Automated Segmentation of TMTV in DLBCL Patients: What About Method Measurement Uncertainty?

TO THE EDITOR: In baseline ^{18}F-FDG PET imaging of patients with diffuse large B-cell lymphoma (DLBCL), Barrington et al. recently confirmed that different outlining methods providing total metabolic tumor volume (TMTV) can be used to predict prognosis (*1*). An automated tool was applied for segmentation, focusing on the need in clinical practice for a fast, easy, and robust method. From the success–failure ratings of the visible-tumor delineation by 2 independent observers, involving minimal user interaction, the method based on a fixed SUV threshold of 4.0 g/mL (i.e., SUV4.0) was recommended for further evaluation, as well as a majority-vote method usually combining SUV4.0 and SUV2.5 (i.e., 2.5 g/mL fixed SUV threshold). Although different methods may provide significantly different TMTV outcomes, the authors suggested that bias in TMTV outcome is clinically less relevant than good reproducibility.

We fully agree with this suggestion but would like to stress that the study did not provide any quantitative information about the reproducibility percentage for each method—a quantification of the closeness of the agreement between TMTV outcomes obtained under changed conditions of measurement (*2*). These changed conditions may consist of different observers, as in Barrington's study, but also, in clinical practice, interscan time, scanning, and patient's conditions (including uptake time). Going further with the suggestion of Barrington et al., we believe that an outlining method providing a biased TMTV estimate—in other words, a surrogate—but accompanied by a significantly lower measurement uncertainty (here, for single scan) than that of SUV4.0 should be preferred for DLBCL prognosis (*2*). As a supporting example, although the ^{18}F-FDG SUV is only a surrogate for the metabolic rate of glucose consumption, its use no longer needs to be justified, because measurement uncertainty and availability are reasonable (*3*). It is noteworthy that such a reduced measurement uncertainty might compensate for the substantial measurement uncertainty expected for the TMTV cutoff from Figure 4 by Barrington, showing poor ($<$0.65) areas under the receiver-operating-characteristic curves (*1,4*). To summarize, the issue of a quick and easy method is indeed relevant in clinical practice, but we believe that it should not dominate the crucial measurement-uncertainty issue, even if too many clicks may affect inter- and intraobserver reproducibility. A 3- to 6-min TMTV measurement for most scans, depending on the method, seems to us a reasonable price to pay for patient management (*1*).

Furthermore, since the Quantitative Imaging Biomarkers Alliance profile for ^{18}F-FDG as an imaging biomarker for treatment-response assessment did not address the prognosis issue from a single scan, we take the opportunity to suggest that a TMTV cutoff for DLBCL staging should involve measurement uncertainty and, hence, be accompanied by asymmetric confidence limits of $100 \times \{\exp[\pm 1.96 \times \mathrm{SD}(d)/\mathrm{sqrt}(2)] - 1\}\%$, where SD($d$) is the SD of the differences in the test–retest TMTV-value logarithms (95% confidence) (*3,4*). Unlike a strict cutoff, these measurement uncertainty–derived upper and lower limits may reduce the number of false-positive and -negative scans for avoiding therapy escalation or undertreatment, respectively. This rationale offers the same flexibility as the use of liver or mediastinum SUV for assessing complete metabolic response in lymphoma patients according to treatment strategy. Strategy may also help to arbitrarily decide whether an outcome is false-positive or -negative when the outcome is close to a limit. The limits may be relevantly adjusted by expert consensus (e.g., changing 1.96 to 1 for 68% confidence).

To conclude, evaluating the best outlining method in clinical practice for assessing TMTV in DLBCL at baseline, along with determining the optimal TMTV cutoff to separate patients according to good or poor prognosis, are important issues for making treatment decisions. However, without any quantitative information about the measurement uncertainty of each method, we believe that recommendations are of limited scope. Repeated comments about the prognostic use of a strict cutoff for a continuous parameter, as well as a proposal for avoiding TMTV computing, might be taken into consideration (*4,5*).

COPYRIGHT © 2021 by the Society of Nuclear Medicine and Molecular Imaging.

REFERENCES

1. Barrington SF, Zwezerijnen BG, de Vet HC, et al. Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful? *J Nucl Med.* July 17, 2020 [Epub ahead of print].
2. Evaluation of measurement data—guide to the expression of uncertainty in measurement. Bureau International des Poids et Mesures website.https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf. Published September 2008. Accessed December 10. 2020.
3. Kinahan PE, Perlman ES, Sunderland JJ, et al. The QIBA profile for FDG PET/CT as an imaging biomarker measuring response to cancer therapy. *Radiology.* 2020; 294:647–657.
4. Laffon E, Marthan R. On the cutoff of baseline total metabolic tumor volume in high-tumor-burden follicular lymphoma. *J Clin Oncol.* 2017;35:919–920.
5. Laffon E, Marthan R. Could we avoid computing TMTV of DLBCL patients in routine practice? *Eur J Nucl Med Mol Imaging.* 2018;45:2235–2237.

Eric Laffon*
Roger Marthan
**Hôpital Haut-Lévèque*
Avenue de Magellan
33604 Pessac, France
E-mail: elaffon@u-bordeaux.fr

Published online Oct. 9, 2020.
DOI: 10.2967/jnumed.120.256214

REPLY: We thank Laffon and Marthan for their interest in our work (*1*) and for acknowledging that bias in metabolic tumor volume (MTV) outcome is less clinically relevant than good reproducibility. We agree that estimation of the reproducibility of MTV measurement methods is important to determine measurement uncertainty. We reported that agreement between observers for assessment of

MTV measurements using the same software was 91% for the method that uses 41% of maximum SUV and more than 95% for all other methods, and we considered this to be good agreement (1). The success rate of MTV measurement was unaffected by scanning conditions (whether compliant or not with the EANM Research Ltd. harmonization program) and the presence or absence of subsequent disease progression. The uptake time influenced the success rate of measurements for the method that uses 41% of maximum SUV and the method that uses majority vote 3, which were less successful with longer uptake times.

Laffon and Marthan propose that MTV cutoffs derived from PET data to guide discrimination of prognosis should be accompanied by upper and lower confidence limits based on measurement uncertainty. The main purpose of our work was not to derive cutoffs to discriminate prognosis but to take a first step to answer a methodologic question, which was to determine the optimal automatic segmentation method or methods for MTV to apply in a larger cohort. The criteria in our study focused on 2 aspects. First, did the MTV measurement methods generate plausible total tumor burden segmentations? This was prioritized over precision, as good repeatability does not necessarily provide meaningful results. Thereby, whether such (known) precision should subsequently be used to define a threshold uncertainty or gray zone is a matter of effect size in the studied population and the intended use of the biomarker. Second, to apply a method clinically or in trials, the segmentation and workflow should be fast and easy to use and have minimal observer interaction. By applying these criteria, we identified 2 candidate methods (majority vote 2 and the method based on a fixed SUV threshold of 4.0 g/mL) that can be considered for further MTV biomarker validation. For individual patient assessment to guide prognosis and when the ultimate goal is to offer personalized treatment, MTV should ideally be assessed as a continuous variable. Then, cut points and measurement errors or misclassification become less relevant.

We presented data on discriminatory power to confirm similarity for the different segmentation methods as shown previously (2) and to support the argument that choice of method can be based on ease of use and success rates in giving plausible volumes under various conditions. For the current study, we used a case-control design to test parameters that might influence the best segmentation method—meaning that the patient population and any derived cutoffs would not be representative of usual clinical practice. We are progressing with MTV measurement in a large warehouse of clinical and scan data in patients with non-Hodgkin lymphoma (https://petralymphoma.org/). Sufficient data are required to derive robust optimal MTV cutoffs for training, validation, and test datasets. In these studies, measurement error, confidence limits, and uncertainty will be considered.

Finally, MTV is a robust predictor of prognosis in diffuse large B-cell lymphoma but will likely need to be factored into an algorithm with baseline clinical factors, including the international prognostic index (3), and potentially with emerging biomarkers that reflect tumor dissemination and molecular heterogeneity (4,5) and dynamic response markers (3,4).

DISCLOSURE

Sally Barrington acknowledges support from the National Institute for Health Research (NIHR) (RP-2-16-07-001). King's College London and UCL Comprehensive Cancer Imaging Center 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.

REFERENCES

1. Barrington SF, Zwezerijnen BG, de Vet HC, et al. Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful? *J Nucl Med.* July 17, 2020 [Epub ahead of print].
2. Ilyas H, Mikhaeel NG, Dunn JT, et al. Defining the optimal method for measuring baseline metabolic tumor volume in diffuse large B cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2018;45:1142–1154.
3. Mikhaeel NG, Smith D, Dunn JT, et al. Combination of baseline metabolic tumor volume and early response on PET/CT improves progression-free survival prediction in DLBCL. *Eur J Nucl Med Mol Imaging.* 2016;43:1209–1219.
4. Kurtz DM, Scherer F, Jin MC, et al. Circulating tumor DNA measurements as early outcome predictors in diffuse large B-cell lymphoma. *J Clin Oncol.* 2018; 36:2845–2853.
5. Cottereau AS, Nioche C, Dirand AS, et al. ^{18}F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. *J Nucl Med.* 2020;61: 40–45.

Sally F. Barrington*
Ben G.J.C. Zwezerijnen
Henrica C.W. de Vet
Martijn W. Heymans
Ronald Boellaard*
**St. Thomas Hospital,*
London SE1 7EH, U.K.
E-mail: sally.barrington@kcl.ac.uk

Published online Oct. 30, 2020.
DOI: 10.2967/jnumed.120.257030

Data-Driven Motion Correction in Clinical PET: A Joint Accomplishment of Creative Academia and Industry

TO THE EDITOR: I read with great interest the recent *JNM* article by Walker et al. comparing data-driven and hardware-driven motion correction technologies in PET (1). The former is an important innovation, and its transition into the marketplace is exciting to see. Publications such as this one play a pivotal role in the technology's acceptance and broader dissemination. However, this work is very similar to work from our group published in 2016 (2), and unfortunately, our publication was not properly referenced.

Like Walker et al., we compared nongated, software-gated, and hardware-gated images head-to-head in a large set of clinical PET scans, using quantitative analysis of lesion uptake and qualitative masked reviewer scoring of image quality, with similar results—a statistically significant preference for software-gated images over hardware-gated images and with similar ratios of performance metrics. There are, of course, subtle differences between the gating approaches, and Walker et al. note that their work validates newly available commercial technology. Given that this work focused on commercial product testing, it should add scientific context to note that the *key points* they presented also describe our earlier findings.