Journal of Nuclear Medicine, published on April 12, 2019 as doi:10.2967/jnumed.119.227249

TIME TO PREPARE FOR RISK ADAPTATION IN LYMPHOMA

BY STANDARDISING MEASUREMENT OF METABOLIC TUMOUR BURDEN.

Sally F Barrington (1) and Michel Meignan (2)

(1) King's College London and Guy's and St Thomas' PET Centre, School of Biomedical Engineering and Imaging Sciences, King's College London, King's Health Partners, London, UK, (2) Lymphoma study association Imaging (LYSA-IM), Henri Mondor University Hospitals, Functional Imaging and Therapeutics Department, University Paris Est Créteil, Créteil, France

Corresponding author: Sally Barrington, School of Biomedical Engineering and Imaging Sciences

4th floor Lambeth Wing, St Thomas Hospital, Westminster Bridge Road, London SE1 7EH UK

00 44 207 188 8364 (phone) 00 44 207 620 0790 (fax)

sally.barrington@kcl.ac.uk

michel.meignan@aphp.fr

Word count : 6349

Running title : Time to standardise MTV measurement

ABSTRACT

Increased tumour burden is associated with inferior outcomes in many lymphoma subtypes. Surrogates of tumour burden that are easy to measure, such as the maximum tumour dimension of the 'bulkiest' lesion on CT have been used as prognostic indices for many years. Recently, the total metabolic tumour volume (MTV) and tumour lesion glycolysis (TLG) have emerged as promising and robust biomarkers of outcome in various lymphomas. The median MTV value and the optimal cut-off points to separate patients into risk groups in a study population are however, highly dependent on the population characteristics and the delineation method used to outline tumour in the PET image. This has precluded the use of MTV for risk stratification in trials and clinical practice. Standardisation of the methodology is timely to allow the potential for risk adaptation to be explored in addition to response adaptation using PET.

Meetings between representatives from research groups active in the field were held under the auspices of the PET international lymphoma and myeloma workshop. A summary of those discussions, which included a review of the literature and a practical assessment of methods used for outlining, including various software options is presented.

Finally, a proposal is made to perform a technical validation of MTV measurement enabling benchmark reference ranges to be derived for published delineation approaches used for outlining with various softwares.

This process would require i) collation of representative imaging datasets of the most common lymphoma subtypes, ii) agreement on pragmatic criteria for the selection of lesions, iii) generating a range of MTV values with consensus to be reached on final contours in a training set, and iv) developing automated software solutions with a set of minimum functionalities to reduce measurement variability.

Methods developed in the above training exercise could then be applied to another dataset with a final set of contours and values generated. This final dataset would provide a benchmark against which end-users could test their ability to measure MTV consistent with expected values. The dataset and automated software solutions could be shared with manufacturers with the aim to include these in standard workflows to allow standardisation of MTV measurement across the world.

Keywords: lymphoma; positron emission tomography; standardization

INTRODUCTION

The association of tumour burden with resistance to treatment in Hodgkin lymphoma (HL) and inferior patient outcomes has been recognised for 30 years (1). Assessment of tumour volume at that time was performed using clinical examination, chest X-ray and lymphography (1), later replaced by computed tomography (CT) (2) . These studies demonstrated that tumour burden was the single most important prognostic factor at the time of diagnosis for the prediction of treatment failure and disease relapse. The MabThera International Trial (MInT) demonstrated the survival benefits of combining rituximab with chemotherapy in young diffuse large B cell lymphoma (DLBCL) patients with good-prognosis disease (3). In this landmark study, the presence of bulky disease was the only independent clinical risk factor associated with overall survival (OS) with a linear effect observed, using cut-off points from 6cm to 10cm for maximum tumour dimension. In a further trial in young patients with DLBCL, with an age adjusted IPI of 1, a maximum tumour dimension of ≥ 10cm was the only factor associated with OS (4) . Similar findings were reported around the same time in follicular lymphoma (FL), where the longest diameter of the largest involved node was identified as an independent predictor of progression-free survival (PFS) with an optimal cut-off point of 6cm (5).

The time involved and the complexity of measuring the entire tumour volume in individual patients on CT scans has meant that surrogates for the total tumour burden have been relied upon as predictive factors. Disease stage, number of involved nodes, involvement of extranodal sites and the presence of bulk have been included in prognostic indices that are commonly used in Hodgkin and non-Hodgkin lymphoma (NHL) (5-9). These prognostic indices, however, do not classify patients at high risk of treatment failure very effectively. Tumour volumes on PET and CT are routinely assessed for the purposes of radiotherapy planning, but this is generally limited to assessment of one site or a few sites rather than evaluation of the total tumour burden.

The introduction of positron emission tomography (PET) has made measurement of the total metabolically active volume of tumour more feasible. Tumour locations that accumulate 18F-fluorodeoxyglucose (FDG) can be outlined and summed together to calculate the total metabolic tumour volume (MTV). Tumour lesion glycolysis (TLG) can also be assessed, which is the MTV multiplied by the mean standardised uptake value (SUV) in the entire volume and takes account of both the extent and intensity of tracer uptake. Multiple reports from large studies performed in retrospective cohorts or retrospective analyses of prospective trials have demonstrated that MTV and/or TLG is associated with progression free survival (PFS) and some with overall survival (OS) in subtypes including HL, DLBCL, FL, primary mediastinal B cell and T cell lymphomas (10-18). Highly effective PET adapted treatment may have contributed to the inability to show an association with OS in some studies. MTV appears to be a robust prognosticator irrespective of the method used for measurement. However the median MTV and/or optimal cut-off point that separates high from low risk groups varies according to the patient population and the method of analysis. This has precluded the use of metabolic volumes for risk stratification in clinical trials in haematological malignancies to date.

Standardisation of the methodology for the assessment of metabolic tumour burden is required to validate this promising biomarker to enable inclusion in patient management. Standardisation of response assessment using FDG-PET has previously been successful, using the Deauville criteria (19) which are widely applied (20) and used for PET-response adapted treatment (21). This was undertaken as a sequential process, firstly agreeing simple rules for reporting and measuring concordance rates between reviewers using international cohorts of patients with HL and NHL (technical validation) (22,23). Secondly, the criteria were evaluated against patient outcomes in retrospective cohorts (24-27) and prospectively tested in clinical trials (clinical validation) (21,28,29). A similar approach to standardise the measurement of metabolic tumour burden is now proposed to enable testing of PET '*risk-adapted'* as well as '*response-adapted*' strategies.

METHODS

Meetings were convened with representatives from research groups active in the field under the auspices of the PET International Lymphoma and Myeloma (PILM) Workshop; https://www.lymphomapet.com/ . A review of the literature and studies in progress was undertaken with presentations and face-to-face meetings in Paris on 01/02/2018 and Menton on 04/10/2018. A proposal was developed to perform a benchmarking exercise for the technical validation of MTV and TLG in FDG PET-CT images.

The group acknowledged uncertainties regarding i) what structures to include, ii) the best delineation method(s) to apply and iii) which software package(s) to use to outline tumour.

The following section summarises the results of discussions and potential ways forward. The term 'cut-off point' is used to mean the MTV cut-off that separates patients into different risk groups. 'Threshold' is used to mean the threshold applied in the segmentation method to delineate tumour.

RESULTS

What should be included in the Assessment of MTV

Measurement begins with visual assessment of the scan, as occurs in routine clinical practice, noting the location of abnormal focal uptake in nodal and extranodal sites, ensuring that all relevant areas are imaged. Images should be scaled to a fixed SUV display and colour table (20). Lymphomatous uptake can be distinguished from physiological uptake and disease unrelated to lymphoma according to the distribution and/or CT characteristics with knowledge of the lymphoma subtype by a trained observer (20).

A pragmatic approach is required for measurement of MTV and TLG to be feasible in clinical practice, with the intention to capture the main areas of tumour bulk. It may not be possible or desirable to include every small involved node or areas that are difficult to measure, for example, diffuse disease in the bone marrow. A minimum volume, perhaps 2 or 3 mL at baseline, is suggested to avoid including multiple small regions that may be time consuming to measure when a manual method is used, but which do not contribute much to the overall volume (30,31). Smaller volumes may however need to be measured at the point of response assessment, as tumour residuals may be small. Acknowledging the uncertainties of this approach, technical validation could include measurement of the volume within compartments e.g. nodal, splenic and bone marrow compartments as well as the total volume.

It is proposed to include in the assessment of MTV/TLG

- viable areas in lymph nodes with increased FDG uptake above a specified threshold
- focal uptake in the spleen, irrespective of splenic size
- diffuse increased uptake in the spleen, in the absence of reactive changes in bone marrow, greater than the liver uptake (i.e. where there is a reversed hepato-splenic ratio)
- focal uptake in the bone marrow .

It is uncertain exactly how to classify an abnormal hepato-splenic ratio. Splenic uptake greater than 1.5 times the liver uptake has been used previously but has not been validated (10,32,33). It is our experience that reactive changes in the bone marrow are often accompanied by similar changes in the spleen and it is suggested that diffuse uptake in the spleen should not be included in the volume in this situation.

Diffuse uptake in the bone marrow occurs in approximately 1 in 5 patients with HL (34) almost always due to reactive change and it was considered that it should not be included in the MTV. In DLBCL, diffuse uptake is more likely to indicate reactive change than lymphomatous involvement in the bone marrow compartment, however where diffuse uptake is due to bone marrow involvement, this is usually reflective of diffuse low volume, sometimes discordant cellular infiltration (35-37) which probably has less of an impact on prognosis (38,39) than areas of tumour bulk. In FL, diffuse cellular infiltration of the bone marrow involvement is commonly missed by FDG-PET (40). Patients referred for PET scanning with FL are typically patients with high tumour burden who are being considered for immunochemotherapy, in whom the inclusion of bone marrow with diffuse cellular infiltration may be less important. For these reasons it is suggested to include focal uptake only for the computation of MTV for the three most common lymphoma subtypes of HL, DLBCL and FL. In occasional cases in DLBCL there may be mainly marrow-based disease, with intense abnormal diffuse FDG uptake, confirmed on biopsy to represent bone marrow involvement; then diffuse marrow uptake should be included in the measurement (Fig. 1).

Which threshold(s) should be applied to segment MTV

Satisfactory image quality and accurate quantification are key to ensure reliable measurement of metabolic tumour burden. Solutions to deal with uncertainties in technical and biological factors (41) are included in international guidance (42) and are commonly applied in trials and clinical practice for tumour imaging.

The segmentation of tumour in patients with lymphoma is considerably more complex than with solid tumours. There may be multiple sites of involvement in nodes and different extranodal sites, with large variability in lesion size, shape, heterogeneity of uptake and number (Fig. 2). Various contouring thresholds have been applied to outline tumour in lymphoma patients, perhaps because of this complexity. Results have been reported using absolute SUV thresholds applied to the entire image. The threshold may be fixed e.g. SUV equal to or greater than 2.5 (13,16,43,44) , SUV equal to or greater than 4.0 (45) or relative to a reference region such as the liver and/or mediastinum (46,47) as suggested in the PET response criteria in solid tumours (PERCIST) (48). Results using percentage thresholds have also been reported e.g. outlining 41% (49-51) or 25% of the maximum SUV in individual lesions then summing them to calculate MTV (14,15). More complex image processing methods including gradient thresholds based on the changes in intensity of uptake at the edges of lesions (52), source to background corrected contours (53) and statistical methods, such as clustering (54), FLAB (55) and others have been proposed, but not applied much in lymphoma nor possibly providing any clinical advantage over simpler methods (56).

The success of any delineation method will be influenced by tumour and imaging characteristics. The minimum mean and maximum SUV in the tumour and the spatial distribution will affect quantification (53,57). Significant underestimation of visible tumour may occur with absolute thresholds if many voxels in a tumour mass have low uptake, less than the threshold value (Fig. 3) and conversely overestimation of tumour, if tumour lies adjacent to areas with high physiological uptake with spillover of counts into normal tissues (47). Underestimation occurs with percentage thresholds when there is a high maximum SUV and heterogeneity of uptake, with a large number of voxels that have uptake which is lower than the threshold (Fig. 3) and conversely overestimation when the maximum SUV in the tumour is low but significant (e.g. SUV 4) and many voxels in the surrounding background are included in the contour.

Image noise, the matrix size, image resolution and reconstruction will also affect SUVs (58,59) although the impact of varying these parameters will be relatively more important in patients with smaller tumour volumes and have less impact in patients with advanced disease and large tumour volumes (53). MTV and TLG are much less sensitive to these influences than baseline metrics such as the maximum and peak SUV, and MTV is less affected by these imaging characteristics than the TLG (which is the product of MTV and the mean SUV in the entire volume).

Irrespective of these challenges and the various thresholding methods applied to outline tumour in lymphoma, MTV and TLG remain strong, prognostic indicators of patient outcomes (47). The different thresholds also appear to have good reproducibility between observers (47). Importantly however, the use of different thresholds leads to different median values in study populations and consequently to different optimal cut-off points to separate patients into high and low risk groups (Table 1). The characteristics of the study population including the range of volumes and the efficacy of treatment also influence the cut-off points (60). This means that the optimal cut-off points for prediction of risk using MTV and TLG may be unique to the particular patient characteristics, lymphoma subtype and treatment and need to be derived for specific situations.

Each thresholding method clearly has limitations and currently it may not be possible to decide on a single best method. It may be worthwhile to investigate approaches proposed in radiation oncology, to reduce interobserver variation, whereby more than one threshold is combined using semAi-automated contouring to outline tumours. These methods include the STAPLE algorithm (https://www.ncbi.nlm.nih.gov/pubmed/15250643) and the majority vote, where only voxels selected using the majority of segmentation methods are included in the final outline (61). Artificial intelligence methods also appear promising, with selection of imaging features used as the basis for choosing one of several segmentation methods in an individual patient (e.g. ATLAAS algorithm)(62). The rationale is that no single thresholding method will perform optimally in every patient, however combining voxels included in the tumour outline by more than one delineation method will be close to the best performing method in the majority. Evaluation of absolute and percentage thresholds is likely to be required in a benchmarking exercise for the technical validation of MTV.

Which software packages should be used and are manual or automatic approaches better

Given that all thresholds appear to perform in a similar way to predict patient outcomes, the most important requirements for a suitable measurement method are high success rates for segmenting visible tumour, ease of use and provision of quick, consistent results, suitable for testing in multicentre trials and ultimately clinical application.

Various software options exist for measuring MTV/TLG and some work better than others using different thresholds. Broadly speaking, most use some form of automatic segmentation which can then by adjusted manually. This may comprise the observer 'point-picking' areas of tumour whilst avoiding areas of physiological uptake, or fully automated selection of regions of uptake, applying one or more thresholds with subsequent removal of physiological uptake by the observer.

The former using seed growing algorithms for 'point-picking' is often easier when there are few areas of tumour present which are well separated from areas with high uptake such as the brain, heart or urinary system (Fig 2C and 2D). In this scenario, the total MTV can be measured rapidly without the need for further editing but is more observer dependent and time consuming when there is multifocal tumour than fully automatic segmentation.

Fully automated segmentation is easier with multiple tumour regions (Fig. 2E) but always requires removal of physiological uptake. Cropping to avoid slices at the top (e.g. including brain uptake) and bottom of the image (e.g. bladder uptake) may reduce the amount of editing required, if the tumour distribution allows, but this is sometimes difficult, especially in FL.

The software that performs best will therefore vary by disease distribution and threshold chosen and the two approaches should be combined in the same software package. Academic groups have developed shareware for research, recognising that automation is highly desirable. These include LIFEx; https://www.lifexsoft.org (63) , FIJI; https://fiji.sc/ and ACCURATE (64) tools. Ultimately though, engagement with manufacturers is important for regulatory approval and maintenance and development of the software for clinical use. Proprietary software solutions for measuring metabolic tumour burden using adaptive thresholding have been approved and whilst very useful for general reporting, are not widely applicable.

DISCUSSION

Where to go from here

It is proposed to collect representative baseline scans from patients with early and advanced HL, early and advanced DLBCL and FL with high tumour burden. Scans could be collated from existing international published datasets, the number in each group to be decided, which are representative of the variation in FDG uptake and image quality seen in clinical practice using a range of available technologies (Fig. 4).

Consensus criteria for inclusion of lesions in MTV and TLG could be formulated, based on pragmatic choices, as suggested in this manuscript. Measurement could be undertaken using available automated software developed by academic groups or with a new consensus method with region preselection based on the commonly used absolute and percentage thresholds with minimum volumes to be agreed upon based on similar work in radiation oncology (61). Using consensus criteria and automated selection of regions, MTV values and ranges could be generated for a training dataset using two or more thresholds by observers from international groups. The final consensus contours should be agreed upon by an expert panel. Detailed instructions based on this training dataset will allow reference MTV values to be generated for a separate test dataset. This dataset could provide a benchmark against which end-users in trials and clinical practice could test their ability to measure MTV consistent with the expected values.

Automated software solutions could be shared with manufacturers, with a set of minimum functionalities required to minimise MTV measurement variability. Manufacturers should be encouraged to include these tools in standard work packages. This technical validation is the first step that needs to be taken prior to testing MTV and TLG as prognostic markers in specific patient populations to define suitable cut-off points for risk stratification of patients treated with standard (and experimental) therapy in prospective studies or retrospective analyses of prospectively acquired datasets.   Risk stratification using MTV will likely involve integration with other baseline parameters such as clinical prognostic scores (10,13,50), possibly as continuous variables and perhaps in combination with response assessment (13,65).

CONCLUSION

We believe that segmentation of MTV

- should require minimal observer interaction (although this is inevitable in some cases)
- should not be vendor specific and work in different software environments
- needs commercial support and regulatory approval
- should be ideally integrated into the clinical workflow of all platforms, without the requirement to purchase separate packages for volume measurement
- should comply with the proposed benchmark standard as suggested in this paper

If these requirements are fulfilled, different softwares implementing the same delineation methods and used with the same settings should give MTV values within an acceptable pre-specified range everywhere in the world.

DISCLOSURES

Professor Barrington acknowledges support from the National Institute of Health Research (NIHR) [RP-2-16-07-001]. King's College London and UCL Comprehensive Cancer Imaging Centre is funded by the CRUK and EPSRC in association with the MRC and Department of Health (England). Professor Meignan is the scientific coordinator of the LYSA Imaging. LYSA is a cooperative group of haematological centres involved in Lymphoma care. It is a non-profit organization agreed by the French National Cancer Institute. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR the Department of Health UK or the LYSA. No other potential conflicts of interest relevant to this article exist.

ACKNOWLEDGEMENTS

The authors wish to thank our colleagues who contributed to this manuscript, including the

experts who either participated in the focus group meeting in Paris:

R Boellaard, I Buvat, O Casasnovas, S Chauvie, AS Cottereau, A Gallamini, S Kanoun, C Kobe, NG Mikhaeel, L Pike, R Ricci, A Versari

and/or presented data or chaired sessions about this topic during the main PET International Lymphoma and Myeloma (PILM) Workshop in Menton :

A Alizadeh, R Boellaard, I Buvat O Casasnovas ,C Haioun, M Hutchings, S Kanoun , L Kostakoglu, D Kurtz, NG Mikhaeel, A Moskowitz, J Trotman, L Rigacci, H Schoder, L Schwartz, S Stroobants, H Tilly, J Zaucha

and other members of the scientific committee who contributed to the closed expert session about this topic at PILM:

M Andre, A Avigdor, S Becker, A Biggi, F Bodere, C Bodet-Milin, L Ceriani, M Coronado, E Dann, U Dührsen, V Edeline, T El Galaly, M Federico, R Fonti, MJ Garcia Velloso, C Gisselbrecht, M Gregianin, C Haioun, K Herrmann, A Hüttmann, R Klüge, A Lister, S Luminari, M Martelli, P Moreau, C Nanni, W Oyen, M Picardi, A Riedinger, J San Miguel, S Stroobants, I Tal, JN Talbot, C Thieblemont, H Tilly, E Van Zele, T Vander Borght, E Zamagni, J Zijlstra, E Zucca.

The authors thank R Boellaard, I Buvat, S Chauvie, C Kobe and A Versari for review and appraisal of the manuscript.

Key points
Question: What steps are required to standardise measurement of metabolic tumour volume (MTV) for patients with lymphoma?
Findings: A technical validation of MTV measurement is proposed, which will require a training set of patients with lymphoma to be developed and tumour volumes delineated to make MTV measurements, using agreed selection criteria and automated software solutions. Methods developed in the training exercise will be used to create a benchmark dataset with tumour contours and MTV measurements.
Implications for patient care: End-users, including software manufacturers, can then test their ability to measure MTV consistent with expected values with the aim to include MTV measurement in standard workflows to allow standardisation of MTV measurement across the world.

REFERENCES

1. Specht L, Nordentoft AM, Cold S, Clausen NT, Nissen NI. Tumor burden as the most important prognostic factor in early stage Hodgkin's disease. relations to other prognostic factors and implications for choice of treatment. *Cancer*. 1988;61:1719-1727.

2. Gobbi PG, Ghirardelli ML, Solcia M, et al. Image-aided estimate of tumor burden in Hodgkin's disease: Evidence of its primary prognostic importance. *J Clin Oncol*. 2001;19:1388-1394.

3. Pfreundschuh M, Ho AD, Cavallin-Stahl E, et al. Prognostic significance of maximum tumour (bulk) diameter in young patients with good-prognosis diffuse large-B-cell lymphoma treated with CHOP-like chemotherapy with or without rituximab: An exploratory analysis of the MabThera international trial group (MInT) study. *Lancet Oncol*. 2008;9:435-444.

4. Recher C, Coiffier B, Haioun C, et al. Intensified chemotherapy with ACVBP plus rituximab versus standard CHOP plus rituximab for the treatment of diffuse large B-cell lymphoma (LNH03-2B): An open-label randomised phase 3 trial. *Lancet*. 2011;378:1858-1867.

5. Federico M, Bellei M, Marcheselli L, et al. Follicular lymphoma international prognostic index 2: A new prognostic index for follicular lymphoma developed by the international follicular lymphoma prognostic factor project. *J Clin Oncol*. 2009;27:4555-4562.

6. Hasenclever D, Diehl V. A prognostic score for advanced Hodgkin's disease. international prognostic factors project on advanced Hodgkin's disease. *N Engl J Med*. 1998;339:1506-1514.

7. Solal-Celigny P, Roy P, Colombat P, et al. Follicular lymphoma international prognostic index. *Blood*. 2004;104:1258-1265.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

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

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

16. 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.

17. 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.

18. 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.

19. Meignan M, Gallamini A, Meignan M, Gallamini A, Haioun C. Report on the first international workshop on interim-PET-scan in lymphoma. *Leuk Lymphoma*. 2009;50:1257-1260.

20. 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.

21. Johnson P, Federico M, Kirkwood A, et al. Adapted treatment guided by interim PET-CT scan in advanced Hodgkin's lymphoma. *N Engl J Med*. 2016;374:2419-2429.

22. Itti E, Meignan M, Berriolo-Riedinger A, et al. An international confirmatory study of the prognostic value of early PET/CT in diffuse large B-cell lymphoma: Comparison between deauville criteria and DeltaSUVmax. *Eur J Nucl Med Mol Imaging*. 2013;40:1312-1320.

23. Biggi A, Gallamini A, Chauvie S, et al. International validation study for interim PET in ABVD-treated, advanced-stage Hodgkin lymphoma: Interpretation criteria and concordance rate among reviewers. *J Nucl Med*. 2013;54:683-690.

24. Gallamini A, Barrington SF, Biggi A, et al. The predictive role of interim positron emission tomography for Hodgkin lymphoma treatment outcome is confirmed using the interpretation criteria of the deauville five-point scale. *Haematologica*. 2014;99:1107-1113.

25. Trotman J, Luminari S, Boussetta S, et al. Prognostic value of PET-CT after first-line therapy in patients with follicular lymphoma: A pooled analysis of central scan review in three multicentre studies. *Lancet Haematol*. 2014;1:e17-27.

26. Nols N, Mounier N, Bouazza S, et al. Quantitative and qualitative analysis of metabolic response at interim positron emission tomography scan combined with international prognostic index is highly predictive of outcome in diffuse large B-cell lymphoma. *Leuk Lymphoma*. 2014;55:773-780.

27. Kobe C, Goergen H, Baues C, et al. Outcome-based interpretation of early interim PET in advanced-stage Hodgkin lymphoma. *Blood*. 2018;132:2273-2279.

28. Trotman J, Barrington SF, Belada D, et al. Prognostic value of end-of-induction PET response after first-line immunochemotherapy for follicular lymphoma (GALLIUM): Secondary analysis of a randomised, phase 3 trial. *Lancet Oncol*. 2018;19:1530-1542.

29. Mamot C, Klingbiel D, Hitz F, et al. Final results of a prospective evaluation of the predictive value of interim positron emission tomography in patients with diffuse large B-cell lymphoma treated with R-CHOP-14 (SAKK 38/07). *J Clin Oncol*. 2015;33:2523-2529.

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

31. Tylski P, Stute S, Grotus N, et al. Comparative assessment of methods for estimating tumor volume and standardized uptake value in (18)F-FDG PET. *J Nucl Med*. 2010;51:268-276.

32. 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.

33. Casasnovas O, Collin A, Kanoun S, et al. Spleen involvement identified on baseline PET imaging influences outcome of young patients with high risk diffuse large B-cell lymphoma treated with R-CHOP14 but not R-ACVBP. *Hematol Oncol*. 2015;33:211.

34. Voltin CA, Goergen H, Baues C, et al. Value of bone marrow biopsy in Hodgkin lymphoma patients staged by FDG PET: Results from the German Hodgkin study group trials HD16, HD17, and HD18. *Ann Oncol*. 2018;29:1926-1931.

35. Cerci JJ, Gyorke T, Fanti S, et al. Combined PET and biopsy evidence of marrow involvement improves prognostic prediction in diffuse large B-cell lymphoma. *J Nucl Med*. 2014;55:1591-1597.

36. Alzahrani M, El-Galaly TC, Hutchings M, et al. The value of routine bone marrow biopsy in patients with diffuse large B-cell lymphoma staged with PET/CT: A danish-canadian study. *Ann Oncol*. 2016;27:1095-1099.

37. Khan AB, Barrington SF, Mikhaeel NG, et al. PET-CT staging of DLBCL accurately identifies and provides new insight into the clinical significance of bone marrow involvement. *Blood*. 2013;122:61-67.

38. Campbell J, Seymour JF, Matthews J, Wolf M, Stone J, Juneja S. The prognostic impact of bone marrow involvement in patients with diffuse large cell lymphoma varies according to the degree of infiltration and presence of discordant marrow involvement. *Eur J Haematol*. 2006;76:473-480.

39. Sehn LH, Scott DW, Chhanabhai M, et al. Impact of concordant and discordant bone marrow involvement on outcome in diffuse large B-cell lymphoma treated with R-CHOP. *J Clin Oncol*. 2011;29:1452-1457.

40. Luminari S, Biasoli I, Arcaini L, et al. The use of FDG-PET in the initial staging of 142 patients with follicular lymphoma: A retrospective study from the FOLL05 randomized trial of the fondazione italiana linfomi. *Ann Oncol*. 2013;24:2108-2112.

41. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med*. 2009;50 Suppl 1:11S-20S.

42. 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.

43. 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.

44. Song MK, Chung JS, Shin HJ, 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.

45. 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.

46. Kostakoglu L, Martelli M, Sehn LH, et al. Baseline PET-derived metabolic tumor volume metrics predict progression-free and overall survival in DLBCL after first-line treatment: Results from the phase 3 GOYA study. *Blood*. 2017;130:824-824.

47. 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.

48. 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 Suppl 1:122S-50S.

49. Cottereau AS, Lanic H, Mareschal S, et al. Molecular profile and FDG-PET/CT total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-cell lymphoma. *Clin Cancer Res*. 2016;22:3801-3809.

50. 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.

51. 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.

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

53. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J Nucl Med*. 2005;46:1342-1348.

54. Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. *Med Phys*. 2010;37:1309-1324.

55. 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.

56. 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.

57. 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.

58. Kuhnert G, Boellaard R, Sterzer S, et al. Impact of PET/CT image reconstruction methods and liver uptake normalization strategies on quantitative image analysis. *Eur J Nucl Med Mol Imaging*. 2016;43:249-258.

59. Krak NC, Boellaard R, Hoekstra OS, Twisk JW, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Imaging*. 2005;32:294-301.

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

61. 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.

62. Berthon B, Marshall C, Evans M, Spezi E. ATLAAS: An automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. *Phys Med Biol*. 2016;61:4855-4869.

63. Nioche C, Orlhac F, Boughdad S, et al. LIFEx: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res*. 2018;78:4786-4789.

64. Boellaard R. Quantitative oncology molecular analysis suite: ACCURATE. *Journal of Nuclear Medicine*. 2018;59:1753.

65. Cottereau AS, Versari A, Luminari S, et al. Prognostic model for high-tumor-burden follicular lymphoma integrating baseline and end-induction PET: A LYSA/FIL study. *Blood*. 2018;131:2449-2453.

66. 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.

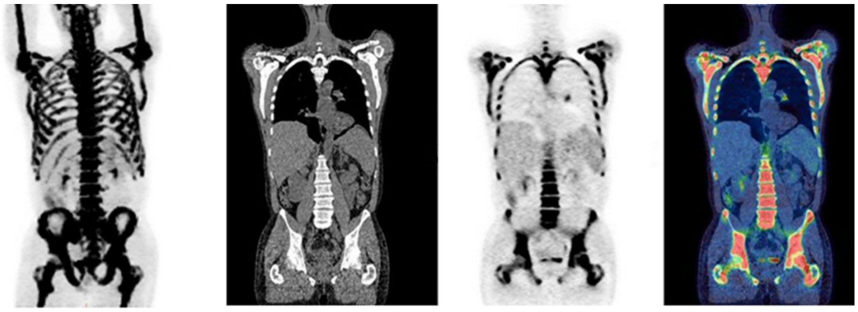

Figure 1: MIP image (left) CT, PET and fused coronal images of a patient with intense abnormal diffuse uptake in the bone marrow in DLBCL and minimal nodal involvement at the left lung hilum. In this case the bone marrow involvement, which was confirmed on bone marrow biopsy would seem appropriate to include in the measurement of MTV.

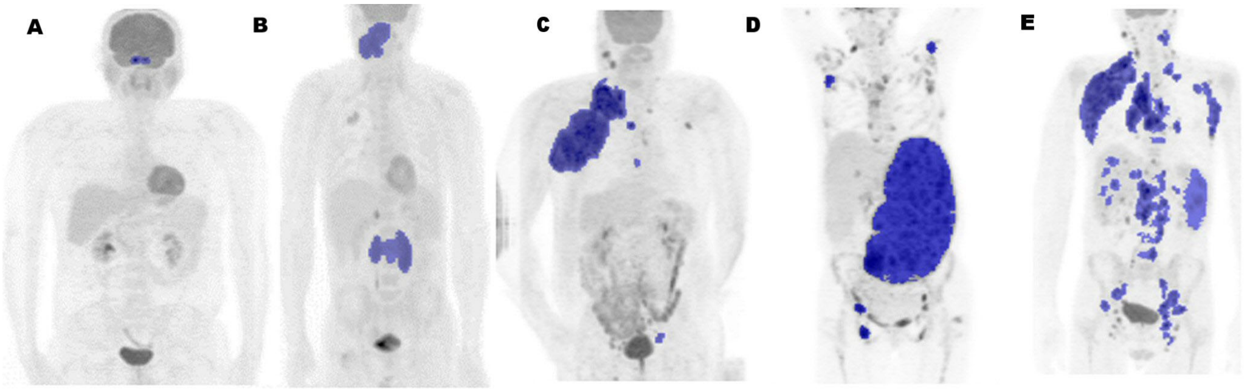

Figure 2: Examples of patients with DLBCL with different size and distribution of tumour.

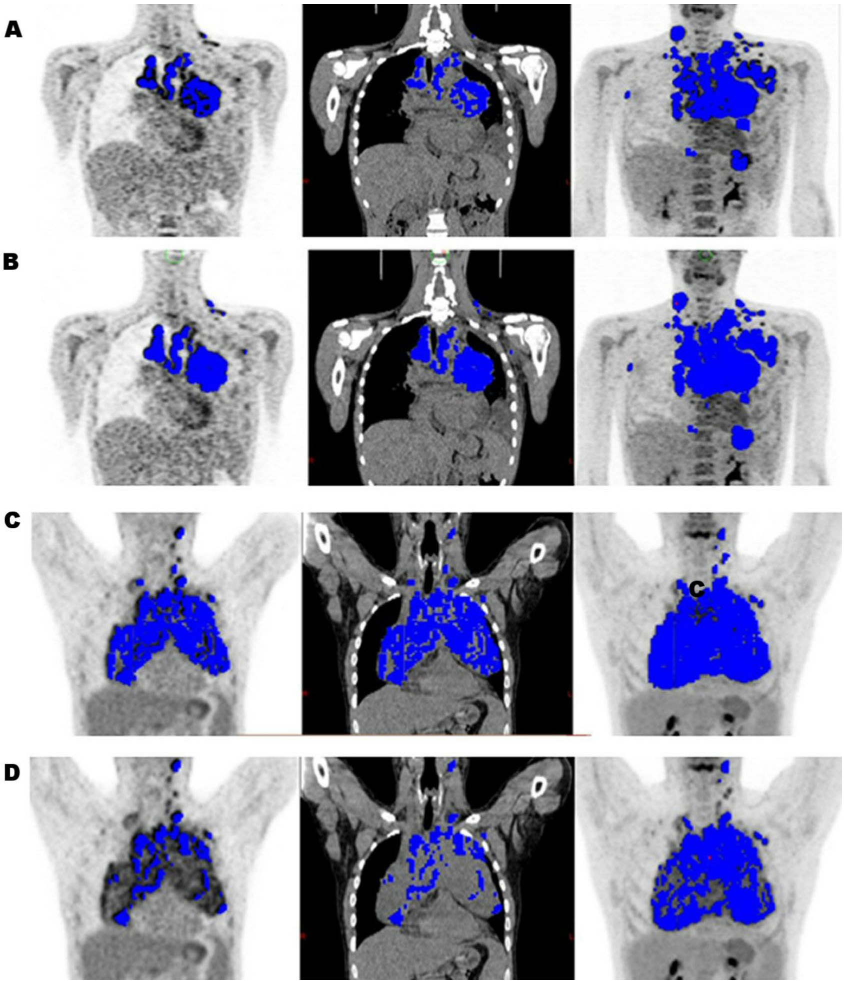

Figure 3: Representative coronal images of PET (left) CT (middle) and MIP images (right) are shown for two patients. Visible tumour is underestimated in one patient using a percentage threshold (A) compared to an absolute threshold (B). In another patient visible tumour is adequately assessed using a percentage threshold (C) but underestimated using an absolute threshold (D). The same absolute threshold (SUV of 4) and percentage threshold (41% of maximum SUV) were used for both patients.

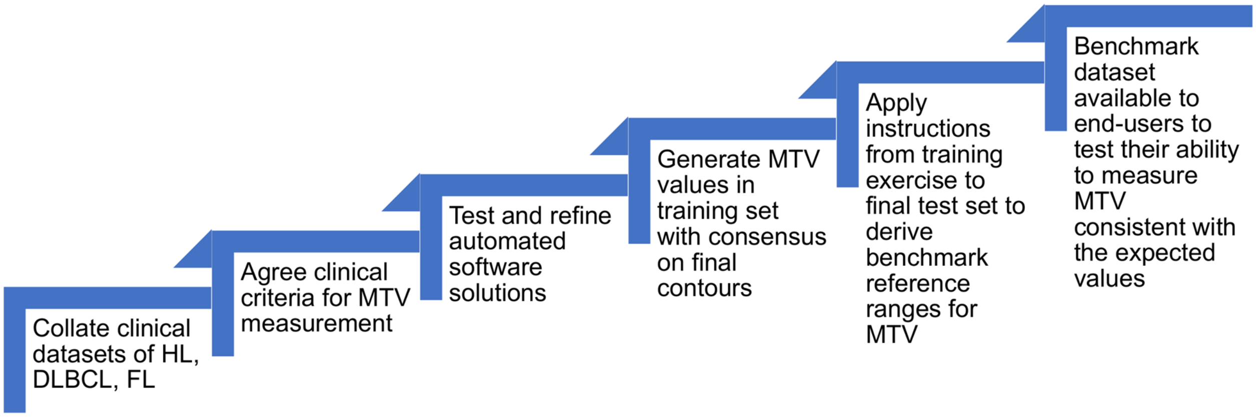
Collate clinical datasets of HL, DLBCL, FL
Agree clinical criteria for MTV measurement
Test and refine automated software solutions
Generate MTV values in training set with consensus on final contours
Apply instructions from training exercise to final test set to derive benchmark reference ranges for MTV
Benchmark dataset available to end-users to test their ability to measure MTV consistent with the expected values

Figure 4: Road map for possible benchmarking exercise

Table 1: Thresholds and study population characteristics contribute to different median values and optimal cut-off points to separate patients into risk groups demonstrated by these reports in diffuse large B cell lymphoma.

	N	PFS and OS	≥ 60y %	Advanced stage %	Bulk	IPI %	PS ≥ 2 %	Threshold	Median (IQR) (cm^3)	Cut-off point (cm^3)
Song 2012(44)	169	At 3y PFS 74 OS 76	60	41	4% ≥ 5cm	26 ≥ 3	25	SUV ≥ 2.5	198 (5–1991)	220
Sasanelli 2014 (51)	114	NA	31	82	36% ≥ 10 cm	65 ≥2 (aaIPI)	30	≥ 41% SUV max	315 (4–2654)	550
Song 2016 (43)	107	NA	67	100	19%	81 ≥ 4 (NCCN-IPI)	16	SUV ≥ 2.5	527 (15–3549)	600
Cottereau 2016 (49)	81	At 5y PFS 60 OS 63	63	80	40% ≥ 10 cm	68 ≥2 (aaIPI)	30	≥ 41% SUV max	320 (106–668)	300
Mikhaeel 2016 (13)	147	At 5y PFS 65 OS 74	48	69	40% ≥ 10 cm	69 ≥2	30	SUV ≥ 2.5	595 (2–7337)	400
Tout 2017 (66)	108	At 4y PFS 76 OS 82	Median age 49	80	NA	60 ≥ 3 (modified IPI)	NA	≥ 41% SUV max	313.5 (NA)	NA

Key: (aa)IPI - (age adjusted) international prognostic index; IQR - interquartile range; max - maximum; N- number; NA – not available ;NCCN - national comprehensive cancer network; OS - overall survival; PS - performance status; PFS - progression free survival; SUV - standardised uptake value