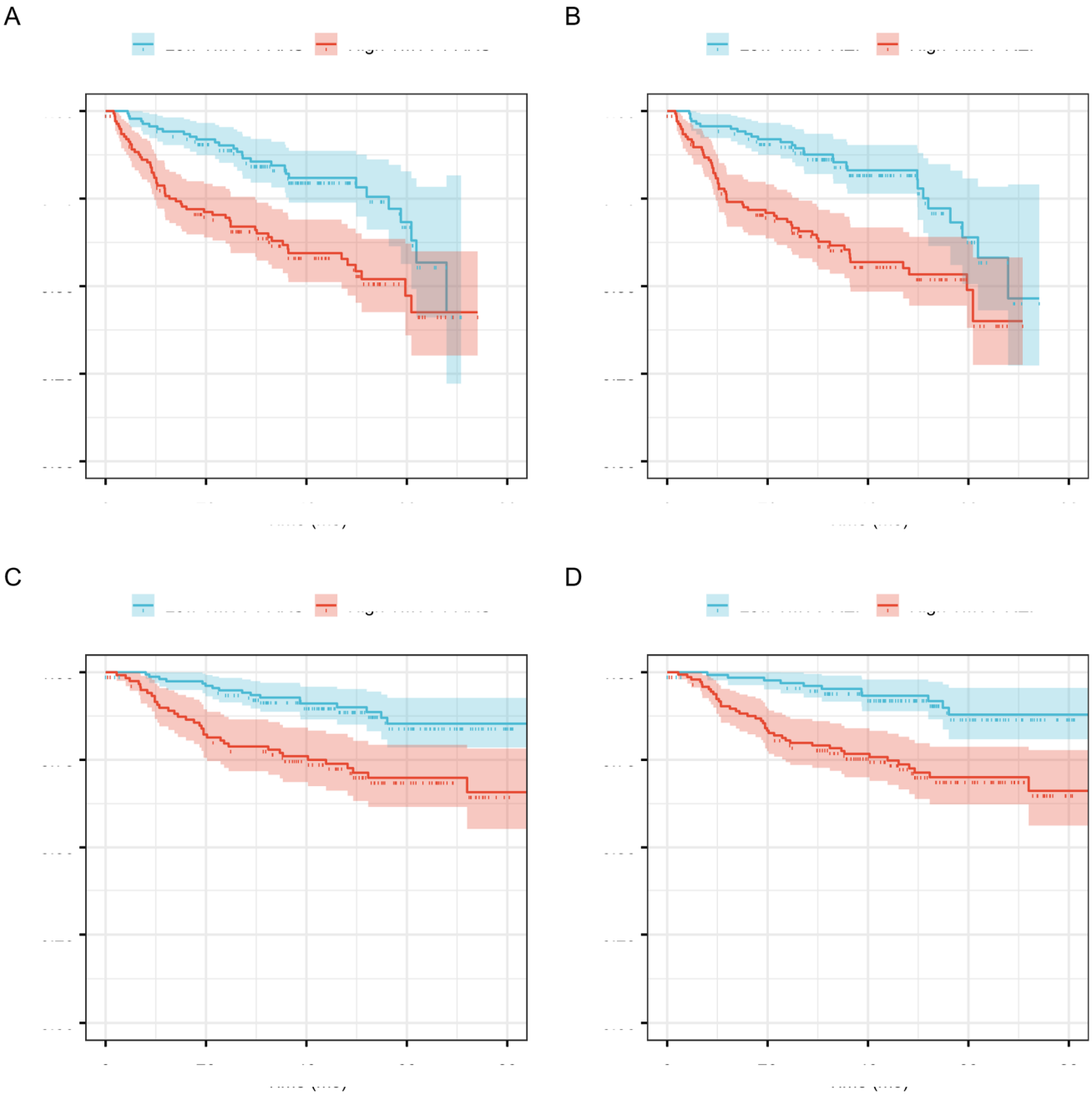


Figure 4. Survival curves for the low- and high-TMTV groups for fully automated and reference TMTV estimations.

Kaplan–Meier survival curves for PFS (A: $TMTV_{PARS}$, B: $TMTV_{REF}$) and OS (C: $TMTV_{PARS}$, D: $TMTV_{REF}$).

TABLES

Table 1. Patient characteristics

Patient characteristics	**Total number = 280 (%)**
Sex	
Female	119 (42.5)
Male	161 (57.5)
Age (median, ranges) years	68 (58-80)
Ann Arbor Stage	
I	1 (0.4)
II	25 (8.9)
III	57 (20.4)
IV	197 (70.4)
Performance status (ECOG)	
0	113 (40.4)
1	119 (42.5)
2	39 (13.9)
3	2 (0.7)
4	2 (0.7)
Missing	5 (1.8)
IPI	
1	6 (2.1)
2	73 (26.1)
3	97 (34.6)
4	81 (28.9)
5	19 (6.8)
Missing	4 (1.4)
Elevated LDH (>Upper limit of normal*)	
No	111 (39.6)
Yes	165 (58.9)
Missing	4 (1.4)

*LDH upper limit set specifically for each laboratory

Table 2: Descriptive statistics of TMTV obtained using the software prototype PARS ($TMTV_{PARS}$) and the reference method for the 280 subjects included in the study

TMTV Estimation	Mean	STD	Min	Q1 (25%)	Median	Q3 (75%)	Max
$TMTV_{PARS}$ (mL)	235.2	347.6	0.0	32.9	110.2	280.8	2471.9
$TMTV_{REF}$ (mL)	433.7	571.3	2.27	80.0	240.0	529.3	3832.7

Table 3: Results associated with ROC analysis of TMTV, Kaplan–Meier estimation of four-year survival rates, Cox regression hazard ratio, and Wald test p-values for PFS and OS for the 280 subjects included in the study.

TMTV estimation	AUC*	Cutoff (mL)	Hazard ratio (95% CI)	High TMTV 4-y Survival	Low TMTV 4-y Survival	*P*
Progression-free Survival						
$TMTV_{PARS}$	0.61	110	2.4 (1.5–3.7)	58%	81%	0.00016
$TMTV_{REF}$	0.64	242	2.6 (1.6–4.1)	55%	83%	0.00004
Overall survival						
$TMTV_{PARS}$	0.64	148	2.8 (1.6–5.1)	74%	90%	0.00044
$TMTV_{REF}$	0.66	223	3.7 (1.9–7.2)	74%	93%	0.00012

*Area under the ROC curve

Supplemental Data

Supplemental Table 1: PET/CT scan characteristics.

PET/CT study characteristics	**Total number = 280**
Injected dose (MBq)	309 ± 87 (mean ± std)
Post injection scan delay (min)	71.7 ± 14.1 (mean ± std)
PET slice thickness (mm)	Median: 3.7; min-max: 2.0–5.0
PET pixel spacing (mm)	Median: 4.0; min-max: 2.3–5.5
CT slice thickness (mm)	Median: 3.00; min-max: 1.25–8.00
CT pixel spacing (mm)	Median: 1.17; min-max: 0.86–1.52
PET/CT scanner model	
General Electric (all)	72
Discovery 690	40
Discovery STE	14
Discovery ST	8
Discovery RX	4
Discovery 600	3
Discovery 710	2
Discovery LS	1
Siemens (all)	105
Biograph HiRez (1080)	40
Biograph Truepoint (1093,1094)	27
Biograph mCT	25
Biograph LSO (1023,1024)	8
Biograph BGO (1062)	5
Philips (all)	103
Gemini TF	38
Gemini GXL	36
Allegro Body	19
Unspecified (Philips)	10

Supplemental Table 2: Results associated with the classification of high-uptake ROIs for two different groups of ROIs obtained with two different settings of the multi-foci segmentation algorithm.

	SUVmax > 2.5 (vol>2mL)	SUVpeak > Blood Pool (vol>2mL)
Total number of MFS* findings (ROI_{PARS})	18674	6737
Average number of ROI_{PARS} per subject (min–max)	66.7 (6–242)	24.1 (2–91)
Median number of ROI_{PARS} findings per subject (IQR)	59.0 (39.0–86.0)	19.0 (13.0–31.2)
Average misclassified number of ROI_{PARS} per subject (min–max)	6.9 (0–73)	3.6 (0–60)
Median misclassified number of ROI_{PARS} per subject (IQR)	4.0 (2.0–9.0)	2.0 (1.0–5.0)
Overall accuracy	0.90	0.85
Overall sensitivity	0.79	0.80
Overall specificity	0.92	0.88
Average classification accuracy per subject (min–max)	0.90 (0.40–1.00)	0.87 (0.34–1.00)
Median classification accuracy per subject (IQR)	0.93 (0.86–0.97)	0.89 (0.81–0.96)

*Multi-foci segmentation

Supplemental Table 3: Results associated with total metabolic tumor volume obtained using two different settings of the high-uptake region detection algorithm (multi-foci segmentation).

	SUVmax > 2.5 (vol>2mL)	SUVpeak > Blood Pool (vol>2mL)
Mean TMTV (min–max)	258.2 (0.0–2544.1)	235.2 (0.0–2471.9)
Median TMTV (IQR)	126.8 (37.8–295.0)	110.2 (32.9–280.8)
Average Dice with respect to the patient set of ROI_{REF} (min-max)	0.59 (0.00–0.99)	0.60 (0.00–0.99)
Median Dice with respect to the patient set of ROI_{REF} (IQR)	0.71 (0.31–0.86)	0.73 (0.33–0.86)
Average recall with respect to the patient set of ROI_{REF} (min-max)	0.56 (0.00–1.00)	0.53 (0.00–0.99)
Median recall with respect to the patient set of ROI_{REF} (IQR)	0.66 (0.23–0.85)	0.62 (0.20–0.81)
Average precision with respect to the patient set of ROI_{REF} (min-max)	0.79 (0.00–1.00)	0.89 (0.00–1.00)
Median precision with respect to the patient set of ROI_{REF} (IQR)	0.88 (0.72–0.96)	0.96 (0.86–0.99)
Spearman correlation coefficient with respect to reference TMTV	0.73	0.76

Supplemental Table 4: Results associated with the classification of high-uptake ROIs for four groups of ROIs obtained with two different settings of the multi-foci segmentation algorithm both with and without the neglection of ROIs with a volume between 0.1mL and 2mL.

	SUVmax>2.5 (vol>2mL)	SUVmax>2.5 (vol>0.1mL)	SUVpeak>Blood Pool (vol>2mL)	SUVpeak>Blood Pool (vol>0.1mL)
Total number of MFS* findings (ROI_{PARS})	18674	82114	6737	16717
Number of ROI_{PARS} per subject, average (min-max)	66.7 (6-242)	293.3 (11-1952) †	24.1 (2-91)	59.7 (2-689) †
Number of ROI_{PARS} per subject, median (IQR)	59.0 (39.0-86.0)	191.0 (91.8-428.5) †	19.0 (13.0-31.2)	39.5 (23.0-72.5) †
Classification accuracy per subject, average (min-max)	0.90 (0.40-1.00)	0.89 (0.46-1.00) ‡	0.87 (0.34-1.00)	0.85 (0.38-1.00) †
Classification accuracy per subject, median (IQR)	0.93 (0.86-0.97)	0.93 (0.83-0.97) ‡	0.89 (0.81-0.96)	0.87 (0.78-0.94) †
TMTV, average (min-max)	258.2 (0.0-2544.1)	275.9 (0.0-2571.9) †	235.2 (0.0-2471.9	244.8 (0.0-2488.3) †
TMTV, median (IQR)	126.8 (37.8-295.0)	142.2 (42.9-340.1) †	110.2 (32.9-280.8)	123.3 (35.9-295.6) †
Dice with respect to the patient set of ROI_{REF}, average (min-max)	0.59 (0.00-0.99)	0.60 (0.00-0.99) †	0.60 (0.00-0.99)	0.62 (0.00-0.99) †
Dice with respect to the patient set of ROI_{REF}, median (IQR)	0.71 (0.31-0.86)	0.71 (0.35-0.85) †	0.73 (0.33-0.86)	0.74 (0.39-0.88) †

*Multi-foci segmentation

†$p < 0.05$, ‡$p > 0.05$, Wilcoxon signed-rank test compared to the same variable obtained by neglecting ROIs with a volume below 2mL

Supplemental Table 5: Descriptive statistics related to the number of ROI_{PARS} and ROI_{REF} in the 280 subjects included in the study.

	ROI_{PARS}	ROI_{REF}
Total number of ROI	6737	7996
Average number of ROI per subject (min–max)	24.1 (2–91)	28.6 (1-201)
Median number of ROI per subject (IQR)	19.0 (13.0–31.2)	16.0 (6.0-38.0)

Supplemental Table 6: Results associated with the classification of ROIs with uptake significantly above the blood pool and volume above 2mL, when different levels of overlap are required to consider a ROI as matching the reference TMTV region.

	Overlap≥25%	Overlap≥50%	Overlap≥75%
Overall accuracy	0.85	0.85	0.84
Overall sensitivity	0.79	0.80	0.81
Overall specificity	0.91	0.88	0.85
Average misclassified number of ROI_{PARS} per subject (min–max)	3.5 (0-61)	3.6 (0-60)	3.9 (0-53)
Median misclassified number of ROI_{PARS} per subject (IQR)	2.0 (1.0-4.0)	2.0 (1.0-5.0)	2.0 (1.0-5.0)
Average classification accuracy per subject (min–max)	0.88 (0.33-1.00)	0.87 (0.34-1.00)	0.86 (0.42-1.00)
Median classification accuracy per subject (IQR)	0.90 (0.82-0.96)	0.89 (0.81-0.96)	0.88 (0.80-0.94)

Supplemental Table 7. Performance of the prediction of the occurrence of an event for both PFS and OS based on the TMTV cutoff thresholds selected by maximizing Youden's J index.

	Accuracy	**Sensitivity**	**Specificity**	**NPV***	**PPV†**
$TMTV_{PARS}$ PFS	0.60	0.66	0.57	0.79	0.41
$TMTV_{REF}$ PFS	0.61	0.67	0.58	0.80	0.42
$TMTV_{PARS}$ OS	0.63	0.67	0.62	0.89	0.28
$TMTV_{REF}$ OS	0.58	0.78	0.53	0.92	0.27

*Negative predictive value
†Positive predictive value