Journal of Nuclear Medicine, published on January 6, 2022 as doi:10.2967/jnumed.121.263067

The impact of semi-automatic segmentation methods on metabolic tumor volume, intensity and dissemination radiomics in ^{18}F-FDG PET scans of patients with classical Hodgkin lymphoma

Julia Driessen[1*], Gerben J.C. Zwezerijnen[2*], Heiko Schöder[3], Esther E.E. Drees[4], Marie José Kersten[1], Alison J. Moskowitz[5], Craig H. Moskowitz[6], Jakoba J. Eertink[7], Henrica C.W. de Vet[8], Otto S. Hoekstra[2], Josée M. Zijlstra[7], Ronald Boellaard[2].
**These authors contributed equally.*

[1]Department of Hematology, Amsterdam UMC, University of Amsterdam, LYMMCARE (Lymphoma and Myeloma Center, Amsterdam), Cancer Center Amsterdam, Amsterdam, The Netherlands; [2]Department of Radiology and Nuclear Medicine, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, Amsterdam, The Netherlands; [3]Department of Molecular Imaging and Therapy Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA; [4]Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, Amsterdam, The Netherlands. [5]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA; [6]Department of Medicine, Sylvester Comprehensive Cancer Center, Miami, Florida, USA; [7]Department of Hematology, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, Amsterdam, The Netherlands. [8]Department of Epidemiology and Data Science, Amsterdam Public Health research institute, Amsterdam, Netherlands.

Short title: PET Segmentation in Hodgkin lymphoma

Funding: This work was financially supported by SHOW (Dutch Foundation of hemato-oncological research).

Statement of prior presentation: Not applicable

First author: Julia Driessen, BSc
In training: Medical student and PhD candidate
j.driessen@amsterdamumc.nl
ORCID: 0000-0001-9364-2501
Department of Hematology
Amsterdam UMC, University of Amsterdam
Meibergdreef 9
1105AZ Amsterdam, Netherlands
Phone: +31615831843

Corresponding author: Ronald Boellaard, MD, PhD
r.boellaard@amsterdamumc.nl
ORCID: 0000-0002-0313-5686
Department of Radiology and Nuclear Medicine
Amsterdam UMC, Vrije Universiteit Amsterdam,
De Boelelaan 1117
1081HV Amsterdam, Netherlands
Phone: +31(0)2044449638

ABSTRACT

Introduction: Consensus about a standard segmentation method to derive metabolic tumor volume (MTV) in classical Hodgkin lymphoma (cHL) is lacking, and it is unknown how different segmentation methods influence quantitative PET features. Therefore, we aimed to evaluate the delineation and completeness of lesion selection and the need for manual adaptation with different segmentation methods, and to assess the influence of segmentation methods on the prognostic value of MTV, intensity and dissemination radiomics features in cHL patients. **Methods:** We analyzed a total of 105 ^{18}F-FDG PET-CT scans from patients with newly diagnosed (n=35) and relapsed/refractory (n=70) cHL with six segmentation methods: two fixed thresholds on SUV4.0 and SUV2.5, two relative methods of 41% of SUVmax (41max), and a contrast-corrected 50% of SUVpeak (A50P) and two combination 'majority vote' methods (MV2, MV3). Segmentation quality was assessed by two reviewers based on pre-defined quality criteria: completeness of selection, the need for manual adaptation and delineation of lesion borders. Correlations and prognostic performance of resulting radiomics features were compared among the methods. **Results:** SUV4.0 required the least manual adaptation but tended to underestimate MTV and often missed small lesions with low FDG uptake. SUV2.5 most frequently included all lesions but required minor manual adaptations and generally overestimated MTV. In contrast, few lesions were missed when using 41max, A50P, MV2 and MV3, but these segmentation methods required extensive manual adaptation and overestimated MTV in most cases. MTV and dissemination features significantly differed among the methods. However, correlations among methods were high for MTV and most intensity and dissemination features. There were no significant differences in prognostic performance for all features among the methods. **Conclusions:** There is a high correlation between MTV, intensity and most dissemination features derived with the different segmentation methods and the prognostic performance is similar. Despite frequently missing small lesions with low

FDG avidity, segmentation with a fixed threshold of SUV4.0 required the least manual adaptation, which is critical for future research and implementation in clinical practice. However, the importance of small, low-avid lesions should be addressed in a larger cohort of cHL patients.

Keywords: Hodgkin lymphoma, Segmentation methods, ^{18}F-FDG PET-CT, Outcome prediction, Radiomics

INTRODUCTION

The ^{18}F-fluorodeoxyglucose (FDG) - positron emission tomography (PET) – computed tomography (CT) scan is standard of care for staging and response evaluation in the treatment of classical Hodgkin lymphoma (cHL) *(1).* Optimizing baseline risk-stratification contributes to implement individualized treatment strategies aiming to lower toxicity in patients with favorable prognostic characteristics, and identifying patients with unfavorable prognostic characteristics early for treatment with other therapies *(2-4).* The use of quantitative PET features to improve risk stratification could be implemented in clinical practice if workflows are optimized.

Several studies have shown that metabolic tumor volume (MTV) is a potential prognostic marker in newly diagnosed (ND) and relapsed/refractory (R/R)-cHL *(4-11).* However, there are different methods for assessing MTV and there is no consensus which method performs best in cHL patients in terms of prognostic performance, ease of use and interobserver variability *(12)*. MTV assessment is especially challenging in disseminated diseases such as lymphoma. cHL is a heterogeneous disease that is typically localized in the mediastinal and para-aortic regions, mainly affecting young patients who frequently show high physiological FDG uptake in brown fat and muscles *(1)*. These regions with high physiological FDG uptake impede accurate delineation of tumor lesions nearby. Therefore, it is important to evaluate different segmentation methods specifically for cHL.

Although manual segmentation is the current standard for determining MTV, it is very time-consuming and prone to interobserver variability *(12)*. Semi-automatic segmentation includes algorithms that select regions with high FDG uptake above the threshold of a certain standard uptake value (SUV). Segmentation of the MTV can be performed by either pre-defining regions of interest in which lesions will be automatically selected, or by starting with automatic segmentation and deleting regions with high physiological FDG uptake (e.g. brain, liver, kidneys) thereafter. Although the segmentation method applied can significantly impact

the MTV, it is unknown how each method affects other quantitative PET-radiomics features, such as patient-level dissemination parameters *(13-17)*. Besides, no comparative studies have been performed that address representativeness of the segmented MTV with the visual interpretation of the MTV in cHL patients.

The aim of our research was to evaluate the delineation and completeness of lesion selection, and the need for manual adaptation with six different semi-automatic segmentation methods, and to assess the influence of the segmentation method on the prognostic value of MTV, intensity and dissemination radiomics features in scans of cHL patients.

MATERIALS AND METHODS

Study Population

PET-CT scans from ND-cHL patients were collected from study cohorts of the Amsterdam UMC (n=35)*(2,18)*. PET-CT scans of patients with RR-cHL were collected from three clinical trials conducted in Amsterdam UMC, the Netherlands (n=47) and Memorial Sloan Kettering Cancer Center, NY, USA (n=23)*(2-4)*. All patients had biopsy-proven cHL and the PET-CT scan was performed before start of therapy. All patients provided written informed consent for participation in the clinical trials (NCT02280993, NCT00255723, NCT01508312) or biobank cohort (*18*) of which the study protocols were approved by Institutional Review Boards and Ethics Committees of the centers that conducted the trials. For secondary use of data for this analysis a waiver was obtained from the Ethics Committee.

^{18}F-FDG PET-CT Scans and Quality Control

The PET-CT systems used to perform the scans were EARL (Europe) or ACR (USA) accredited *(19)*. PET-CT scans were de-identified at the participating centers and centrally collected. PET scans that did not meet the following four criteria, described by EANM guidelines, were excluded from analysis: 1) plasma glucose <11mmol/L; 2) reconstruction of attenuation corrected PET according to guidelines described by EARL or ACR; 3) total image activity (MBq) between 50-80% of the total injected FDG activity or liver SUVmean between 1.3-3.0; and 4) essential PET acquisition data and clinical data available *(19)*.

Segmentation of the Volume of Interest

Attenuation-corrected PET scans were analyzed using the ACCURATE tool *(20)*. Six different semi-automatic methods were used for each scan to select the Volume of Interest (VOI): two fixed thresholds of SUV4.0 and SUV2.5, two relative thresholds of 41% of SUVmax

(41max) and a contrast corrected 50% of SUVpeak (A50P), and two 'majority vote' (MV) methods selecting voxels that are selected with ≥2 (MV2) and ≥3 (MV3) of the previously mentioned fixed or relative methods, respectively. The VOI was delineated by automatic preselection of FDG avid structures using the six different segmentation methods and a volume threshold of ≥3mL. Non-tumor regions were deleted and lymphoma lesions <3mL were added with single mouse clicks. If tumor regions were adjacent to non-tumor FDG avid regions (e.g. heart, liver, bladder), non-tumor regions were either removed manually, or tumor segmentation was restricted by placing a border or mask, which prevented selection of lesions outside the border (Figure 1A). Only focal extranodal and splenic lesions were included in the VOI. A global increase in FDG uptake of the spleen or bone marrow, was not included in the VOI. Delineations were performed by JD under supervision of a nuclear medicine physician (GJCZ or HS).

Quality Scores of Representativeness of Segmentations Compared to Visual Judgement

The quality of the segmentation by the six different methods was assessed using three 'quality score' (QS) criteria (Table 1):

1) completeness of selection of the VOI (i.e. were all tumor-lesions selected);

2) requirement of manual adaptation after semi-automatic segmentation (i.e. manual removal of non-tumor regions);

3) delineation quality of the VOI (i.e. does the VOI border reflect the visual interpretation of the FDG avid tumor area on the PET scan).

Two reviewers (JD and GJCZ or HS) performed the QS assessment for each of the six segmentations for all scans, blinded for patient outcome. Completeness of selection and delineation QS were assessed independently, followed by a consensus meeting in which the reviewers reached a consensus on all discrepancy scores and assigned a final QS to each segmentation. The manual adaptation QS was assessed in consensus between the reviewers

during review of the segmentation of scans. An example of the QS assessment by the six segmentation methods is included in Figure 1B.

Radiomics Feature Extraction

RaCat software was used to extract 18 patient-level dissemination features from the complete MTV at patient level *(21)*. Dissemination features included several novel features addressing inter-lesional heterogeneity based on distance, volume, SUVmax and SUVpeak (i.e. the 1mL with the highest SUV within the VOI). In addition, MTV, SUVmax, SUVpeak, SUVmean and total lesion glycolysis (TLG) were extracted from the VOI. An overview of all features and its definitions are provided in Supplemental Table 1.

Statistical Analysis

QS of segmentations were analyzed descriptively and compared using chi-square tests for the whole cohort and separately for ND-cHL and RR-cHL patients. MTV, intensity and dissemination radiomics features were compared between the ND-cHL and RR-cHL cohorts using Wilcoxon rank sum test for non-parametric data. Further analyses were performed on the whole cohort. Correlations of MTV, intensity and dissemination radiomics features among the six different segmentation methods were assessed using Spearman's rank coefficients correlation. Receiver-operating characteristics (ROC) analysis was used to calculate the area under the curve (AUC) for each feature per segmentation method on the whole cohort. An event was defined as the occurrence of progressive disease within 3 years and patients who died without progression were excluded. AUC curves were compared using a paired t-test as described by DeLong *et al*.*(22)*.

Statistical analysis was performed using R software version 4.0.3. A *P*-value of <0.05 was considered statistically significant.

RESULTS

Patient Characteristics

A total of 105 PET-CT scans of patients with ND-cHL (n=35) and RR-cHL (n=70) were included in the analysis (Supplemental Table 2). A comparison of radiomics features between ND-cHL and RR-cHL showed no significant differences for most features, except for MTV, SUVpeak and Dvol (i.e. the maximum difference in volume between lesions), which were all higher in ND patients compared to RR patients (Supplemental Table 3).

Quality Scores of Segmentations

Agreement of QS assessment between the two reviewers was high (91% for segmentation quality and 82% for delineation quality).

Segmentation resulted in complete selection of all lesions in the majority of cases (Figure 2A; Supplemental Table 4). SUV2.5 showed the highest rate of complete selection, followed by 41max, MV2, A50P and MV3, while SUV4.0 frequently missed minor (59%) and major (10%) lesions. Using the SUV4.0 method, 91% of scans could be segmented without any manual adaptation (Figure 2B). The SUV2.5 method required minor adaptations in 37% of scans and 7% major adaptations. Using the 41max and MV2 methods, only 30% and 34% of scans could be segmented without manual adaptation, and in 47% and 33% of cases, major manual adaptations were required, respectively. Using A50P and MV3, about 50% of scans did not require manual adaptation. None of the methods resulted in a high percentage of representative delineation of tumor borders (Figure 2C). SUV4.0, SUV2.5 and MV3 resulted in representative delineation in about 50% of cases, while SUV4.0 tended to underestimate the MTV and SUV2.5 and MV3 tended to overestimate the MTV in the remaining cases. The 41max, A50P and MV2 methods resulted in representative delineation in less than 30% and usually overestimated the MTV.

No significant differences were observed for QS between ND and RR patients, except for completeness of selection in which complete selection rates were higher in RR patients compared to ND patients with 41max, A50P or MV3 (Supplemental Figure 1).

Comparison of Features

MTV differed significantly among the segmentation methods. The median MTV per method ranged between 44-143 mL (Figure 3; Supplemental Table 5). SUV4.0 resulted in significant lower MTV compared to all other segmentation methods ($p<0.001$). The number of lesions was significantly lower with 41max and MV2 compared to SUV4.0 and SUV2.5 segmentation methods ($p<0.05$). Dmax (i.e. the maximum distance between two lesions) was not significantly different among the segmentation methods.

MTV, the number of lesions and Dmax showed high correlations among most methods (Figure 4; Supplemental Table 6). For MTV and the number of lesions, the highest correlations were observed between the two fixed methods (SUV4.0 and SUV2.5), and between the relative and majority vote methods, with lower correlations between the fixed and relative or majority vote methods. SUVmax and SUVpeak had identical median values and were strongly correlated (R=1) across all methods. Dissemination features addressing differences in volume or SUVpeak among lesions showed lower correlations between SUV4.0 and the other five segmentation methods (Supplemental Table 6).

To assess the effect of incomplete selection of lesions, several features derived with SUV4.0 were plotted against SUV2.5 (Supplemental Figure 2). Scans that missed major lesions with SUV4.0 did not show large deviations in the correlation between SUV4.0 and SUV2.5 when compared with scans that had complete selection or missed only minor lesions.

Prognostic Performance per Method

Except for MV2, the AUC of the ROC did not differ significantly among the segmentation methods for all features assessed (Figure 5; Supplemental Table 7). The highest AUC's were observed for MTV (range 0.62-0.65), TLG (0.63-0.65), number of lesions (0.55-0.63), Spread in volume (VolSpread)(0.58-0.65) and the difference in SUVpeak between the hottest lesion and all other lesions (DSUVpeakSumHot)(0.56-0.63). Of all methods MV2 showed the lowest AUC for the various features (median AUC of all variables: 0.55). The other five methods showed comparable median AUC's with the highest median AUC of all variables of 0.62 for SUV4.0.

DISCUSSION

MTV has shown prognostic value in cHL, but the use of different segmentation methods hampers direct comparisons between studies *(4-10)*. This is especially true if a cutoff for MTV is used to divide patients in low- and high-risk groups, since absolute MTV values significantly differ between methods. Harmonization of MTV assessment enables evaluating MTV as prognostic marker in cHL in multi-cohort setting. The same holds for other quantitative PET-features including dissemination features.

We evaluated the completeness of lesion selection, need for manual adaptations and delineation quality of six semi-automatic segmentation methods to assess MTV and dissemination features in 105 cHL patients. Segmentation with SUV4.0 required the least manual adaptations because this method, in contrast to other methods, rarely floods into regions with high physiological FDG uptake. SUV2.5 often required minor adaptations, but seldomly major adaptations. Although segmentation using SUV4.0 frequently did not include all lesions (missing those with a SUV<4.0), these lesions were often small and scans with major lesions missing did not cause significant deviations in the correlation between SUV4.0 and SUV2.5, which was the most complete method. Additionally, the prognostic performance between all methods was similar, and SUV4.0 and SUV2.5 showed the highest AUCs for most variables.

This suggests that small lesions with low SUV uptake, that are frequently not included with SUV4.0, probably do not contain critical prognostic information. This could be partly explained by the low contribution to total MTV of small lesions. However, small lesions could still influence dissemination features, of which the prognostic value needs to be established in a larger set of patients with more progression events. Additionally, small low-uptake lesions are potentially of higher importance in the response-assessment situation, thus SUV4.0 may be less suitable for quantitative interim PET analyses in cHL *(1)*.

All segmentation methods, except SUV4.0, frequently overestimated the MTV assessed by visual interpretation. This may be less relevant when using only patient-level features, as correlations

among methods are high, however, lesion-based radiomics analysis involving texture features may be adversely affected by over-segmentation, i.e. by selection of voxels that are not part of the tumor *(23)*. Methods that tended to overestimate the MTV also showed a lower number of lesions, as lesions close to each other were frequently clustered into one lesion, as illustrated in Figure 1. This explains the discrepancy that SUV4.0 often misses small or low uptake lesions, but still shows the highest number of lesions (Figure 3).

In a recent comparison of six segmentation methods in diffuse large B-cell lymphoma (DLBCL), a fixed threshold of SUV4.0 was considered the best method to derive MTV *(24)*. Similar to our findings, MTV significantly differed among the methods but the prognostic performance was comparable. Interestingly, method performance in DLBCL at interim PET has been shown to depend on the lesional SUVmax, in which lesions with SUVmax<10 were delineated most successfully using MV3, while SUV4.0 was most successful in lesions with SUVmax>10 *(25)*. Correlations for MTV were significantly higher in our cohort than previously described for DLBCL, which may be explained by the fact that our correlations were assessed following manual adaptation *(24,25)*. Additionally, and contrary to our findings, the 41max, A50P, and MV3 methods yielded lower exact MTV values than SUV4.0 in baseline DLBCL. This shows that performance of different methods can be disease-dependent. In our cohort, 41max resulted in the highest MTV, which can be explained by the lower SUV in our cHL cohort (median SUVmax 11.3), compared to DLBCL patients (median SUVmax 22.6) (*26*). Since SUVmax is a patient-level feature, and cHL shows heterogeneous FDG uptake, other lesions within a patient may have a much lower SUVmax, resulting in overestimation of the MTV and flooding with relative methods such as 41max.

Methods based on relative thresholds (e.g. 41max and A50P) are less suitable for assessing MTV in diseases with heterogeneous FDG uptake, such as cHL, because a high lesional SUVmax may exclude the lower avid voxels of the lesion causing under-segmentation, whereas a low lesional SUVmax results in a low threshold leading to ‘flooding’ into regions with

physiological FDG uptake. The ‘majority vote’ methods could not overcome this disadvantage of the relative methods. MV2 frequently uses voxels that are being selected with 41max and A50P, and while MV3 needs a third method this did not result in better segmentation compared to methods with a fixed threshold.

Although the 41max method is recommended for MTV segmentation and has been used in several lymphoma studies, this method requires extensive manual adaptation, which is time-consuming and more susceptible to inter-observer variation *(13,15,19)*. Additionally, the recommendation for 41max is based on solid malignancies rather than disseminated diseases such as cHL, and 41max has not been compared directly to a fixed threshold of SUV4.0 *(27-29).* Therefore this recommendation should be reconsidered for cHL.

CONCLUSION

For PET-CT segmentation in cHL, we showed a high correlation among MTV and most intensity and dissemination features derived with different segmentation methods, except for dissemination features addressing differences in volume and SUVmax/peak. The prognostic performance of all features is comparable among the methods. The SUV4.0 method required the least manual adaptation, which is critical for future research and implementation in clinical practice. Although segmentation with SUV4.0 often missed small low avid lesions, which may in particular affect dissemination features such as the Dmax, this seemed not to influence the prognostic performance of most features, including Dmax. However, to be conclusive about recommending SUV4.0 for cHL segmentation, the prognostic importance of small lesions with low uptake should be evaluated in a larger cohort of cHL patients with more progression events.

DISCLOSURE

This work was financially supported by SHOW (a non-profit donation fund of Amsterdam UMC). There is no financial support for this work that could have influenced the outcomes described in the manuscript. However, particular authors report a potential conflict of interest:

RB: scientific advisor and chair of the EARL accreditation program.

MJK: Consultancy: BMS/Celgene, Kite/Gilead, Miltenyi Biotech, Novartis, Takeda. Honoraria: Kite/Gilead, Novartis, Roche. Research funding: Kite/Gilead, Takeda.

CHM: Advisor and research funding: Celgene, Genentech, Merck, Seattle Genetics.

AJM: Consultancy: Takeda, Imbrium Therapeutics, Janpix, Merck, Seatle Genetics. Research funding: Incyte, Merck, Seattle Genetics, ADC Therapeutics, Beigene, Miragen, Bristol-Myers Squibb.

JMZ: Research funding: Takeda.

All remaining authors have declared no conflict of interest.

ACKNOWLEDGEMENTS

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

AUTHORSHIP CONTRIBUTIONS

JD, GJCZ, RB and JMZ designed the study. JD, EEED, MJK, JMZ, HS, AJM and CHM collected the data. JD and GJCZ performed the MTV segmentation analysis. GJCZ and HS reviewed the PET scans and supervised the segmentation analysis. JD performed the statistical analysis. JD drafted the manuscript with contributions from all authors. All authors interpreted the data, read, commented on, and approved the final version of the manuscript.

KEY POINTS

QUESTIONS: which segmentation method provides the best delineation and completeness of lesion selection with the least manual adaptation in scans of classical Hodgkin lymphoma (cHL) patients, and what is the influence of the segmentation method on the prognostic value of MTV, intensity and dissemination radiomics features?

PERTINENT FINDINGS: 1) Segmentation with a fixed threshold of SUV4.0 required the least manual adaptation, with SUV2.5 resul ting in the most complete selection of all lesions. 2) The prognostic performance of features was comparable per segmentation method, and there was a high correlation for MTV and intensity features, but not for all dissemination features, assessed with the different methods.

IMPLICATIONS FOR PATIENT CARE: semi-automated estimation of MTV, intensity and dissemination radiomics features of cHL patients is feasible using a method with a fixed threshold.

REFERENCES

1. Cheson BD, Fisher RI, Barrington SF, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. *J Clin Oncol.* 2014;32:3059-3068.

2. Kersten MJ, Driessen J, Zijlstra JM, et al. Combining brentuximab vedotin with dexamethasone, high-dose cytarabine and cisplatin as salvage treatment in relapsed or refractory Hodgkin lymphoma: the phase II HOVON/LLPC Transplant BRaVE study. *Haematologica.* 2021;106:1129-1137.

3. Moskowitz CH, Matasar MJ, Zelenetz AD, et al. Normalization of pre-ASCT, FDG-PET imaging with second-line, non-cross-resistant, chemotherapy programs improves event-free survival in patients with Hodgkin lymphoma. *Blood.* 2012;119:1665-1670.

4. 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.

5. Albano D, Mazzoletti A, Spallino M, et al. Prognostic role of baseline 18F-FDG PET/CT metabolic parameters in elderly HL: a two-center experience in 123 patients. *Ann Hematol.* 2020;99:1321-1330.

6. Milgrom SA, Elhalawani H, Lee J, et al. A PET Radiomics Model to Predict Refractory Mediastinal Hodgkin Lymphoma. *Sci Rep.* 2019;9:1322.

7. Rogasch JMM, Hundsdoerfer P, Hofheinz F, et al. Pretherapeutic FDG-PET total metabolic tumor volume predicts response to induction therapy in pediatric Hodgkin's lymphoma. *BMC Cancer.* 2018;18:521.

8. 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.

9. Procházka V, Gawande RS, Cayci Z, et al. Positron Emission Tomography-Based Assessment of Metabolic Tumor Volume Predicts Survival after Autologous Hematopoietic Cell Transplantation for Hodgkin Lymphoma. *Biol Blood Marrow Transplant.* 2018;24:64-70.

10. Song MK, Chung JS, Lee JJ, et al. Metabolic tumor volume by positron emission tomography/computed tomography as a clinical parameter to determine therapeutic modality for early stage Hodgkin's lymphoma. *Cancer Sci.* 2013;104:1656-1661.

11. Mettler J, Muller H, Voltin CA, et al. Metabolic Tumour Volume for Response Prediction in Advanced-Stage Hodgkin Lymphoma. *J Nucl Med.* 2018;60:207-211.

12. 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.

13. Tutino F, Puccini G, Linguanti F, et al. Baseline metabolic tumor volume calculation using different SUV thresholding methods in Hodgkin lymphoma patients: interobserver agreement and reproducibility across software platforms. *Nucl Med Commun.* 2021;42:284-291.

14. Martín-Saladich Q, Reynés-Llompart G, Sabaté-Llobera A, Palomar-Muñoz A, Domingo-Domènech E, Cortés-Romera M. Comparison of different automatic methods for the delineation of the total metabolic tumor volume in I-II stage Hodgkin Lymphoma. *Sci Rep.* 2020;10:12590.

15. Camacho MR, Etchebehere E, Tardelli N, et al. Validation of a Multifocal Segmentation Method for Measuring Metabolic Tumor Volume in Hodgkin Lymphoma. *J Nucl Med Technol.* 2020;48:30-35.

16. 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.

17. Weisman AJ, Kim J, Lee I, et al. Automated quantification of baseline imaging PET metrics on FDG PET/CT images of pediatric Hodgkin lymphoma patients. *EJNMMI Phys.* 2020;7:76.

18. Drees EEE, Roemer MGM, Groenewegen NJ, et al. Extracellular vesicle miRNA predict FDG-PET status in patients with classical Hodgkin Lymphoma. *J Extracell Vesicles.* 2021;10:e12121.

19. 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.

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

21. Pfaehler E, Zwanenburg A, de Jong JR, Boellaard R. RaCaT: An open source and easy to use radiomics calculator tool. *PLoS One.* 2019;14:e0212223.

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837-845.

23. Pfaehler E, Beukinga RJ, de Jong JR, et al. Repeatability of (18) F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med Phys.* 2019;46:665-678.

24. Barrington SF, Zwezerijnen B, de Vet HCW, et al. Automated Segmentation of Baseline Metabolic Total Tumor Burden in Diffuse Large B-Cell Lymphoma: Which Method Is Most Successful? A Study on Behalf of the PETRA Consortium. *J Nucl Med.* 2021;62:332-337.

25. Zwezerijnen GJ, Eertink JJ, Burggraaff CN, et al. Interobserver agreement in automated metabolic tumor volume measurements of Deauville score 4 and 5 lesions at interim (18)F-FDG PET in DLBCL. *J Nucl Med.* 2021;62:1531-1536.

26. Eertink JJ, van de Brug T, Wiegers SE, et al. 18F-FDG PET baseline radiomics features improve the prediction of treatment outcome in diffuse large B-cell lymphoma. *European Journal of Nuclear Medicine and Molecular Imaging.* 2021;Epub ahead of print.

27. Frings V, de Langen AJ, Smit EF, et al. Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. *J Nucl Med.* 2010;51:1870-1877.

28. 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.

29. 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.

Tables

TABLE 1: Definitions of quality scores for visual assessment of segmentation quality

Quality score	Level	Definition
Completeness of selection	Complete	All visible tumor lesions are selected
	Missing minor lesions	Missing lesions are <3mL and within the selected VOI region (e.g. considered not to influence the Dmax)
	Missing major lesions	Lesions are missing that are either ≥3mL or outside of the selected VOI region (e.g. considered to influence the Dmax)
Manual adaptation	No adaptation	No manual adaptation is required. Adding lesions with single mouse clicks is not considered manual adaptation
	Minor adaptation	Manual adaptation is required in order to obtain a representative selection of the VOI by removing max 1 non-tumor region
	Major adaptation	Extensive manual adaptation is required by removing >1 non-tumor region
Delineation	Representive	Delineation of VOI borders is representative of the visual interpretation of the tumor
	Underestimation	Delineation of VOI borders is underestimated
	Overestimation	Delineation of VOI borders is overestimated

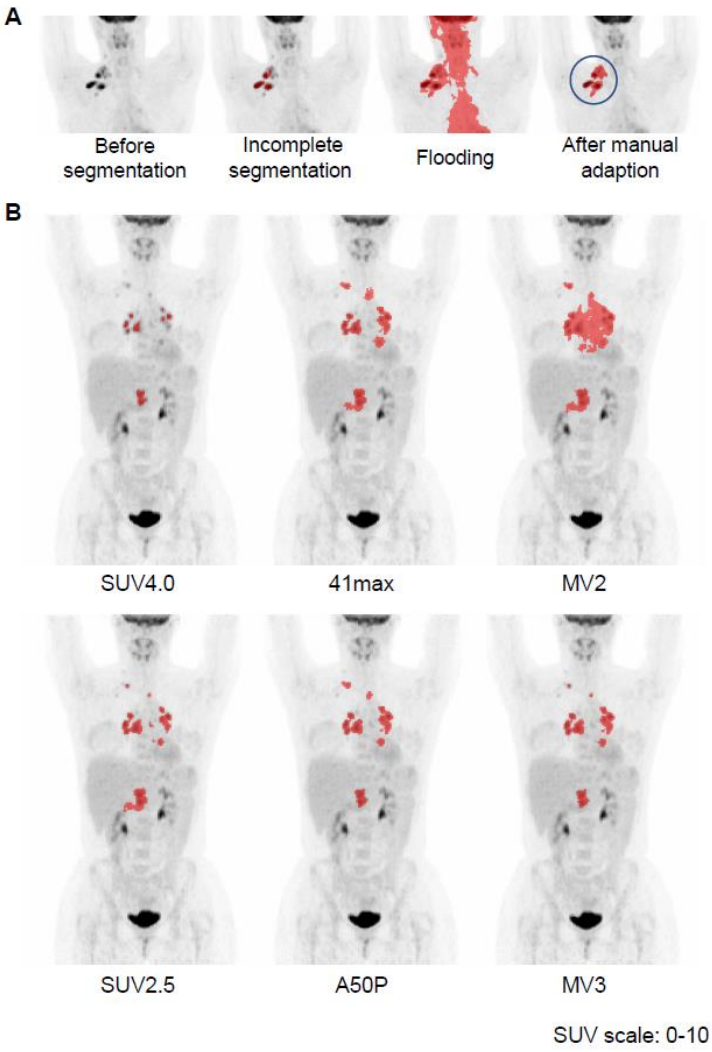


FIGURE 1.

Examples of semi-automatic segmentation. (A) Minimal intensity projection (MIP) of the PET scan before segmentation; automatic selection with the 41max method missed multiple lesions; adding missing lesions resulted in flooding into the heart, tonsils and brain; manual adaptation by placing a border around the volume of interest before segmentation resulted in complete selection. (B) Segmentation with SUV4.0 was scored as 'missing minor lesions' and 'representative delineation'. Segmentation with SUV2.5, 41max, A50P, MV2 and MV3 were scored as 'complete segmentation' with 'overestimation of delineation'. Segmentation with 41max flooded into the heart and required minor manual adaptation. Segmentation with MV2 flooded into the heart and liver and required major manual adaptations.

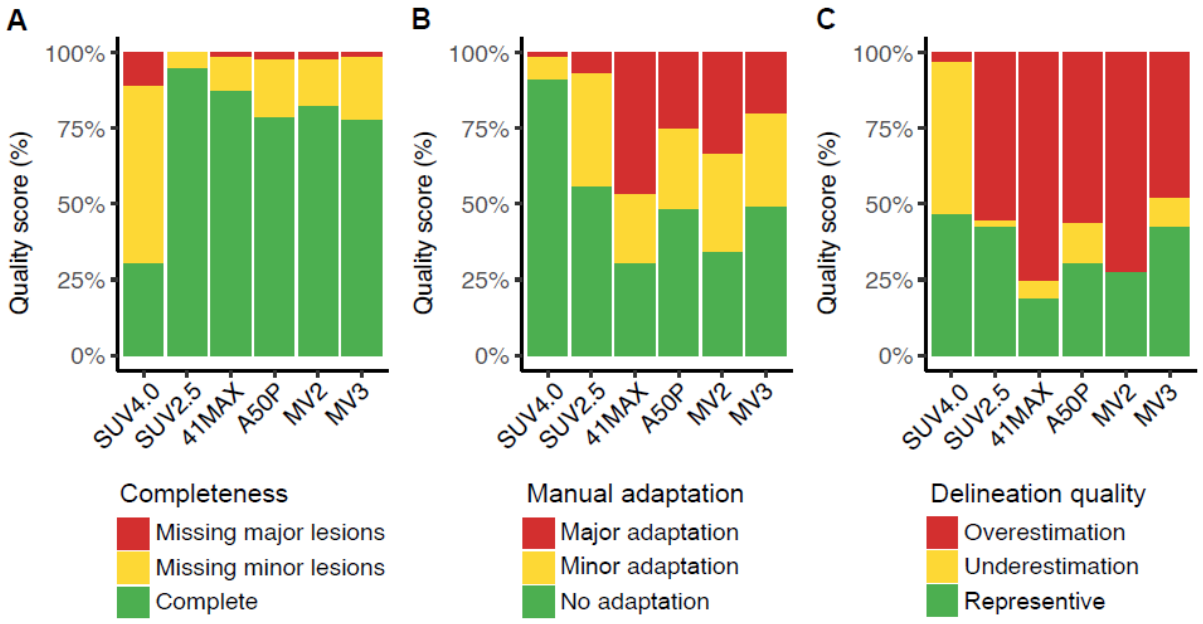


FIGURE 2.

Quality scores (QS) of segmentation methods. (A) Completeness of selection. (B) Manual adaptations required for representive segmentation. (C) Delineation of tumor borders.

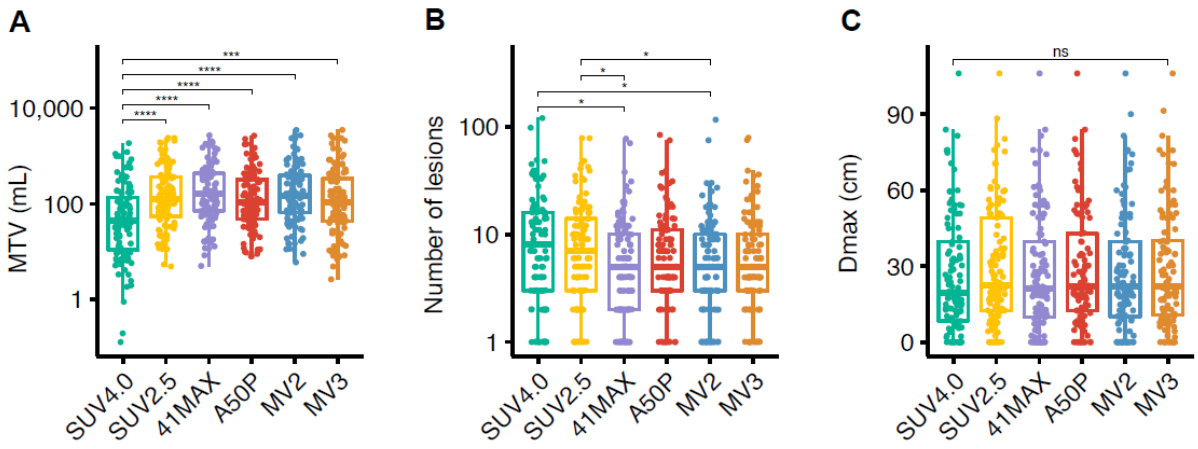


FIGURE 3.

Radiomics features derived with six different semi-automatic segmentation methods. (A) Metabolic tumor volume (MTV) in mL. (B) Number of lesions. (C) Maximum distance (Dmax) in cm. * p<0.05; ** p<0.01; *** p<0.001; ****p<0.0001; ns, not significant.

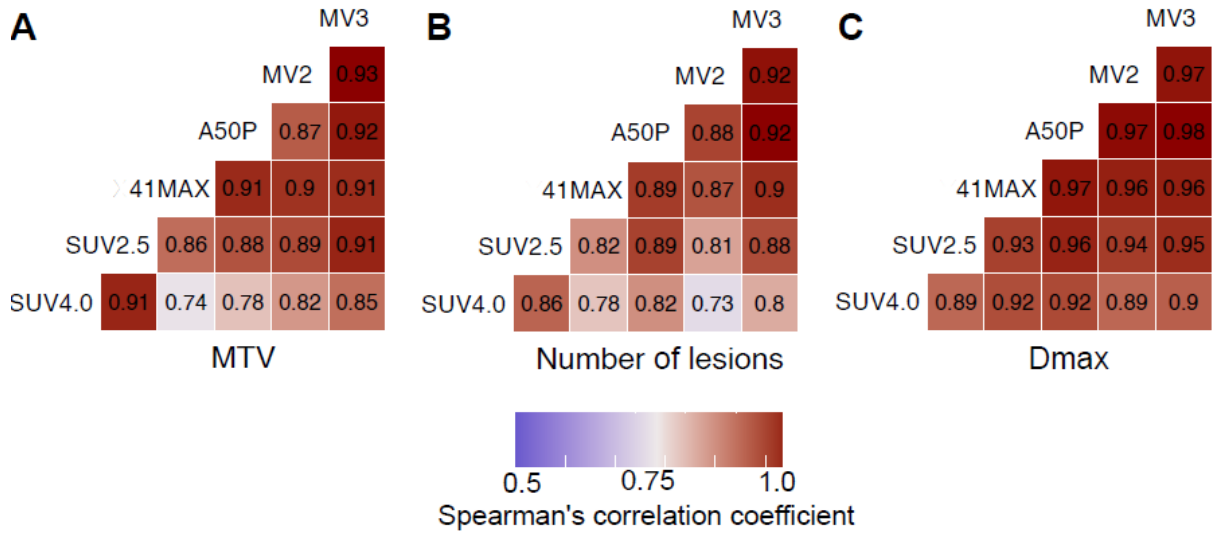


FIGURE 4.

Spearman's rank correlation coefficients for radiomics features among different segmentation methods. (A) Metabolic tumor volume (MTV). (B) Number of lesions. (C) Maximum distance (Dmax). All correlations assessed had a *P*-value of <0.01.

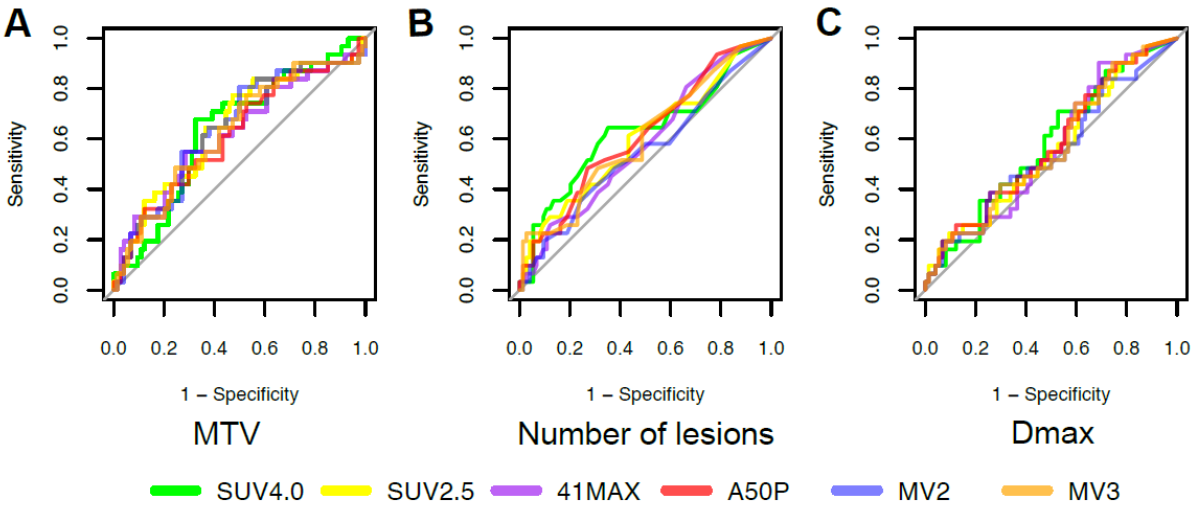
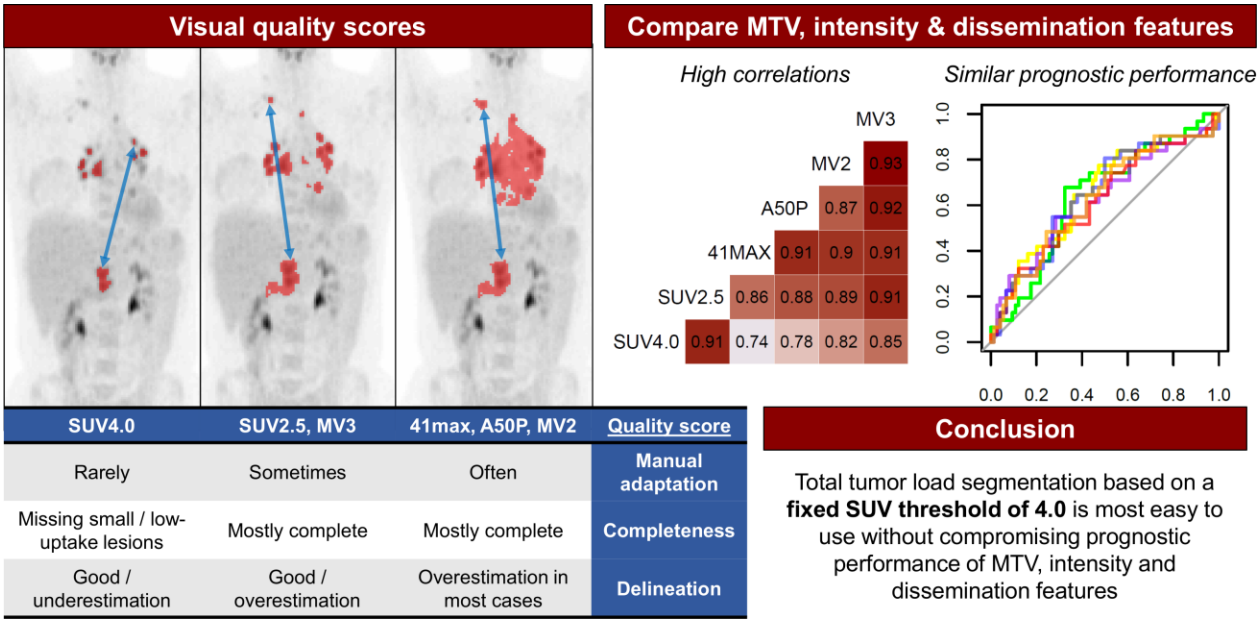


FIGURE 5.

Prognostic performance of radiomics features per method assessed by Area Under the Curve of Receiver Operating Characteristics analysis. (A) Metabolic tumor volume (MTV). (B) Number of lesions. (C) Maximum distance (Dmax).



Graphical Abstract

SUPPLEMENTAL MATERIAL

The impact of semi-automatic segmentation methods on metabolic tumor volume, intensity and dissemination radiomics in ^{18}F-FDG PET scans of patients with classical Hodgkin lymphoma

Julia Driessen[1*], Gerben J.C. Zwezerijnen[2*], Heiko Schöder[3], Esther E.E. Drees[4], Marie José Kersten[1], Alison J. Moskowitz[5], Craig H. Moskowitz[6], Jakoba J. Eertink[7], Henrica C.W. de Vet[8], Otto S. Hoekstra[2], Josée M. Zijlstra[7], Ronald Boellaard[2].

**These authors contributed equally.*

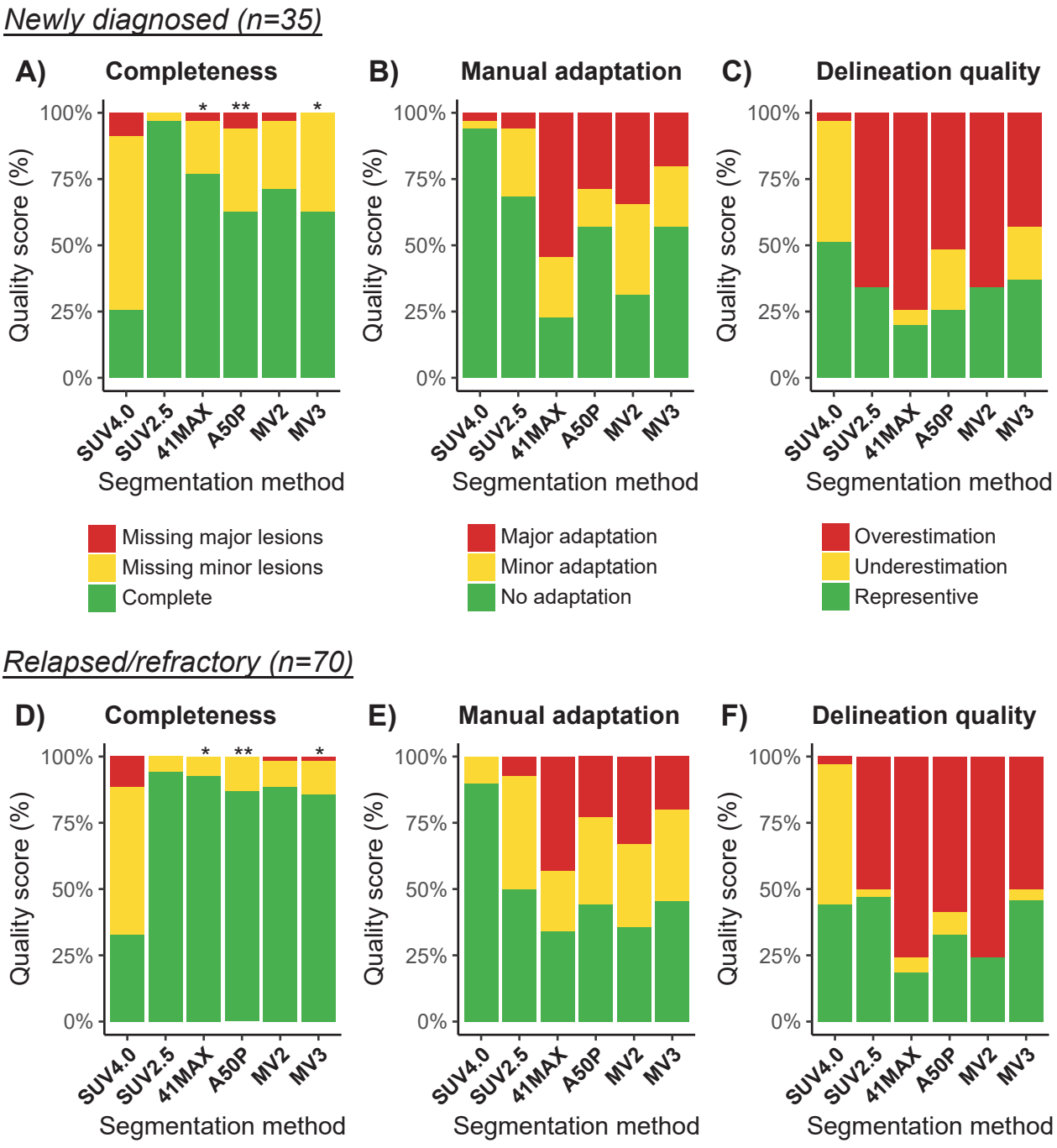
[1]Department of Hematology, Amsterdam UMC, University of Amsterdam, LYMMCARE (Lymphoma and Myeloma Center, Amsterdam), Cancer Center Amsterdam, Amsterdam, The Netherlands; [2]Department of Radiology and Nuclear Medicine, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, Amsterdam, The Netherlands; [3]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA; [4]Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, Amsterdam, The Netherlands. [5]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA; [6]Department of Medicine, Sylvester Comprehensive Cancer Center, Miami, Florida, USA; [7]Department of Hematology, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, Amsterdam, The Netherlands. [8]Department of Epidemiology and Data Science, Amsterdam Public Health research institute, Amsterdam, Netherlands.

Index

Figures:

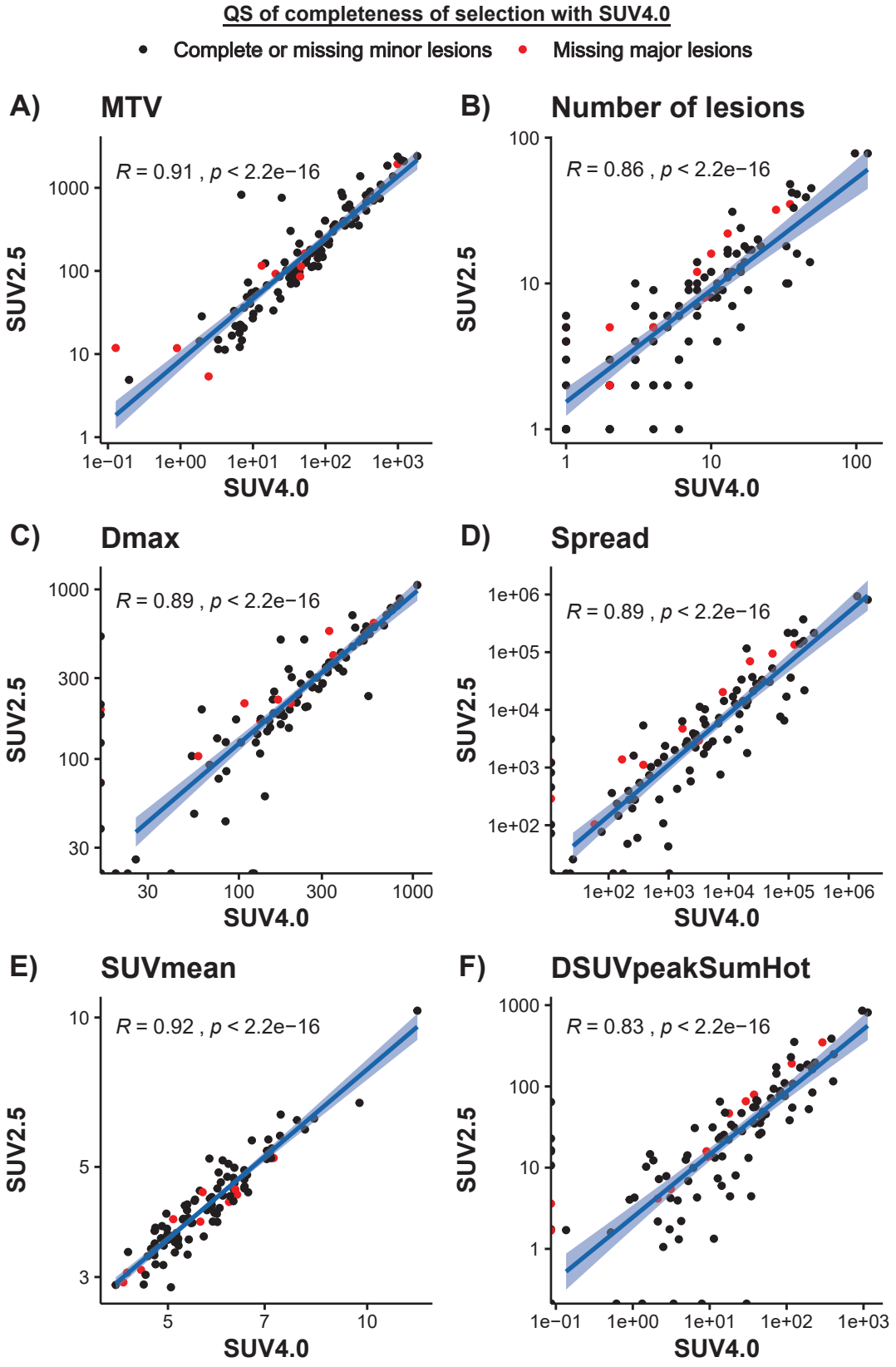
Tables:

Supplemental Figure 1



Description: Quality scores (QS) of segmentation methods, stratified for newly diagnosed (ND) and relapsed/refractory (RR) HL patients. **A+D)** QS of completeness of selection; **B+E)** Manual adaptations after semi-automatic segmentation; **C+F)** QS of delineation of the tumor borders; for ND (A,B,C) and RR (D,E,F) patients, respectively. P-values represent comparisons of QS=complete selection, no manual adaptations, or good delineation for ND vs RR. *P<0.05;
**P<0.01.
Abbreviations: N, number; SUV, standard uptake value; 41MAX, 41% of SUVmax; A50P, a 50% of contrast corrected SUVpeak; MV, majority vote.

Supplemental Figure 2



Description: Scatter plots and Spearman's rank correlation coefficients of radiomics features derived with the SUV4.0 method versus the SUV2.5 method. Black dots represent scans that had a complete selection of all lesions, or missed minor lesions with the SUV4.0 method. Red dots represent scans that missed major lesions with the SUV4.0 method. The SUV2.5 method resulted in complete selections of lesions in all cases.

Supplemental Table 1: Definitions of PET- and radiomics features

Variable	Definition
MTV	The FDG-avid tumor volume
TLG	MTV * SUVmean
SUVmean	The mean SUV value of the VOI
SUVmax	The SUV of the voxel with the highest SUV within the VOI
SUVpeak	The SUV of the 3mL with the highest SUV within the VOI (global peak)
Number of lesions	The number of separated lesion selections within the VOI
Dmax	The maximum distance between two lesions
DmaxBulk	The maximum distance between the largest lesion and any other lesion
Spread	The sum of the distance between all lesions
SpreadBulk	The sum of the distance between the largest lesion and all other lesions
Dvol	The difference in volume between the largest and the smallesfft lesion
VolSpread	The sum of the differences in volume between all lesions
VolSpreadBulk	The sum of the differences in volume between the largest lesion and all other lesions
DSUVmax	The difference in SUVmax between the lesion with the highest SUVmax and the lesion with the lowest SUVmax
DSUVmaxSum	The sum of the differences in SUVmax of all lesions
DSUVmaxBulk	The differences in SUVmax between the largest lesion and all other lesions
DSUVmaxSumBulk	The sum of the differences in SUVmax between the largest lesion and all other lesions
DSUVmaxSumHot	The sum of the differences in SUVmax between the lesion with the highest SUVmax and all other lesions
DSUVpeak	The difference in SUVpeak between the lesion with the highest SUVpeakmax and the lesion with the lowest SUVpeak
DSUVpeakSum	The sum of the differences in SUVpeak of all lesions
DSUVpeakBulk	The differences in SUVpeak between the largest lesion and all other lesions
DSUVpeakSumBulk	The sum of the differences in SUVpeak between the largest lesion and all other lesions
DSUVpeakSumHot	The sum of the differences in SUVpeak between the lesion with the highest SUVpeak and all other lesions

Supplemental Table 2: Patient Characteristics

Variable [n; (%)]	Newly diagnosed *(N=35)*	Relapsed/refractory *(N=70)*	Total *(N=105)*
Sex			
Female	21 (60%)	37 (53%)	58 (55%)
Age			
Median (min, max)	34 (19, 66)	30 (13, 64)	30 (13, 66)
Relapse status*			
Primary refractory	NA	32 (46%)	NA
Relapse	NA	38 (54%)	NA
Ann Arbor stage			
I	2 (6%)	6 (9%)	8 (8%)
II	17 (49%)	25 (36%)	42 (40%)
III	3 (8%)	14 (20%)	17 (16%)
IV	13 (37%)	25 (36%)	38 (36%)
Extranodal disease			
Yes	14 (40%)	26 (37%)	40 (39%)
Progression			
Yes	17 (49%)**	14 (20%)	31 (30%)

Description: Patient characteristics of included PET-CT scans.

*Primary refractory disease was defined as no complete response on first line treatment or relapse <3 months.

** This includes n=15 patients from the RR-cHL cohort of whom the PET-CT scans at primary diagnosis were retrospectively collected. Two other patients of the remaining n=20 newly diagnosed patients showed progression on first-line treatment but were not included in the RR cohort. Therefore, the percentage of patients with progression during or after first-line treatment is not representive for the general population of primary diagnosed cHL patients.

Supplemental Table 3: Radiomics features per patient group

Variable [ND – RR; p-value]	Method					
	SUV4.0	SUV2.5	41MAX	A50P	MV2	MV3
MTV	112.4 - 34.7; p=0.001	352.1 - 101.2; p<0.001	398.2 - 108.8; p<0.001	321.9 - 94.5; p=0.001	360.8 - 98.7; p<0.001	285.3 - 84.9; p=0.001
TLG	716.3 - 183.9; p=0.001	1436 - 401; p<0.001	1618 - 460.1; p<0.001	1300 - 382.6; p<0.001	1432 - 441.4; p<0.001	988.5 - 406.1; p<0.001
SUVmean	5.8 - 5.5; p=0.465	4.3 - 4; p=0.22	4 - 3.8; p=0.726	4.3 - 4; p=0.364	4.1 - 3.9; p=0.364	4.3 - 4.3; p=0.613
SUVmax	11.6 - 9.9; p=0.294	11.6 - 9.9; p=0.294	11.6 - 9.9; p=0.303	11.6 - 9.9; p=0.291	11.6 - 9.9; p=0.263	11.6 - 9.9; p=0.294
SUVpeak	9.4 - 7.4; p=0.016	9.4 - 7.4; p=0.017	9.4 - 7.4; p=0.017	9.4 - 7.4; p=0.017	9.4 - 7.4; p=0.017	9.4 - 7.4; p=0.017
Number of lesions	6 - 8; p=0.892	6 - 7.5; p=0.751	3 - 6; p=0.227	5 - 6.5; p=0.832	3 - 5.5; p=0.267	4 - 6; p=0.67
Dmax	18.1 - 20.8; p=0.477	21 - 24.2; p=0.716	19 - 22.2; p=0.374	19.4 - 22.9; p=0.582	18.4 - 23.4; p=0.372	19.4 - 24.4; p=0.605
DmaxBulk	15 - 17.8; p=0.516	17.9 - 20; p=0.696	18.1 - 18.4; p=0.401	18.3 - 19.6; p=0.596	17.3 - 17.8; p=0.414	18.3 - 19.5; p=0.624
Spread	2033 - 2821; p=0.916	1536 - 3221; p=0.729	497.7 - 2019; p=0.29	889.9 - 2163; p=0.913	575.1 - 1891; p=0.273	889.9 - 2209; p=0.541
SpreadBulk	58.3 - 69.4; p=0.9	50.3 - 98.8; p=0.634	29.7 - 62.2; p=0.335	39.9 - 64.7; p=0.865	30.4 - 63.3; p=0.273	32.2 - 63.9; p=0.663
Dvol	45.4 - 16.4; p=0.04	146.9 - 39.2; p=0.004	135.7 - 55.1; p=0.063	88.1 - 42.8; p=0.013	121 - 48.6; p=0.145	115.2 - 45.9; p=0.013
VolSpread	412.3 - 154.8; p=0.122	686 - 423.9; p=0.159	357.1 - 280.9; p=0.598	361.3 - 240.3; p=0.106	660.2 - 315.8; p=0.577	368.3 - 311; p=0.211
VolSpreadBulk	236.2 - 89.4; p=0.072	658 - 263.4; p=0.087	237.2 - 182.3; p=0.356	335.5 - 197.2; p=0.049	246.6 - 192.5; p=0.438	362.5 - 204; p=0.106
DSUVmax	6.1 - 5; p=0.793	7.5 - 6.5; p=0.654	5.2 - 5.4; p=0.49	7 - 5.4; p=0.536	5.6 - 5.7; p=0.698	6.9 - 6.3; p=0.77
DSUVmaxSum	38.2 - 49.6; p=0.731	54.1 - 72.7; p=0.86	10.5 - 32.8; p=0.207	39.4 - 36.7; p=0.916	19.3 - 26; p=0.344	25.1 - 42.9; p=0.754
DSUVmaxBulk	5.2 - 4.7; p=0.841	7.3 - 6.4; p=0.534	4.7 - 4.2; p=0.965	6.1 - 4.9; p=0.498	4.7 - 5; p=0.989	6.5 - 5.4; p=0.412
DSUVmaxSumBulk	26 - 18.6; p=0.788	35.4 - 28.6; p=0.822	8.3 - 13.3; p=0.408	17.9 - 14.5; p=0.545	14.3 - 16.2; p=0.426	19.1 - 15.7; p=0.708
DSUVmaxSumHot	27.6 - 27.3; p=0.804	35.4 - 31.2; p=0.775	8.3 - 12.6; p=0.685	28.2 - 12.7; p=0.369	13.8 - 12.9; p=0.653	19.9 - 15.9; p=0.713
DSUVpeak	4.1 - 3; p=0.284	5.2 - 4.2; p=0.187	4.4 - 4.2; p=0.783	5.6 - 4.2; p=0.292	4.7 - 4.2; p=0.843	5.5 - 4; p=0.167
DSUVpeakSum	26.5 - 26; p=0.46	36 - 39; p=0.754	8.8 - 27.1; p=0.293	31.7 - 25.3; p=0.61	18.6 - 20; p=0.505	21.9 - 32.1; p=0.854
DSUVpeakBulk	3.8 - 2.9; p=0.346	5.2 - 3.9; p=0.111	4.4 - 3.1; p=0.702	5.1 - 3.3; p=0.187	4.2 - 3.2; p=0.635	5.3 - 3.6; p=0.092
DSUVpeakSumBulk	20.9 - 11.4; p=0.424	28.1 - 16.3; p=0.45	8.3 - 12; p=0.701	19.5 - 12.3; p=0.308	12.8 - 12.6; p=0.648	18.7 - 12.7; p=0.426
DSUVpeakSumHot	20.9 - 14.2; p=0.416	28.1 - 19.3; p=0.501	8.3 - 12.2; p=0.69	23.7 - 11.3; p=0.392	13.9 - 12.5; p=0.624	19.5 - 12.7; p=0.45

Description: PET and radiomics features per method, stratified for newly diagnosed (ND) and relapsed/ refractory (RR) HL patients. Numbers represent median values for ND and RR patients. P-values are derived with wilcoxon rank sum test. P-values <0.05 are highlighed in green.

Supplemental Table 4: Quality Scores per method and per patient group

	Quality Score n (%)		SUV4.0	SUV2.5	41MAX	A50P	MV2	MV3
All patients (n=105)	Completeness	Complete	32 (30%)	100 (95%)	92 (88%)	83 (79%)	87 (83%)	82 (78%)
		Missing minor lesions	62 (59%)	5 (5%)	12 (11%)	20 (19%)	16 (15%)	22 (21%)
		Missing major lesions	11 (10%)	0 (0%)	1 (1%)	2 (2%)	2 (2%)	1 (1%)
	Manual adaptation	No adaptation	96 (91%)	59 (56%)	32 (30%)	51 (49%)	36 (34%)	52 (50%)
		Minor adaptation	8 (8%)	39 (37%)	24 (23%)	28 (27%)	34 (32%)	32 (30%)
		Major adaptation	1 (1%)	7 (7%)	49 (47%)	26 (25%)	35 (33%)	21 (20%)
	Delineation quality	Representive	49 (47%)	45 (43%)	20 (19%)	32 (30%)	29 (28%)	45 (43%)
		Underestimation	53 (50%)	2 (2%)	6 (6%)	14 (13%)	0 (0%)	10 (10%)
		Overestimation	3 (3%)	58 (55%)	79 (75%)	59 (56%)	76 (72%)	50 (48%)
Newly diagnosed (n=35)	Completeness	Complete	9 (26%)	34 (97%)	27 (77%)	22 (63%)	25 (71%)	22 (63%)
		Missing minor lesions	23 (66%)	1 (3%)	7 (20%)	11 (31%)	9 (26%)	13 (37%)
		Missing major lesions	3 (9%)	0 (0%)	1 (3%)	2 (6%)	1 (3%)	0 (0%)
	Manual adaptation	No adaptation	33 (94%)	24 (69%)	8 (23%)	20 (57%)	11 (31%)	20 (57%)
		Minor adaptation	1 (3%)	9 (26%)	8 (23%)	5 (14%)	12 (34%)	8 (23%)
		Major adaptation	1 (3%)	2 (6%)	19 (54%)	10 (29%)	12 (34%)	7 (20%)
	Delineation quality	Representive	18 (51%)	12 (34%)	7 (20%)	9 (26%)	12 (34%)	13 (37%)
		Underestimation	16 (46%)	0 (0%)	2 (6%)	8 (23%)	0 (0%)	7 (20%)
		Overestimation	1 (3%)	23 (66%)	26 (74%)	18 (51%)	23 (66%)	15 (43%)
Relapsed/refractory (n=70)	Completeness	Complete	23 (33%)	66 (94%)	65 (93%)	61 (87%)	62 (89%)	60 (86%)
		Missing minor lesions	39 (56%)	4 (6%)	5 (7%)	9 (13%)	7 (10%)	9 (13%)
		Missing major lesions	8 (11%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	1 (1%)
	Manual adaptation	No adaptation	63 (90%)	35 (50%)	24 (34%)	31 (44%)	25 (36%)	32 (46%)
		Minor adaptation	7 (10%)	30 (43%)	16 (23%)	23 (33%)	22 (31%)	24 (34%)
		Major adaptation	0 (0%)	5 (7%)	30 (43%)	16 (23%)	23 (33%)	14 (20%)
	Delineation quality	Representive	31 (44%)	33 (47%)	13 (19%)	23 (33%)	17 (24%)	32 (46%)
		Underestimation	37 (53%)	2 (3%)	4 (6%)	6 (9%)	0 (0%)	3 (4%)
		Overestimation	2 (3%)	35 (50%)	53 (76%)	41 (59%)	53 (76%)	35 (50%)

Description: Quality scores (QS) of segmentation for 6 different segmentation methods in: all classical Hodgkin lymphoma (cHL) patients, complementary to Figure 2, newly-diagnosed cHL patients, complementary to Supplemental Figure 1A/B/C, and relapsed/refractory cHL patients complementary to Supplemental Figure 1D/E/F.

Abbreviations: N, number; SUV, standard uptake value; 41MAX, 41% of SUVmax; A50P, a 50% of contrast corrected SUVpeak; MV, majority vote.

Supplemental Table 5: Summary statistics for radiomics features per method

Variable [Median (min – max)]	Method					
	SUV4.0	SUV2.5	41MAX	A50P	MV2	MV3
MTV	43.7 (0.1 - 1,853)	123.5 (4.9 - 2,402)	161.0 (5.0 - 2,694)	107.0 (8.0 - 2,655)	143.4 (6.0 - 3,481)	107.8 (2.6 - 3,489)
TLG	252.9 (0.5 - 10,704)	554.5 (14.1 - 12,481)	609.5 (33.3 - 11,463)	448.4 (28.5 - 13,055)	635.7 (24.1 - 13,159)	458.8 (14.1 - 13,205)
SUVmean	5.6 (4.2 - 11.9)	4.1 (2.9 - 10.3)	3.9 (1.8 - 16.4)	4.1 (2.1 - 16.8)	3.9 (1.8 - 10.1)	4.3 (2.2 - 13.6)
SUVmax	11.3 (4.2 - 28.3)	11.3 (4.2 - 28.3)	11.3 (4.2 - 28.3)	11.3 (4.2 - 28.3)	11.3 (4.2 - 28.3)	11.3 (4.2 - 28.3)
SUVpeak	8.0 (4.2 - 24.2)	8.0 (2.9 - 24.2)	8.0 (2.5 - 24.2)	8.0 (2.5 - 24.2)	8.0 (2.5 - 24.2)	8.0 (2.9 - 24.2)
Number of lesions	8 (1 - 120)	7 (1 - 78)	5 (1 - 77)	5 (1 - 84)	5 (1 - 116)	5 (1 -79)
Dmax	19.6 (0 - 106.1)	22.4 (0 - 106.1)	21.4 (0 - 106.1)	21.9 (0 - 106.1)	21.9 (0 - 106.1)	21.9 (0 - 106.1)
DmaxBulk	16.9 (0 - 94.0)	18.4 (0 - 88.6)	18.2 (0 - 70.6)	18.6 (0 - 88.6)	17.8 (0 - 88.6)	18.4 (0 - 70.6)
Spread	2,726 (0 - 2,075,620)	2,530 (0 - 931,350)	1,568 (0 - 884,185)	1,646 (0 - 1,053,960)	1,079 (0 - 1,450,100)	1,760 (0 - 915,957)
SpreadBulk	67.9 (0 - 3,755)	73.7 (0 - 2,543)	57.8 (0 - 2,357)	55.1 (0 - 1,924)	49.6 (0 - 2,046)	56.4 (0 - 2,587)
Dvol	19.9 (0 - 1,850)	63.9 (0 - 1,758)	74.4 (0 - 1,822)	52.5 (0 - 1,606)	53.3 (0 - 3,331)	55.7 (0 - 3,319)
VolSpread	196.8 (0 - 121,389)	456.4 (0 - 149,436)	282.9 (0 - 119,711)	274.0 (0 - 123,206)	353.4 (0 - 301,834)	317.5 (0 - 177,103)
VolSpreadBulk	124.7 (0 - 45,648)	301.2 (0 - 72,054)	211.9 (0 - 65,922)	226.0 (0 -57,759)	197.8 (0 - 264,307)	243.3 (0 - 160,844)
DSUVmax	5.3 (0 -20.9)	7.2 (0 -25.5)	5.3 (0 -25.2)	6.1 (0 -25.2)	5.7 (0 -25.2)	6.5 (0 -25.2)
DSUVmaxSum	45.7 (0 - 27,678)	69.3 (0 - 12,303)	26.3 (0 - 12,180)	39.4 (0 - 15,088)	25.5 (0 - 19,915)	32.7 (0 - 12,584)
DSUVmaxBulk	5.0 (0 -20.9)	6.6 (0 -25.5)	4.7 (0 -25.2)	5.1 (0 -25.2)	4.9 (0 -25.2)	5.6 (0 -25.2)
DSUVmaxSumBulk	25.4 (0 - 1,324)	30.8 (0 - 917.1)	13.1 (0 - 831.7)	15.7 (0 - 705.7)	14.8 (0 - 1,580)	17.4 (0 - 823.6)
DSUVmaxSumHot	27.6 (0 - 1,32)	31.2 (-1 - 963.3)	12.4 (-20.7 - 831.7)	13.7 (-16.8 - 739.8)	13.7 (-5.3 - 1,592)	17.6 (-3.7 - 1,031)
DSUVpeak	3.4 (0 -16.9)	4.8 (0 -21.5)	4.2 (0 -22.2)	4.7 (0 -22.2)	4.3 (0 -22.2)	4.3 (0 -21.5)
DSUVpeakSum	26.3 (0 -18,667)	36.0 (0 - 8,755)	22.2 (0 - 9,583)	28.8 (0 - 11,831)	18.6 (0 - 11,523)	24.3 (0 - 8,635)
DSUVpeakBulk	3.1 (0 - 16.9)	4.3 (0 - 21.5)	3.3 (0 - 22.2)	4.0 (0 - 22.2)	3.6 (0 - 22.2)	4.0 (0 - 21.5)
DSUVpeakSumBulk	13.1 (0 - 1,136)	22.5 (0 - 813)	11.4 (0 - 759)	12.6 (0 - 757)	12.7 (0 - 1,137)	14.5 (0 - 749)
DSUVpeakSumHot	15.3 (0 - 1,136)	22.8 (0 - 854.9)	11.9 (0 - 759.0)	12.6 (0 - 834.3)	12.8 (0 - 1,448)	14.2 (0 - 894.6)

Description: Summary statistics of PET and radiomics features, stratified per segmentation method. Numbers represent median values and ranges. Volumes are in mL, distances are in cm.

Supplemental Table 6: Correlation coefficients for radiomics features among different methods

Variable [*R*]	Spearman's correlation between methods														
Method 1	SUV4.0	SUV4.0	SUV4.0	SUV4.0	SUV4.0	SUV2.5	SUV2.5	SUV2.5	SUV2.5	41MAX	41MAX	41MAX	A50P	A50P	MV2
Method 2	SUV2.5	41MAX	A50P	MV2	MV3	41MAX	A50P	MV2	MV3	A50P	MV2	MV3	MV2	MV3	MV3
MTV	0.91	0.74	0.78	0.82	0.85	0.86	0.88	0.89	0.91	0.91	0.90	0.91	0.87	0.92	0.93
TLG	0.94	0.86	0.88	0.89	0.92	0.94	0.95	0.94	0.96	0.96	0.96	0.95	0.93	0.96	0.96
SUVmean	0.92	0.69	0.71	0.66	0.62	0.77	0.77	0.71	0.71	0.89	0.85	0.81	0.81	0.81	0.86
SUVmax	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SUVpeak	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Number of lesions	0.86	0.78	0.82	0.73	0.80	0.82	0.89	0.81	0.88	0.89	0.87	0.90	0.88	0.92	0.92
Dmax	0.89	0.92	0.92	0.89	0.90	0.93	0.96	0.94	0.95	0.97	0.96	0.96	0.97	0.98	0.97
DmaxBulk	0.84	0.89	0.89	0.87	0.87	0.89	0.93	0.92	0.93	0.95	0.95	0.94	0.96	0.97	0.96
Spread	0.89	0.84	0.87	0.81	0.85	0.86	0.91	0.85	0.90	0.93	0.91	0.94	0.93	0.95	0.94
SpreadBulk	0.90	0.84	0.88	0.82	0.86	0.87	0.93	0.87	0.92	0.92	0.91	0.94	0.94	0.96	0.95
Dvol	0.70	0.64	0.61	0.63	0.67	0.84	0.87	0.84	0.89	0.82	0.85	0.87	0.80	0.85	0.86
VolSpread	0.83	0.81	0.83	0.77	0.82	0.90	0.96	0.89	0.94	0.92	0.89	0.93	0.90	0.94	0.92
VolSpreadBulk	0.80	0.79	0.80	0.73	0.79	0.89	0.94	0.86	0.92	0.90	0.87	0.91	0.87	0.92	0.89
DSUVmax	0.78	0.73	0.76	0.74	0.76	0.82	0.90	0.83	0.90	0.85	0.89	0.88	0.83	0.89	0.89
DSUVmaxSum	0.87	0.80	0.84	0.76	0.81	0.84	0.91	0.83	0.89	0.90	0.90	0.92	0.90	0.93	0.92
DSUVmaxBulk	0.73	0.68	0.70	0.69	0.69	0.78	0.85	0.79	0.86	0.83	0.86	0.81	0.79	0.83	0.84
DSUVmaxSumBulk	0.84	0.79	0.82	0.76	0.78	0.82	0.89	0.82	0.87	0.90	0.90	0.90	0.89	0.93	0.90
DSUVmaxSumHot	0.84	0.78	0.74	0.76	0.76	0.80	0.79	0.79	0.85	0.81	0.84	0.86	0.77	0.81	0.86
DSUVpeak	0.74	0.72	0.74	0.72	0.72	0.79	0.89	0.81	0.88	0.84	0.90	0.87	0.85	0.89	0.87
DSUVpeakSum	0.87	0.79	0.83	0.75	0.81	0.83	0.89	0.82	0.88	0.90	0.90	0.92	0.89	0.93	0.92
DSUVpeakBulk	0.69	0.65	0.66	0.65	0.66	0.77	0.85	0.79	0.86	0.83	0.86	0.81	0.80	0.85	0.84
DSUVpeakSumBulk	0.83	0.80	0.83	0.75	0.78	0.82	0.89	0.83	0.87	0.90	0.90	0.90	0.88	0.93	0.90
DSUVpeakSumHot	0.83	0.77	0.80	0.76	0.77	0.81	0.84	0.82	0.86	0.86	0.87	0.90	0.87	0.89	0.89

Color scale of *R*					
0.0-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0

Description: Spearman's rank correlation coefficients for PET- and radiomics features between different segmentation methods. Columns represent correlation coefficients between Method 1 and Method 2. All correlations were significant with $P<0.01$.

Supplemental Table 7: Area Under the Curve for radiomics features per method

Variable (AUC)	Method					
	SUV4.0	SUV2.5	41MAX	A50P	MV2	MV3
MTV	0.64	0.65	0.62	0.62	0.64	0.63
TLG	0.64	0.64	0.64	0.63	0.65	0.63
SUVmean	0.53	0.55	0.57	0.55	0.57	0.56
SUVmax	0.54	0.54	0.54	0.54	0.54	0.54
SUVpeak	0.56	0.56	0.56	0.56	0.56	0.56
Number of lesions	0.63	0.60	0.58	0.62	0.55$^{\delta\lambda}$	0.60
Dmax	0.58	0.56	0.56	0.57	0.55	0.56
DmaxBulk	0.57	0.54$^{\delta}$	0.58	0.59	0.56	0.57
Spread	0.62	0.60	0.58	0.61	0.55	0.60
SpreadBulk	0.62	0.59	0.58	0.61	0.56$^{\delta\lambda}$	0.61
Dvol	0.62	0.64	0.62	0.61	0.56$^{\beta}$	0.62
VolSpread	0.65	0.64	0.62	0.63	0.58$^{\lambda}$	0.64
VolSpreadBulk	0.65	0.63	0.62	0.64	0.57$^{\delta\lambda}$	0.64
DSUVmax	0.54	0.56	0.59	0.55	0.52$^{\gamma}$	0.55
DSUVmaxSum	0.62	0.61	0.59	0.61	0.55$^{\delta\lambda}$	0.61
DSUVmaxBulk	0.53	0.56	0.60	0.56	0.51$^{\gamma}$	0.56
DSUVmaxSumBulk	0.61	0.60	0.60	0.61	0.54$^{\delta\lambda}$	0.62
DSUVmaxSumHot	0.61	0.61	0.62	0.57	0.54$^{\gamma}$	0.59
DSUVpeak	0.57	0.58	0.59	0.59	0.55	0.58
DSUVpeakSum	0.62	0.62	0.59	0.62	0.56$^{\delta\lambda}$	0.62
DSUVpeakBulk	0.56	0.57	0.61	0.59	0.53$^{\gamma}$	0.58
DSUVpeakSumBulk	0.62	0.60	0.61	0.63	0.54$^{\delta\lambda\gamma}$	0.61
DSUVpeakSumHot	0.62	0.61	0.60	0.63	0.56$^{\delta}$	0.61
Median AUC	**0.62**	**0.60**	**0.59**	**0.61**	**0.55**	**0.60**

Description: Area under the curve (AUC) derived from receiver operating characteristic (ROC) analysis for each feature stratified per segmentation method. AUCs were compared between methods using a paired t-test as described by DeLong et al. A p-value of <0.05 was considered statistically significant.

$^{\alpha}$significantly lower compared to SUV4.0
$^{\beta}$significantly lower compared to SUV2.5
$^{\gamma}$significantly lower compared to 41max
$^{\delta}$significantly lower compared to A50P
$^{\kappa}$significantly lower compared to MV2
$^{\lambda}$significantly lower compared to MV3