Original article

Nuclear Medicine Communications

Baseline metabolic tumor volume calculation using different SUV thresholding methods in Hodgkin lymphoma patients: interobserver agreement and reproducibility across software platforms

Francesca Tutino[a], Giulia Puccini[a], Flavia Linguanti[a], Benedetta Puccini[b], Luigi Rigacci[c], Sofya Kovalchuk[b], Roberto Sciagrà[a] and Valentina Berti[a]

Aim: **Although it is not yet used in clinical practice, metabolic tumor volume (MTV) assessed on the baseline FDG-PET has shown consistent prognostic value in various lymphoma types. The aim of our study was to compare interobserver agreement and reproducibility across platforms of MTV calculation using different SUV thresholding methods in a large series of patients with newly diagnosed Hodgkin lymphoma.**

Materials and methods: **We retrospectively studied 121 patients. MTV at baseline FDG-PET was independently computed by three readers with three programs of semi-automatic segmentation, Fiji, LifeX, and Accurate. MTV measurement was performed with different thresholds: SUV >2.5, SUV >4, and SUV >41% of SUV max.**

Results: **At inter-observer agreement analysis all Intraclass Correlation Coefficients (ICCs) were excellent (ICC >0.9), except for Accurate SUV >41% of SUV max (ICC = 0.8). The highest correlations were obtained at the SUV >4 threshold. The second best was SUV >2.5 threshold. Regarding reproducibility across software, we found statistically significant differences between Fiji versus LifeX and Accurate at fixed thresholds and between LifeX and Accurate at SUV >41% of SUV max, while no significant differences emerged between LifeX and Accurate using fixed thresholds.**

Conclusion: **The three SUV thresholds studied are all suitable for MTV calculation in terms of reproducibility. The best reproducibility is achieved using fixed thresholds, both SUV >4 and SUV >2.5. If more than one software has to be used in a study, we suggest the use of fixed thresholds and the platforms LifeX and Accurate. *Nucl Med Commun* 42: 284–291 Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.**

Nuclear Medicine Communications 2021, **42**:284–291

Keywords: FDG-PET/CT, Hodgkin lymphoma, metabolic tumor volume, reproducibility, SUV thresholding

[a]Nuclear Medicine Unit, Department of Experimental and Clinical Biomedical Sciences "Mario Serio", University of Florence, [b]Haematology Unit, Careggi University Hospital, Florence and [c]Haematology Unit, San Camillo Forlanini Hospital, Rome, Italy

Correspondence to Francesca Tutino, MD, Nuclear Medicine Unit, Department of Experimental and Clinical Biomedical Sciences "Mario Serio", University of Florence, Largo Brambilla 3, 50134 Florence, Italy

E-mail: tutinofrancesca82@gmail.com

Received 4 June 2020 **Accepted** 3 November 2020

Introduction

In recent years, the use of metabolic tumor volume (MTV) at baseline PET evaluation has been promoted to complement the currently used clinical prognostic scores, to better stratify patients based on the risk of recurrence and to tailor treatment. MTV has shown to be consistently prognostic across many lymphoma types: diffuse large B cell lymphoma [1,2], Hodgkin lymphoma [3–6], follicular lymphoma [7], and peripheral T cell lymphoma [8].

A number of methods and software platforms was used to calculate MTV: algorithms with fixed SUV thresholds, adaptive thresholds, and more advanced algorithms with 'perimeter of interest' [9]. The most used thresholding methods were the fixed threshold SUV >2.5 and the per-lesion threshold SUV >41% of SUV max, recommended by EAMN [10]. The heterogeneity of the methods led to different cut-offs for prognostic stratification and there is need to standardize MTV measurement before it can be implemented in clinical protocols for the patient benefit [11].

An ideal method for MTV measurement must be highly reproducible, not only inter-observer but also across software, accurate and precise, without forgetting user-friendliness, necessary for application in a busy clinical setting (Boellaard R, Buvat I. *Challenges of (total) metabolic active tumour volume measurements in lymphoma FDG PET/CT studies*; Mikhaeel G. *Can we use MTV in Clinical Practice? What should we expect?* PILM 2018 powerpoint presentations). Reproducibility is of primary interest and appears to be more important than accuracy, since a gold standard for the 'true' volume of disease is lacking in lymphomas (Buvat I. *How to choose a method for MTV measurement in lymphoma?* PILM2018 powerpoint presentation).

Several studies have investigated methodological aspects of MTV measurement. These studies focused mainly on

DOI: 10.1097/MNM.0000000000001324

method comparison, proving that MTV value strongly depends on the threshold used to outline it, and paid attention to inter-observer agreement and software reproducibility. Barrington et al. in 147 patients with diffuse large B cell lymphoma measuring MTV with the thresholds SUV >2.5, SUV >41% of SUV max and SUV > mean liver uptake (PERCIST), found excellent correlations between two readers using each method. The same group, with the threshold SUV >2.5 only, studied the reproducibility between commercial software and its in-house made software, which resulted excellent [1]. Cottereu et al. in a large series of 106 patients with peripheral T cell lymphoma demonstrated the non-inferiority of fixed SUV thresholds compared to adaptive methods on prediction of prognosis [8]. The group of Kanoun et al. studied 59 patients with Hodgkin lymphoma with fixed SUV thresholds (SUV >2.5 and SUV >41% of SUVmax) and thresholds based on SUV max of the liver, finding cutoffs significantly different across the methods but all well related to prognosis. Also in this study, the inter-observer agreement between two readers was excellent, while significant differences in MTV values emerged in software comparison [12].

In the present study, on a large series of patients with newly diagnosed Hodgkin lymphoma, likely to be representative of the population that is met in clinical practice, we aimed to assess extensively and systematically the inter-observer agreement for MTV measurement, computed with three SUV thresholds, SUV >4, SUV >2.5, and SUV >41% of SUVmax, and to evaluate for each threshold the reproducibility among three open-source software platforms, BI for Fiji, Lifex, and Accurate. In addition, our aim was also to assess, despite the differences among software in pre-processing algorithms, if two or more software were able to give analogous MTV results and could be considered interchangeable in multicentric studies.

Materials and methods

Patients

We retrospectively analyzed 121 patients of the Hematology Unit of our hospital, with a first histological diagnosis of Hodgkin lymphoma between 2010 and 2017, both in stage I–II and advanced. The clinical characteristics of the patient population are summarized in Table 1.

All patients were treated according to EORTC/GHSH recommendations.

The staging 18F-FDG PET was the essential requirement for entering the study. Only patients whose MTV was calculable were included.

Image acquisition

PET scans, from the mid skull to pelvis (standard lymphoma), were obtained 60 ± 10 min after FDG injection in a Gemini TF system (Philips Medical System, Cleveland, Ohio, USA). In particular cases, for example, the presence of known distal involvement, the acquisition fields have been extended. A clinical acquisition protocol was used with injection of 0.1 mCi/kg of 18F-FDG after at least 4 h of fasting and documentation of blood glucose <200 mg/dl. Before PET scans, a low dose CT (120 kV; 50–80 mA) was acquired to allow attenuation correction and lesions localization. PET images were reconstructed using an iterative algorithm (3D LOR RAMLA reconstruction with TOF, FOV: 576, matrix: 144 × 144, voxel dimension: 4 × 4 × 4 mm).

Table 1 Patient characteristics

Age (mean ± SD)	40 ± 15.59
Gender (M:F)	1.05
Stage (%)	I = 1.1
	II = 50.5
	III = 30.1
	IV = 18.3
Symptoms[a] (%)	a = 64.5
	b = 35.5

[a]a, asymptomatic; b, B symptoms.

Data analysis

PET images were independently evaluated by three readers (two nuclear medicine physicians and a resident in nuclear medicine), who measured MTV using three open-source software platforms, Beth-Israel (BI) plugin for Fiji [13], LifeX (Orlhac F, Nioche C, Buvat I. LIFEx user guide. LIFEx version 5.nn, Last update: June 12, 2019, https://www.lifexsoft.org) and Accurate [14].

Each measurement was performed with three different SUV thresholding methods, two based on the absolute SUV value, SUV >2.5 and SUV >4, and one relative, SUV >41% of the SUV max of each lesion.

SUV max and SUV peak of the reference lesion were also evaluated, calculated in SUVbw units for both variables.

Using Accurate, SUVpeak was measured as the SUVmean in a 1-ml volume of interest (VOI), positioned such to provide the highest value across all positions within the tumor (SUVpeak at maximum peak). Using Fiji, SUVpeak was calculated as the average SUV included in a 1-cm^3 region of interest (ROI) centered on SUVmax (SUVpeak at maximum). LifeX does not assess SUVpeak.

In order to make segmentation more automatic and less user-dependent as possible, we kept default settings suggested by the manufacturers in the user guides, in particular settings for each software were the following. In BI for Fiji, only SUV threshold (2.5; 4; 41%) and maximum SUV that can be processed (200) were set, without selecting the 'use CT' option. In LifeX, the initial thresholds were set as follows: absolute SUV threshold (set to 2.2) and Pruning Volume (set to 0.5 ml). In Accurate, the lower volume was set to 0.5 ml, as in LifeX. The software default settings have been maintained also for the maximum volume allowed (500 ml Fiji, 1000 ml LifeX, 500 ml Accurate).

In several cases, the operators had to manipulate the automatically outlined ROI to obtain an MTV value suitable for a hypothetical clinical use. The regions containing only physiological FDG uptake (brain, bladder, kidney, intestine, etc.) have been edited out. Furthermore, when the pathological tissue was automatically incorporated into the same ROI with adjacent physiologically active tissue, the ROI was removed and was redesigned by defining the limits of the VOI. This was done by defining the limits of the slices in which the lesion was included and redesigning it semi-automatically in Fiji ('draw' function), by the ROI 3D function, subtype 'click' in Lifex, by the 'mask' function on the 'Volume of interest' page in Accurate. At SUV >41% of the SUV max threshold, Accurate did not identify many lesions, despite SUV above the initial threshold SUV >4, which were added manually, as suggested by the user guide, on the 'Volume of interest' page or with the function 'Mask' or by clicking directly on the lesions in mode ROI 41% max, at the operator's discretion.

Ease of use and differences among software platforms in the segmentation method were also considered.

Statistical analysis

Statistical analysis was performed with the SPSS software (SPSS25). Quantitative variables, MTV, SUV max, and SUV peak, were expressed as means. Inter-observer agreement for MTV values was quantified as correlation consistency through the Intraclass Correlation Coefficient (ICC). To bring out the differences among methods, post-hoc analyses were performed by testing the Pearson coefficient (r), evaluated separately for each software, and for each threshold. Reproducibility across software platforms was also quantified with the ICC and the Pearson coefficients were assessed. A single reader data was used to obtain ICCs among software platforms. Post-hoc analyses were performed for all readers. Finally, MTV and SUV max values were tested with the General Linear Model (GLM) for repeated measurements with Bonferroni correction to highlight and quantify the effect of the reader and of the software.

Results

Patients

MTV was computable in 121 patients with newly diagnosed Hodgkin lymphoma. Ten patients from the original database were excluded, of whom four had disseminated disease, two had diffuse brown fat activation, and one had coexistence of other FDG avid pathology, all interfering with processing, two had inconsistent SUV due to incorrect normalization by the patient weight and one could not be processed by any of the three platforms, due to technical problems. In addition, LifeX with one or more thresholds (i.e., SUV >2.5 and SUV >41%) and Accurate SUV >41% of SUV max were not feasible in seven patients and in two patients, respectively, owing to technical problems.

Inter-observer reproducibility

Average MTV values obtained by the three readers using each thresholding method and ICCs are presented in Tables 2 and 3

All ICCs were excellent (ICC >0.9), except for Accurate SUV >41% of SUV max (ICC = 0.8).

The highest correlations were found for all software platforms at the SUV >4 thresholds, especially using LifeX and Accurate. The second best was the SUV >2.5 threshold.

In the post-hoc inter-observer analysis, Fiji had the strongest Pearson correlation at the SUV >4 threshold (r = 0.994). Fiji did not show the highest correlation between a pair of observers for a given threshold compared to the other software; however, the correlation coefficients were all excellent ($r \geq 0.9$).

LifeX showed the maximum inter-observer agreement at the SUV >4 threshold (r = 1), followed by SUV >2.5 threshold, which had a slightly weaker correlation. We obtained the minimum correlation at SUV > 41% of SUV max threshold ($r \sim 0.8$). Accurate behaved like LifeX, mostly with fixed thresholds: the maximum correlation is at SUV >4 (r = 1), the second-highest correlation was using SUV >2.5. At SUV >41% of SUV max threshold, the agreement among readers was the weakest ($r \sim 0.6$–0.7). In multivariate analysis for repeated measurements, we observed a statistically significant reader effect using Accurate at the SUV >41% of SUV max threshold ($P < 0.001$) (Fig. 1). This effect was related to the fact that reader 2 gave MTV measurements significantly different from reader 1 and reader 3 (difference reader 2 – reader 1 = –30.1 ml and difference reader 2 – reader 3 = –17.4 ml).

Table 2 Average MTV values by the readers

MTV	Reader 1	Reader 2	Reader 3
MTV FIJI 41%	125.7	119.8	127.7
LIFEX 41%	117.2	110.6	112.1
ACCURATE 41%	175.1	102.0	142.0
MTV FIJI 4	100.2	95.0	100.0
LIFEX 4	106.5	106.3	106.3
ACCURATE 4	107.3	106.6	107.1
FIJI 2.5	174.1	166.5	176.1
LIFEX 2.5	222.1	225.5	227.2
ACCURATE 2.5	233.4	232.3	231.5

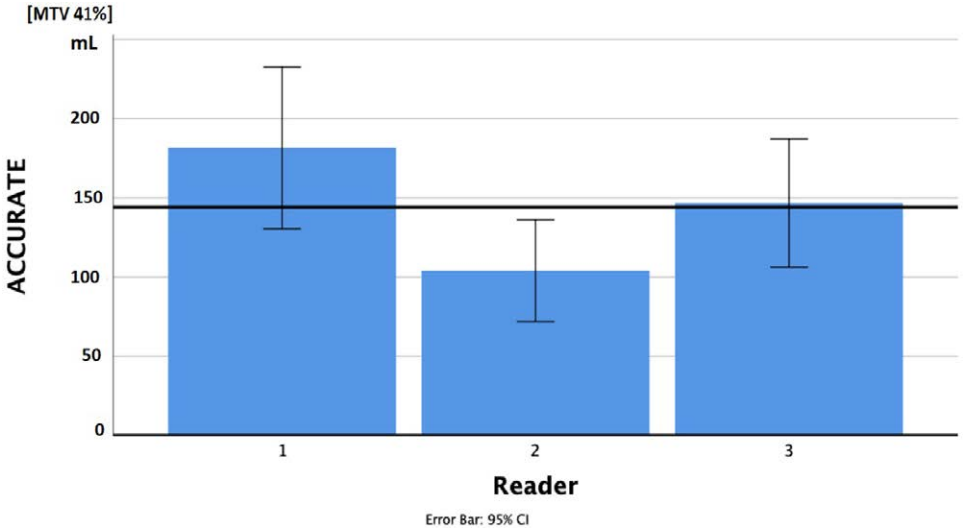
ICC, Intraclass Correlation Coefficient; MTV, metabolic tumor volume.

Table 3 Inter-observer ICC for different thresholding methods (software; threshold)

MTV	ICC
FIJI 41%	0.989
LIFEX 41%	0.961
ACCURATE 41%	0.852
FIJI 4	0.990
LIFEX 4	0.999
ACCURATE 4	0.999
FIJI 2.5	0.989
LIFEX 2.5	0.997
ACCURATE 2.5	0.996

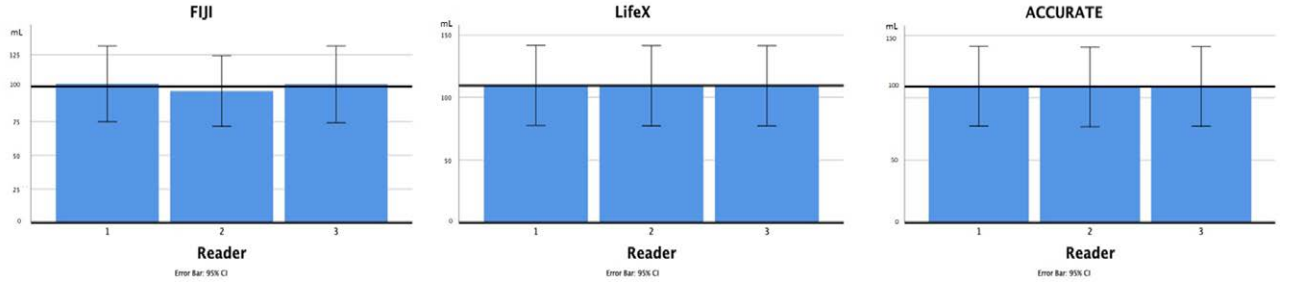
ICC, Intraclass Correlation Coefficient; MTV, metabolic tumor volume.

Fig. 1



Average MTV values by the readers using Accurate SUV >41% of SUV max.

Fig. 2



Average MTV values by the readers at the SUV >4 threshold using each software platform.

Table 4 Average MTV values using the three software platforms (single reader data)

MTV	FIJI	LIFEX	ACCURATE
41%	142.8	123.1	177.5
4	107.5	118.3	114.8
2.5	185.8	235.0	238.1

MTV, metabolic tumor volume.

Using the fixed thresholds, both SUV >4 and SUV >2.5, no statistically significant differences were found among reader measurements with any of the three software platforms ($P > 0.1$) (Fig. 2).

MTV reproducibility across software platforms

The average MTV values using the three software platforms and ICCs at each threshold are summarized in Tables 4 and 5.

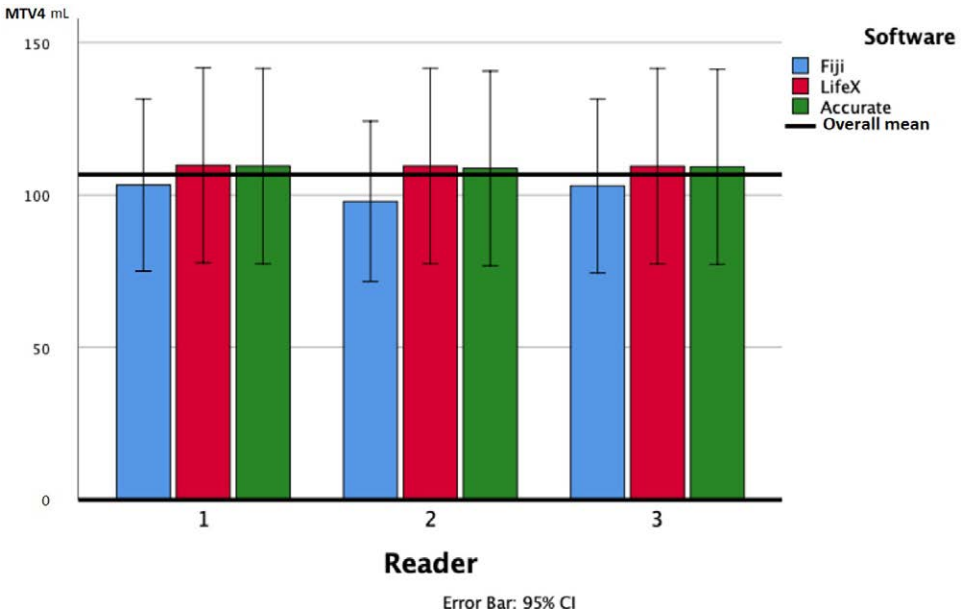
Table 5 ICC among software platforms at different thresholds

Threshold	ICC
41%	0.950
4	0.995
2.5	0.967

ICC, Intraclass Correlation Coefficient.

The SUV >4 threshold showed the highest correlation among software at ICC analysis (ICC >0.99). At post-hoc analysis, reader 1 obtained the maximum correlations at the SUV >4 threshold ($r > 0.99$), using any software, above all between Fiji and Accurate. Reader 2 showed the strongest correlation between LifeX and Accurate at the SUV >4 threshold ($r = 1$), while the second best was between LifeX and Accurate at the SUV >2.5 threshold. Reader 3 measurements showed the maximum correlations at the SUV >4 threshold, the

Fig. 3



Software comparison at the SUV >4 threshold. The bars represent average MTV values by the readers using the three software platforms and the solid line shows the MTV overall mean for the threshold.

highest between LifeX and Accurate (r = 1), followed by Fiji and Accurate.

At SUV >41% of SUV max threshold, multivariate analysis for repeated measurements showed a statistically significant effect of the software on MTV value (P = 0.002), linked to a significant difference between LifeX and Accurate (difference Accurate – LifeX = 27.7 ml).

The multivariate tests showed a software effect on the SUV >4 threshold, due to a statistically significant difference (P = 0.035) between Fiji and LifeX (difference LifeX– Fiji = 8 ml). The variability between Fiji and Accurate was borderline. No statistically significant difference was found between LifeX and Accurate (Fig. 3).

We observed a marked software effect on MTV measurement at SUV >2.5 threshold (P < 0.001), with significantly higher values using LifeX and Accurate compared to Fiji (difference LifeX – Fiji = 61.14 ml and difference Accurate – Fiji = 60.16 ml) (Fig. 4).

SUV reproducibility

Inter-observer agreement for SUV max and SUV peak and the average values by the readers are given in Tables 6–9.

Both SUV max and SUV peak had ICCs all greater than 0.99. The correlation among observers is greater with SUV peak than with SUV max at any threshold.

The GLM model applied to SUV max values showed a statistically significant effect related to the software used (P = 0.02) brought by LifeX, which gave an SUV max value significantly higher than Accurate (difference LifeX – Accurate = 0.05).

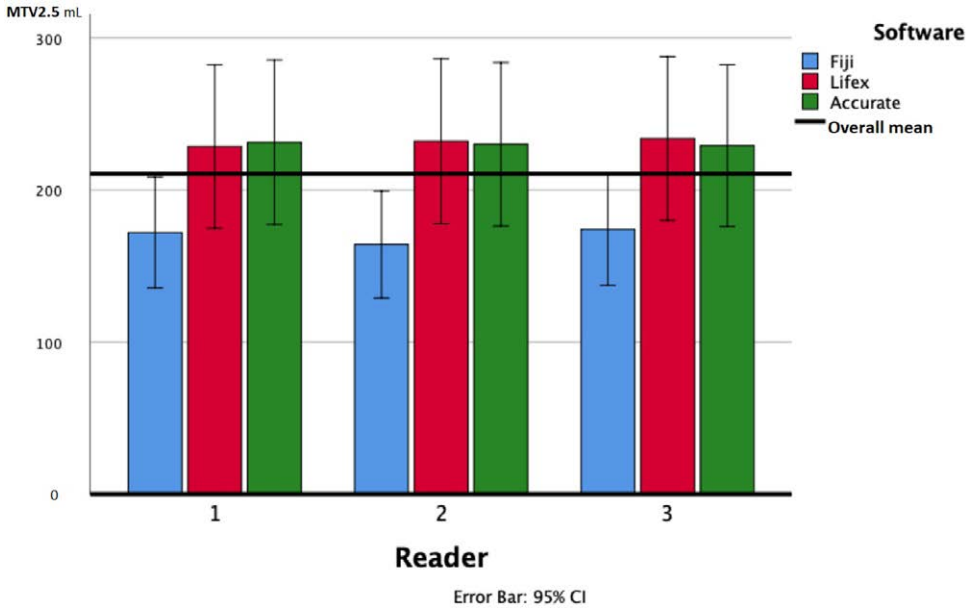
There were no statistically significant effects related to the reader (P > 0.4) or to the method used (P > 0.6).

Discussion

Despite the strong prognostic impact of MTV in lymphomas, a plethora of methods and software platforms are available for its calculation and no consensus has been reached on the best methodology. The preliminary evaluation of the reproducibility of the various thresholding methods is essential for the validation of the best method, but in the literature, an extensive and systematic analysis has not yet been conducted. We designed this study to evaluate the reproducibility of MTV measurements both in terms of interobserver agreement and stability among software platforms. We examined two published SUV thresholding methods, SUV >2.5 and SUV >41% of SUV max and a fixed threshold not yet applied in literature on lymphomas, SUV >4. Ease of use and differences among software platforms and thresholding methods in processing have also been considered.

Regarding differences in the segmentation, VOIs pattern and VOIs number depended on the threshold applied as well as the software used. The SUV >4 segmentation produced a limited number of clean VOIs, well dichotomized between tumor and aspecific uptake, but the less active portions of the tumor were excluded from the segmented volume. VOIs produced by a lower SUV

Fig. 4



Software comparison at the SUV >2.5 threshold. The bars represent average MTV values by the readers using the three software platforms and the solid line shows the MTV overall mean for the threshold.

Table 6 Average SUV max by the readers

SUV MAX	Reader 1	Reader 2	Reader 3
FIJI 41%	11.49	11.38	11.41
LIFEX 41%	11.27	11.19	11.26
ACCURATE 41%	11.27	11.15	11.26
FIJI 4	11.49	11.38	11.37
LIFEX 4	11.44	11.35	11.42
ACCURATE 4	11.35	11.26	11.34
FIJI 2.5	11.44	11.32	11.31
LIFEX 2.5	11.33	11.26	11.31
ACCURATE 2.5	11.22	11.15	11.20

Table 7 Average SUV peak by the readers

SUV PEAK	Reader 1	Reader 2	Reader 3
FIJI 41%	9.01	8.98	8.95
ACCURATE 41%	8.81	8.73	8.79
FIJI 4	9.06	9.04	8.95
ACCURATE 4	8.87	8.83	8.86
FIJI 2.5	8.90	8.87	8.80
ACCURATE 2.5	8.80	8.74	8.79

Table 8 Inter-observer ICC for SUV max values

SUV MAX	ICC
FIJI 41%	0.992
LIFEX 41%	0.999
ACCURATE 41%	0.997
FIJI 4	0.991
LIFEX 4	0.998
ACCURATE 4	0.998
FIJI 2.5	0.991
LIFEX 2.5	0.999
ACCURATE 2.5	0.999

ICC, Intraclass Correlation Coefficient.

Table 9 Inter-observer ICC for SUV peak values

SUV PEAK	ICC
FIJI 41%	0.994
ACCURATE 41%	0.999
FIJI 4	0.993
ACCURATE 4	0.999
FIJI 2.5	0.993
ACCURATE 2.5	0.999

ICC, Intraclass Correlation Coefficient.

threshold of 2.5, instead, were wider and more representative disease volume but with the disadvantage of a more difficult separation of the tumor from adjacent aspecific uptake voxels, in particular, using LifeX. Indeed, large SUV >2.5 VOIs tended to be mixed between tumor/physiological uptake in bulky disease with heterogeneous uptake. Thresholding based on low SUV cutoffs, SUV >2.5 and SUV >41% of SUVmax using LifeX (lower SUV allowable = 2.2), produced a number of aspecific VOIs corresponding to physiological sites of uptake. This required more complex editing out phase than the SUV >4 threshold.

Regarding reproducibility assessment, in this study, the inter-observer agreement for MTV values was excellent with all thresholds. Accurate SUV >41% of SUV max was an exception, showing a lower reproducibility. We obtained the best results with fixed thresholds according to previous comparative studies [1,12]. We had the strongest inter-observer agreement at the SUV >4 threshold, the second best was SUV >2.5. The slightly better

performance of the SUV >4 threshold, compared to SUV >2.5, is linked to its greater ease and automaticity: at SUV >4 threshold software were very specific. They found fewer physiological uptakes, leading to very little operator intervention. However, at SUV >4 threshold, MTV values resulted significantly lower compared with SUV >2.5 threshold. Indeed SUV >4 threshold considered only the areas of very intense hypermetabolism, equal to twice the liver uptake and underestimated the volume of the disease. An SUV >2.5 threshold, which should consider all the areas of hypermetabolism above liver uptake, seemed to allow a more accurate anatomical definition of the disease volume.

The Fiji software, even if it did not gain the strongest agreement among readers, in general showed the best inter-observer reproducibility. Indeed, correlations among readers were excellent at any threshold. This is related to favorable program features: preselection of lesions, the effectiveness of the editing-out phase, good spatial resolution and ease of use, which minimize operator intervention. An important determinant of inter-observer variability was the ROI manipulation by the operators, when lesions were automatically included in the same volume with adjacent physiological uptakes. For example: parapharyngeal lymph nodes/oropharynx, external iliac lymphadenopathy/bladder, or mediastinal bulky disease/bone, in presence of diffuse bone-marrow uptake. Manual intervention was rarely required using Fiji, while it had a strong impact on a purely thresholding segmentation software, such as LifeX and Accurate.

The SUV >41% of the SUV max threshold was significantly influenced by the reader using Accurate and shows great variability among software platforms. While the fixed thresholds are based on a simple volume calculation on the absolute value of SUV, the relative threshold algorithm is based on a percentage of uptake, resulting in more complex and exposed to ROI drawing variability. We noted a processing pitfall using Accurate SUV >41% of SUV max in bulky disease with heterogeneous uptake. In such cases, it was necessary sometimes to modify the volume semi-automatically or even manually, possibly increasing measurement variability. Therefore, the use of the SUV >41% of SUV max threshold should not be recommended with Accurate. As previously suggested by Barrington and colleagues in a study on patients with DLBCL [1], and also according to the analysis of our data on patients with Hodgkin lymphoma, the SUV >41% of SUV max threshold appears less performant in lesions with marked uptake heterogeneity, such as bulky lesions, since the automatic processing with the relative threshold failed in several cases, making it necessary a semi-automatic or manual intervention. Such a relative threshold appears to be more suitable for the study of solid tumors, than for lymphomas.

The fixed thresholds, both SUV >4 and SUV >2.5 were not influenced by the reader, but only by the software used. MTV values were significantly underestimated by Fiji compared to LifeX and Accurate. This could be due to Fiji's lower sensitivity since it is less performant in detecting small and not intense lesions. The reproducibility between LifeX and Accurate is excellent so that they can be considered interchangeable. Therefore, we could suggest that if using more than one software in a study is needed, in particular in multicentric studies, the use of LifeX and Accurate with fixed thresholds should be considered.

Finally, we found an optimal reproducibility of SUV values both among readers and among software platforms, with a better reproducibility of SUV peak compared to SUV max. The fact that all PET scans were performed on the same PET system and with the same acquisition protocol, certainly contributed to this finding. The greater reproducibility of SUV peak compared to SUV max is in line with data from the literature [15,16]. SUV peak, an average SUV calculated in a fixed-size VOI containing more pixels, was defined to be more reliable for the quantification of uptake than SUV max, which represents a single pixel and may reflect mere statistical fluctuations of activity in relation to the duration of the acquisition.

In conclusion, our study shows that all three thresholds studied are suitable for MTV computation in terms of reproducibility. Our major result is that an excellent inter-observer agreement and the best stability among software platforms is obtained with fixed thresholds, both SUV >4 and SUV >2.5, regardless of the software used. If more than one software has to be used in a study, we suggest the use of fixed thresholds, and the two programs LifeX and Accurate.

References

1 Ilyas H, Mikhaeel NG, Dunn JT, Rahman F, Møller H, Smith D, Barrington SF. 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.

2 Mikhaeel NG, Smith D, Dunn JT, Phillips M, Møller H, Fields PA, *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.

3 Cottereau AS, Versari A, Loft A, Casasnovas O, Bellei M, Ricci R, *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.

4 Moskowitz AJ, Schöder H, Gavane S, Thoren KL, Fleisher M, Yahalom J, McCall SJ, Cadzin BR, Fox SY, Gerecitano J, *et al*. Prognostic significance of baseline metabolic tumor volume in relapsed and refractory Hodgkin lymphoma. *Blood* 2017; **130**:2196–2203.

5 Akhtari M, Milgrom SA, Pinnix CC, Reddy JP, Dong W, Smith GL, *et al*. Reclassifying patients with early-stage Hodgkin lymphoma based on functional radiographic markers at presentation. *Blood* 2018; **131**: 84–94.

6 Rogasch JMM, Hundsdoerfer P, Hofheinz F, Wedel F, Schatka I, Amthauer H, Furth C. Pretherapeutic FDG-PET total metabolic tumor volume predicts response to induction therapy in pediatric Hodgkin's lymphoma. *BMC Cancer* 2018; **18**:521.

7 Meignan M, Cottereau AS, Versari A, Chartier L, Dupuis J, Boussetta S, *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.
8 Cottereau AS, Hapdey S, Chartier L, Modzelewski R, Casasnovas O, Itti E, *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.
9 Carlier T, Bailly C. State-of-the-art and recent advances in quantification for therapeutic follow-up in oncology using PET. *Front Med* 2015; **2**:18.
10 Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, *et al*.; European Association of Nuclear Medicine (EANM). FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging* 2015; **42**:328–354.
11 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.
12 Kanoun S, Tal I, Berriolo-Riedinger A, Rossi C, Riedinger JM, Vrigneaud JM, *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.
13 Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, *et al*. Fiji: an open-source platform for biological-image analysis. *Nat Methods* 2012; **9**:676–682.
14 Boellaard R. Quantitative oncology molecular analysis suite: accurate. *J Nucl Med* 2018; **59**:1753.
15 Sher A, Lacoeuille F, Fosse P, Vervueren L, Cahouet-Vannier A, Dabli D, *et al*. For avid glucose tumors, the SUV peak is the most reliable parameter for [(18)F]FDG-PET/CT quantification, regardless of acquisition time. *EJNMMI Res* 2016; **6**:21.
16 Boellaard R, Krak NC, Hoekstra OS, Lammertsma AA. Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study. *J Nucl Med* 2004; **45**:1519–1527.