REGULAR ARTICLE

blood advances

Prognostic model using ^{18}F-FDG PET radiomics predicts progression-free survival in relapsed/refractory Hodgkin lymphoma

Julia Driessen,[1-3] Gerben J. C. Zwezerijnen,[2,4] Heiko Schöder,[5] Marie José Kersten,[1,3] Alison J. Moskowitz,[6] Craig H. Moskowitz,[7] Jakoba J. Eertink,[2,8] Martijn W. Heymans,[9] Ronald Boellaard,[2,4] and Josée M. Zijlstra[2,8]

[1]Department of Hematology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands; [2]Division of Imaging and Biomarkers, Cancer Center Amsterdam, Amsterdam, The Netherlands; [3]LYMMCARE, Lymphoma and Myeloma Center Amsterdam, Amsterdam, The Netherlands; [4]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, The Netherlands; [5]Department of Radiology and [6]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY; [7]Department of Medicine, Sylvester Comprehensive Cancer Center, Miami, FL; [8]Department of Hematology, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; and [9]Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

Key Points

- Quantitative PET radiomics and clinical features can be used to build a strong prognostic model for 3-year PFS in relapsed/refractory cHL.
- We identified a subgroup of high-risk patients with R/R cHL with inferior PFS and overall survival for whom novel therapies should be considered.

Investigating prognostic factors in patients with relapsed or primary refractory classical Hodgkin lymphoma (R/R cHL) is essential to optimize risk-adapted treatment strategies. We built a prognostic model using baseline quantitative ^{18}F-fluorodeoxyglucose positron emission tomography (PET) radiomics features and clinical characteristics to predict the progression-free survival (PFS) among patients with R/R cHL treated with salvage chemotherapy followed by autologous stem cell transplantation. Metabolic tumor volume and several novel radiomics dissemination features, representing interlesional differences in distance, volume, and standard uptake value, were extracted from the baseline PET. Machine learning using backward selection and logistic regression were applied to develop and train the model on a total of 113 patients from 2 clinical trials. The model was validated on an independent external cohort of 69 patients. In addition, we validated 4 different PET segmentation methods to calculate radiomics features. We identified a subset of patients at high risk for progression with significant inferior 3-year PFS outcomes of 38.1% vs 88.4% for patients in the low-risk group in the training cohort ($P < .001$) and 38.5% vs 75.0% in the validation cohort ($P = .015$), respectively. The overall survival was also significantly better in the low-risk group ($P = .022$ and $P < .001$). We provide a formula to calculate a risk score for individual patients based on the model. In conclusion, we developed a prognostic model for PFS combining radiomics and clinical features in a large cohort of patients with R/R cHL. This model calculates a PET-based risk profile and can be applied to develop risk-stratified treatment strategies for patients with R/R cHL. These trials were registered at www.clinicaltrials.gov as #NCT02280993, #NCT00255723, and #NCT01508312.

Introduction

Classical Hodgkin lymphoma (cHL) mainly affects young adults.[1] Treatment consists of chemotherapy and radiotherapy and is successful in most cases.[2] However, ~10% to 20% of patients still relapse or

Submitted 7 April 2023; accepted 25 August 2023; prepublished online on *Blood Advances* First Edition 18 September 2023; final version published online 2 November 2023. https://doi.org/10.1182/bloodadvances.2023010404.

Researchers may request access to certain deidentified data and related study documents by contacting the corresponding author, Julia Driessen (j.driessen@amsterdamumc.nl).

The full-text version of this article contains a data supplement.

© 2023 by The American Society of Hematology. Licensed under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0), permitting only noncommercial, nonderivative use with attribution. All other rights reserved.

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

are primary refractory, of whom about ~50% to 60% can be cured with salvage chemotherapy and autologous stem cell transplantation (ASCT). The remaining 40% to 50% generally have a very poor prognosis.[3,4] Risk profiling at baseline before starting second-line treatment could be used to identify patients with a high risk of progression, for whom novel (immune) therapies can be considered before the start of salvage chemotherapy, instead of adapting treatment to response assessment after reinduction therapy. ^{18}F-Fluorodeoxyglucose (FDG) positron emission tomography (PET) computed tomography (CT)–adapted treatment has improved outcomes for patients with newly diagnosed cHL.[5-7] Although the prognostic value of a complete metabolic response (CMR) before ASCT in patients with relapsed/refractory (R/R) cHL is well known, there is currently no PET-adapted treatment strategy that is widely applied in the salvage treatment setting.[8-10]

Metabolic tumor volume (MTV) is increasingly studied in cHL and has shown moderate prognostic value as a single biomarker.[9,11-16] In most studies, a different cutoff for MTV is used without validation of results, which impedes the use of MTV as a prognostic marker. The Dmax, ie, the maximum distance between 2 lesions, provides another quantitative PET feature, which has shown prognostic value for newly diagnosed cHL.[17,18] Radiomics is an emerging field of research that uses high-throughput imaging-based data to extract quantitative image features from a predefined volume of interest (VOI), such as FDG-avid tumors on a PET. Differences in FDG intensity of the VOI (tumor), shape, volume, localization, texture, and intratumor and intertumor heterogeneity can be investigated and reinforced with available genomic and clinical data to develop prognostic models.[19-22]

Only a few studies have assessed radiomics in cHL, but most prognostic models lack validation in an independent cohort.[18,23,24] A prognostic radiomics model based on texture features in newly diagnosed cHL showed high prognostic value for predicting refractory disease, but results were not validated in an independent cohort.[23] In addition, texture features, which are calculations based on individual voxels, are susceptible to technical variations, especially in small lesions.[25] Many patients present with small lesions in the R/R setting because of early detection during follow-up after first-line treatment.[10] Therefore, radiomics dissemination and interlesion heterogeneity parameters (eg, the spread or the difference in distance, volume, and FDG uptake between lesions), which are less susceptible to technical variations, could be more suitable for use in disseminated diseases with smaller lesions such as lymphoma.[26] Most other prognostic models that have been developed to predict outcomes in the R/R setting, for example, gene expression–based models,[27] have not yet been implemented in a prospective clinical trial or clinical practice, which can possibly be explained by high costs and time-consuming analyses. Because PETs are already used in clinical practice, information obtained through radiomics may contribute to more accurately predict outcomes among patients with cHL and can be implemented in clinical practice to guide treatment decisions, which, in turn, may improve clinical outcome.[20-22]

Materials and methods

Study population

Patients treated within the following 3 clinical trials were included: (1) Kersten et al,[28] who investigated a combination of brentuximab vedotin (BV) and dexamethasone, high-dose cytarabine and cisplatin (DHAP) followed by ASCT; (2) Moskowitz et al,[9,29] who investigated sequential BV and ifosfamide, carboplatin, and etoposide (ICE), followed by ASCT; and (3) Moskowitz et al,[30] who investigated ICE and optional sequential gemcitabine, vinorelbine, and doxorubicin (GVD) for patients with no CMR, followed by ASCT. A complete overview of treatment regimens is provided in supplemental Table 1. All patients were transplant-eligible and had biopsy-proven cHL, and the PET-CT was performed at baseline, that is, before the start of salvage therapy. Patients were excluded if no PET was available or if the follow-up time was <2 years. An overview of reasons for patient exclusion is provided in supplemental Table 2.

All patients provided written informed consent for participation in the clinical trials (NCT02280993, NCT00255723, and NCT01508312), 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 of Amsterdam University Medical Centers, The Netherlands and the Memorial Sloan Kettering Cancer Center, NY.

^{18}F-FDG PET-CTs and quality control

The PET-CT systems used to perform the scans were accredited by the European Association of Nuclear Medicine Research Ltd. (EARL, Europe) or the American College of Radiology (ACR, United States).[31] PET-CTs were deidentified at the participating centers and centrally collected. Inclusion criteria were (1) plasma glucose < 11 mmol/L; (2) reconstruction of attenuation-corrected PET according to guidelines described by EARL or ACR; (3) total image activity (in megabecquerel) between 50% and 80% of the total injected FDG activity or a liver standard uptake value (SUV) mean (SUV_{mean}) between 1.3 and 3.0; and (4) availability of essential PET acquisition data and clinical data.[31,32]

PET segmentation and radiomics feature extraction

Attenuation-corrected PETs were analyzed using the ACCURATE tool, as described before.[26,33] We published earlier that segmentation using a fixed threshold of an SUV of 4.0 (SUV4.0 method) is most suitable for application in the clinical practice for patients with cHL.[26] However, because this method frequently does not include small lesions with low FDG uptake (SUV < 4.0), we also analyzed scans with a threshold of an SUV of 2.5 (SUV2.5 method) and a combination method ("combimethod") in which segmentation with an SUV of 4.0 is complemented with a threshold of an SUV of 2.5 for missing lesions with low uptake (ie, SUV < 4.0). Additionally, we analyzed all scans with a relative threshold of 41% of the SUV_{max} (41max method) for comparison with those of other studies, because this method has also been used frequently in literature.[9,11]

Only focal extranodal and splenic lesions were included in the VOI. A global increase in the FDG uptake of the spleen or bone marrow was not included in the VOI. Delineations were performed by J.D. under supervision of a nuclear medicine physician (G.J.C.Z. or H.S.). RaCat software was used to extract 18 patient-level dissemination, standard intensity–based, and volume-based features such as MTV, SUV parameters, and total lesion glycolysis (ie, SUV_{mean} multiplied by MTV) from the complete VOI at patient level.[34] An overview of all features and their definitions are provided

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

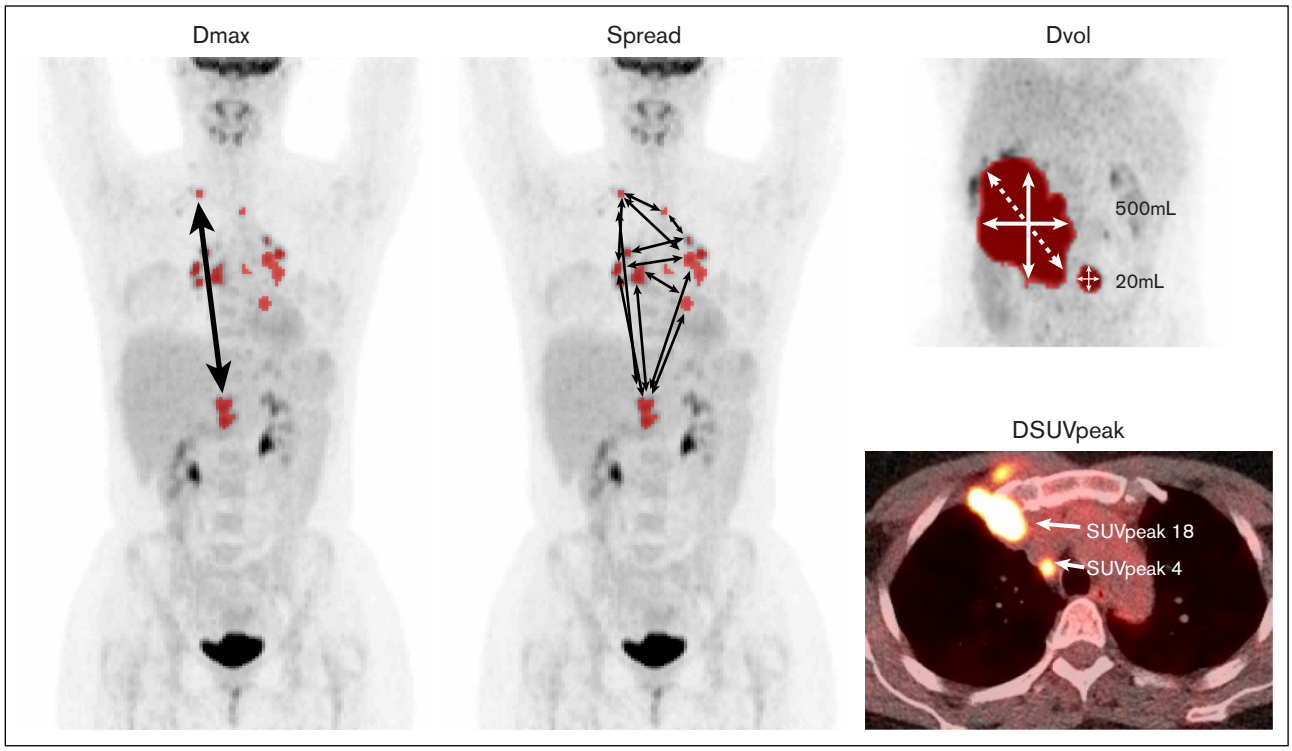


Figure 1. Examples of radiomics features. All definitions of radiomics features are listed in supplemental Table 3. D_{max}, maximum distance between 2 lesions; $DSUV_{peak}$, maximum difference in SUV_{peak} between 2 lesions; Dvol, maximum difference in volume between 2 lesions; spread, sum of the distances between all lesions.

in supplemental Table 3, and examples are given in Figure 1. Dissemination features included several novel features addressing interlesional heterogeneity based on distance, volume, and intensity. Because of the multicenter aspect of this study and the use of different PET systems, we only used robust radiomics features that were not susceptible to technical variations in PET acquisition, such as dissemination features and SUV_{peak} (ie, the average SUV of 1 mL with the highest FDG uptake) instead of SUV_{max} (which represents only the SUV of the highest single voxel, therefore being susceptible to image noise). Additionally, because the SUV_{mean} of the liver was used as a standard quality parameter to compare PETs and was also the reference for a Deauville score of 3, we normalized the SUV_{mean} and SUV_{peak} (ie, the 1 mL with the highest SUV within the VOI) for the liver SUV_{mean} and used the tumor-to-liver ratio (TLR).[32,35-37] The liver SUV_{mean} was estimated on a 3-mL sphere in the right upper lobe of the liver.

End point

The primary end point was to develop a prognostic model for 3-year progression-free survival (PFS) using clinical and radiomics features measured at baseline. PFS was defined as the time from enrollment until progression or death from any cause (binary outcomes: 1 = progression or death and 0 = no event at 3 years). The secondary end point was the 3-year overall survival (OS), defined as the time from enrollment until death from any cause.

Statistical analysis

We analyzed the 18 radiomics dissemination features as listed in supplemental Table 3, and MTV, total lesion glycolysis, $TLR_{SUVmean}$ and $TLR_{SUVpeak}$, and 5 clinical features, that is, age, Ann Arbor stage, extranodal disease, primary refractory disease vs relapsed disease (R/R status), and B symptoms. Radiomics features were log transformed to obtain a linear relationship with the outcome variable. A clinical model was built using only clinical features, a radiomics model was built for each segmentation method, and the final model was built using both clinical and radiomics features using the segmentation method that showed the best performance. We applied a backward feature selection using the stepAIC function of the R package "MASS" version 7.3-53 to select features for each training model and removed features with high multicollinearity. Backward selection was performed separately for each model and could, therefore, result in the selection of different features per model. Cook distance was calculated but identified no extreme outliers. The models were trained using logistic regression on the BV-DHAP and BV-ICE studies (n = 113, training cohort) and validated on the ICE study (n = 69, validation cohort), using the "glm" function of R package "stats" version 4.0.3. Model performance was assessed by calculating the area under the curve (AUC) of the receiver operating characteristics curve on the training and validation cohort, which was also cross-validated on the training cohort using fivefold with 2000 repeats. The significance of the addition of radiomics features to clinical features was calculated using the deltaAUC test of R package "clinfun" version 1.0.15 for comparing the AUC from receiver operating characteristics curves from nested binary regression models.[38] The size of the high-risk group was predefined based on the prevalence of PFS events in the training cohort, which was 26 of 113 (23%). The high-risk group was identified by selecting the top 23% of patients with the highest prediction scores. Another cutoff based on the Youden Index of the cross-validation on the training cohort was

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

Table 1. Patient characteristics of the training and validation cohorts

	Training (n = 113)		Validation (n = 69)		Total (n = 182)		
	No.	%	No.	%	No.	%	*P* value
Study							<.001
BV-DHAP	58	51	0	0	58	32	
BV-ICE	55	49	0	0	55	30	
ICE-GVD	0	0	69	100	69	38	
Female sex	61	54	32	46	93	51	.319
Median age, (range)	30 (13-65)		34 (18-66)		31 (13-66)		.175
Primary refractory	55	50	25	37	80	45	.062
Ann Arbor stage							.002
I	10	9	1	1	11	6	.042
II	46	41	43	62	89	49	.004
III	19	17	2	3	21	12	.004
IV	38	34	23	33	61	34	.970
Extranodal disease	44	39	25	36	69	38	.715
B symptoms	28	25	7	10	35	20	.011

also explored. The Kaplan-Meier method and log-rank test were used to analyze differences in PFS and OS for the high- and low-risk groups. Positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity were calculated, and a Cox proportional hazards regression was performed for the high-risk vs low-risk groups. Statistical analysis was performed using R software version 4.0.3. A P value $< .05$ was considered statistically significant.

Results

Patient characteristics

In total, 231 patients were treated in the 3 studies, of whom n = 49 were excluded from the analysis. A total of 37 (16%) cases were excluded because the PET was of insufficient quality or not compatible with the analysis software (supplemental Table 2). We included 182 patients in the analysis, of whom n = 113 were included in the training cohort (n = 58 treated with BV-DHAP and n = 55 treated with BV-ICE) and n = 69 in the validation cohort (treated with ICE). Patient characteristics are summarized in Table 1. Most clinical characteristics were well distributed across the training and validation cohorts. However, the training cohort consisted of a higher percentage of patients with B symptoms (P = .011) and patients with stage II disease (P = .004), whereas the validation set had more patients with stage III disease (P = .004).

The median follow-up time was 42.4 months (range, 25.5-82.6 months) for the training cohort and 72.3 months (range, 25.5-146.5 months) for the validation cohort. In the training and validation cohort, 26 (23%) and 22 (32%) patients had a 3-year PFS event, and 9 (8%) and 15 patients (22%) died, of whom only 2 (2%) and 1 (1%) died without progressive disease, respectively.

Clinical model

The following clinical patient characteristics were used at time of relapse: age, Ann Arbor stage, presence of extranodal disease, B symptoms, and R/R status. Backward feature selection resulted in selection of 3 variables: stage, B symptoms, and R/R status. The clinical model yielded a cross-validated AUC of 0.729 in the training cohort and an AUC of 0.677 in the validation cohort (Table 2; supplemental Table 4).

Radiomics model with different segmentation methods

For each tumor segmentation method, that is, SUV4.0, SUV2.5, 41max, and combimethod, the backward feature selection was performed on all features as listed in supplemental Table 3. This resulted in 4 prognostic models with different features for each method, of which the SUV4.0 method yielded a cross-validated AUC of 0.691 and highest validated AUC of 0.721 (Table 2; supplemental Figure 1B-E). The AUC values of the SUV2.5 method were comparable with those of the SUV4.0 method, whereas the 41max method yielded lower AUC values (supplemental Table 4). The model of the combimethod, in which segmentation using a threshold of SUV4.0 was combined with a threshold of SUV2.5 for missing lesions with low uptake, did not result in higher AUC values compared with the separate SUV4.0 or SUV2.5 models (validated AUC of 0.712 vs 0.721 and 0.714, respectively). To rule out differences in model performance because of backward feature

Table 2. Model performance

Model	Features*	AUC training cohort (95% CI)	CV-AUC training cohort (95% CI)	AUC validation cohort (95% CI)
Clinical	Ann Arbor stage B symptoms R/R status	0.787 (0.692-0.883)	0.729 (0.724 - 0.734)	0.677 (0.535-0.819)
Radiomics SUV 4.0	No. of lesions VolSpread $TLR_{SUVmean}$	0.719 (0.605-0.833)	0.691 (0.685-0.696)	0.721 (0.580-0.863)
Final model	R/R status B symptoms MTV Spread $TLR_{SUVmean}$	0.837 (0.744-0.930)	0.810 (0.805-0.814)	0.750 (0.627-0.872)
P value of clinical vs final model†		.00094	.0049	<.0001

Spread, the sum of the distances between all lesions; $TLR_{SUVmean}$, tumor-to-liver ratio of the lesion SUVmean and liver SUVmean; VolSpread, the sum of differences in volume between all lesions.

*All radiomics variables are log transformed.

†*P* values represent the added value of the radiomics features to the clinical model. *P* value of the cross-validated (CV)-AUC represents the median *P* value of 2000 repeats of fivefold of cross-validation.

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

selection, the features of the SUV4.0 model selection were also tested on the other methods, but this did not increase the AUCs of these models (supplemental Table 4). Because of high AUC values and a technical validation in an earlier publication, the SUV4.0 method was chosen as the tumor segmentation method for the final model.[26]

Combined prognostic model

For the final prognostic model, backward feature selection was performed using all radiomics features from segmentations with the SUV4.0 method in combination with clinical features. Backward selection resulted in selection of the following features: R/R status, B symptoms, MTV, sum of all distances between all lesions (Spread), and $TLR_{SUVmean}$, and yielded a high cross-validated AUC of 0.810 in the training cohort and an AUC of 0.750 in the validation cohort (Tables 2 and 3). The addition of radiomics features (MTV, Spread, and $TLR_{SUVmean}$) to clinical features showed significant improvement of the AUC in the cross-validated training ($P = .0049$) and validation ($P < .0001$) cohorts (Table 2). Ann Arbor stage was not part of the final prognostic model because it

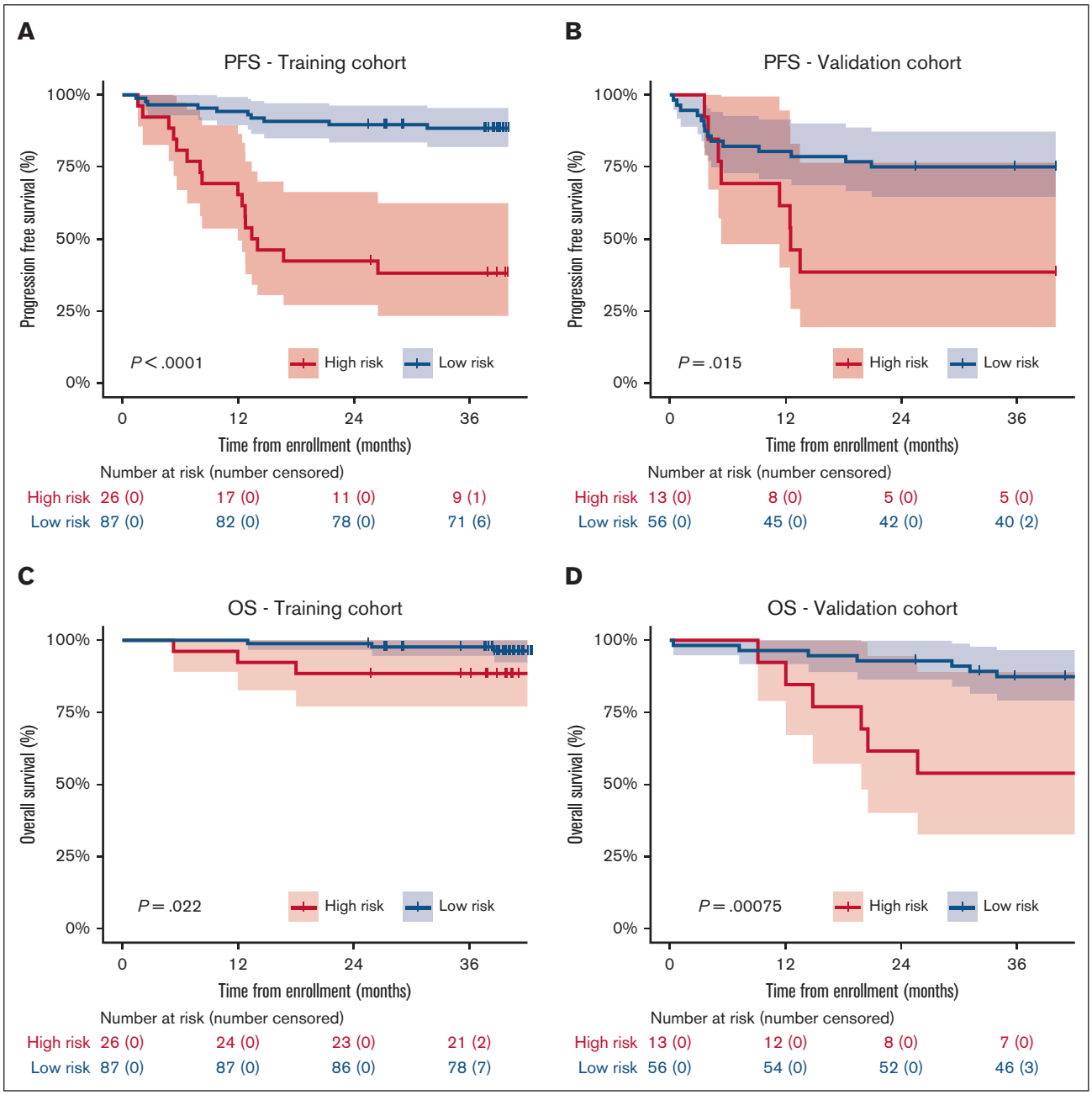


Figure 2. Kaplan-Meier curves of the prognostic model. PFS analysis in the training (A) and independent validation cohort (B). OS analysis in the training (C) and independent validation cohort (D). The size of the high- and low-risk groups were defined according to the percentage of patients with a PFS event in the training cohort, which was 23%. The respective percentage of patients with the highest prediction scores from the logistic regression was classified as high risk.

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

Table 3. Logistic regression results of the model

Features	Estimate	SE	Z value	*P* value
Intercept	-2.5	2.0	-1.2	0.219
R/R (relapsed)	-2.5	0.7	-3.8	0.000
B symptoms	1.0	0.7	1.5	0.136
MTV	-0.4	0.2	-1.6	0.118
Spread	0.4	0.2	2.7	0.007
$TLR_{SUVmean}$	2.4	1.0	2.4	0.018

Logistic regression results of features in the baseline model. Formula of the model: -2.472 − [2.478 * (Relapsed=1, refractory=0)] + [1.010 * (B symptoms = 1, no B symptoms =0)] − [0.384 * log(MTV in uL)] + [0.413 * log(Spread)] + [2.409 * log(SUVmean / liverSUVmean)].

MTV, metabolic tumor volume; R/R, relapsed/refractory; SE, standard error; spread, sum of all distances between all lesions; $TLR_{SUVmean}$, tumor-to-liver ratio of lesion standard uptake value (SUV) mean and the liver.

was being outperformed by the radiomics feature Spread. Replacing Spread for stage resulted in a lower prognostic value for the model (data not shown). Logistic regression results of the model are shown in Table 3.

Based on the predefined cutoff (23% of PFS events), the high-risk group in the training cohort showed a significant inferior PFS compared with patients in the low-risk group, with a 3-year PFS of 38.1% (95% confidence interval [CI], 23-62) vs 88.4% (95% CI; 82-95; $P < .0001$), respectively (Figure 2A; supplemental Table 5). Three-year PFS in the independent validation cohort was 38.5% (95% CI, 19-77) vs 75.0% (95% CI, 65-87; $P = .0153$) for the high- and low-risk groups, respectively (Figure 2B). The 3-year OS was also significantly different between the high- and low-risk groups in the training and validation cohorts (Figure 2C-D). The PPV and NPV for the prediction of 3-year PFS were 61.5% and 88.5%, respectively, in the training cohort, and 61.5% and 75.0%, respectively, in the validation cohort. The PPV and NPV were similar between the 2 studies in the training cohort, that is, the BV-DHAP and BV-ICE studies (Table 4). Results using another exploratory cutoff based on the Youden Index on the cross-validation of the training cohort did not improve the PPV and NPV (supplemental Table 5).

In the training cohort, the CMR rate before ASCT was significantly higher in the low-risk group compared with that of the high-risk group (86% vs 69%; $P = .049$), but this was not the case in the validation cohort. Before ASCT, negative PET result rates seemed higher in the validation cohort because more patients with a positive PET results were excluded during the quality check of PETs. Furthermore, significantly more patients had progressive disease after ASCT in the high-risk groups of both training and validation cohorts (supplemental Table 6).

Table 4. Performance of the model

High- vs low-risk	Training	Validation	BV-DHAP	BV-ICE
Sensitivity	61.5	36.4	61.5	61.5
Specificity	88.5	89.4	93.3	83.3
PPV	61.5	61.5	72.7	53.3
NPV	88.5	75.0	89.4	87.5

Performance of the model shown for the training and validation cohorts. The training cohort consists of the BV-DHAP and BV-ICE studies of which the model performance is also shown separately. The optimal cutoff for high- vs low-risk groups is based on the percentage of PFS events in the training cohort which was 23%.

BV, brentuximab vedotin; DHAP, dexamethasone, high-dose cytarabine and cisplatin; ICE, ifosfamide, carboplatin, etoposide; NPV, negative predictive value; PPV, positive predictive value.

Correlations of clinical and radiomics features

In Figure 3, several radiomics features that were used in the models are stratified for patients with or without a PFS event. MTV was not significantly higher in patients with an event ($P = .12$) (Figure 3A). However, MTV still contributed to the prognostic value of the model because of a complex interaction term with Spread and $TLR_{SUVmean}$, in which Spread showed a higher prognostic value when MTV was high, and in contrast, $TLR_{SUVmean}$ showed a higher prognostic value when MTV was low (Table 3; supplemental Figure 2). A possible explanation for this interaction between Spread, MTV, and $TLR_{SUVmean}$ is that when MTV and Spread are low, a high $TLR_{SUVmean}$ indicates a more aggressive disease and has a worse prognosis, whereas when MTV and Spread are high, the $TLR_{SUVmean}$ is less relevant and the spread of the disease becomes more important to indicate a worse prognosis.

Most radiomics features show moderate to high correlations with other radiomics features. $TLR_{SUVmean}$, which is included in the model, shows the lowest correlations with other radiomics features (Figure 3I). Ann Arbor stage was significantly correlated with MTV, Spread, Dmax, and the number of lesions but not with $TLR_{SUVmean}$ (Figure 3J-N). Patients with B symptoms had significantly higher values of several radiomics features (Figure 3O-S). R/R status did not correlate with any of the radiomics features (data not shown).

Patients with low and high prediction scores

Examples of patients with low and high prediction scores are provided in Figure 4. The formula for the prognostic model can be found in the description of Table 3. Additionally, we created a calculator in Excel format that can be used to calculate the predicted probability for individual patients (supplemental Appendix). For example, patient B from Figure 4 had relapsed disease with no B symptoms, an MTV of 11 mL, Spread of 411 cm, and $TLR_{SUVmean}$ of 2.43, and the model calculated a risk score of 0.05, which is placed in the low-risk group. Correspondingly, this patient had a CMR before ASCT and is still in remission after 41 months of follow-up. In contrast, patient C had primary refractory disease with B symptoms, an MTV of 24 mL, Spread of 677 cm, and $TLR_{SUVmean}$ of 4.59, corresponding to a risk score of 0.88, and relapsed 3 months after ASCT despite an initial CMR on the pre-ASCT PET.

Discussion

There is an unmet need for better risk stratification in the R/R setting for patients with cHL receiving salvage therapy followed by ASCT.[4] Therefore, we have developed a novel prognostic model in R/R cHL for 3-year PFS based on quantitative features from baseline PET scans and clinical characteristics that was validated on an independent data set. The features that were included in our model are robust and not sensitive to technical variations, which makes it more feasible to implement in clinical practice because of the use of different quality of PET scanners across different countries or hospitals.

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

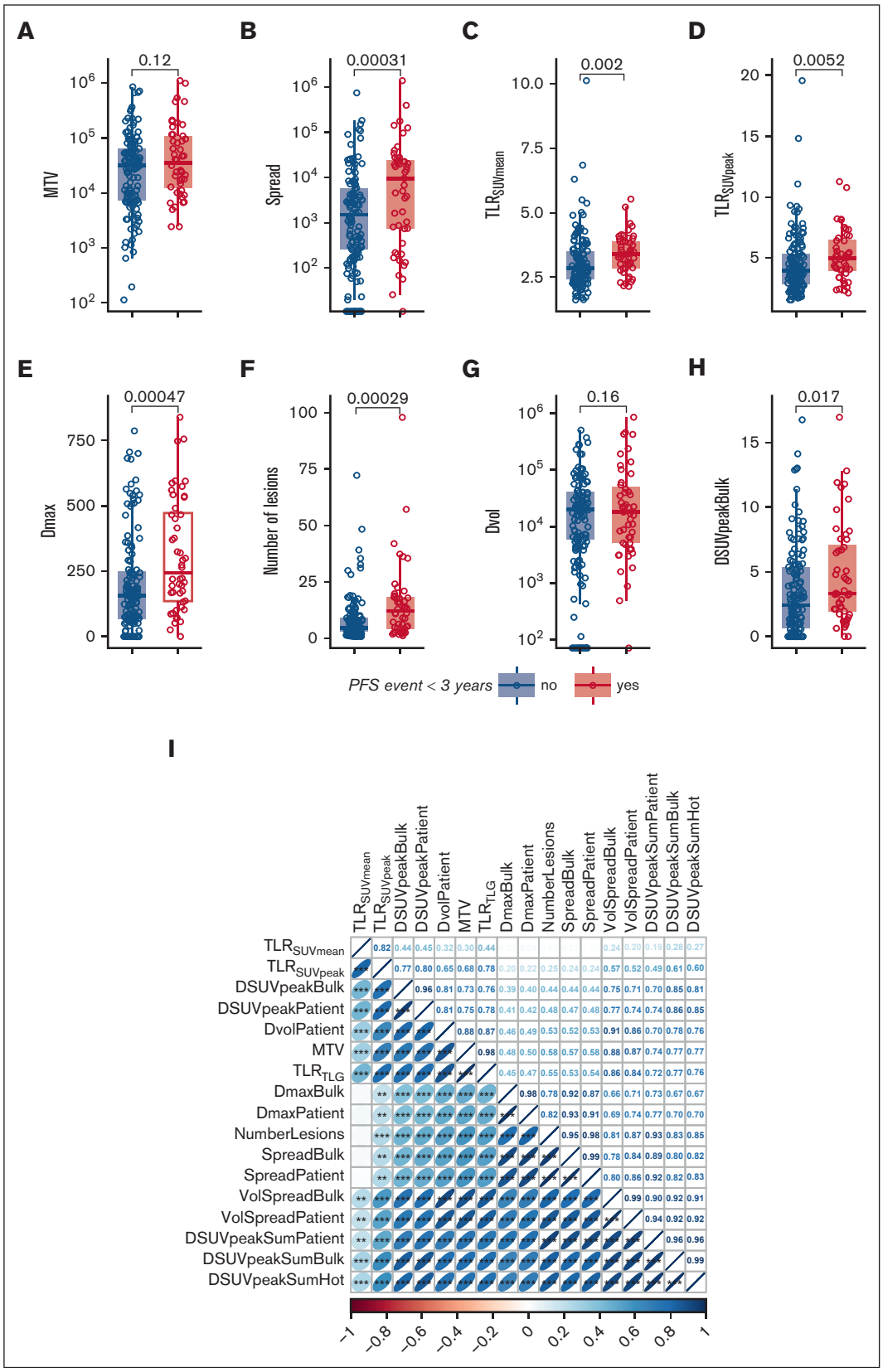


Figure 3. Correlations of radiomics features with PFS outcomes and clinical characteristics and intercorrelations of radiomics features. (A-H) Boxplots of log-transformed radiomics features stratified for patients with or without an event (progression and/or death) on the 3-years PFS. (I) Spearman rank correlation coefficient plot of all

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

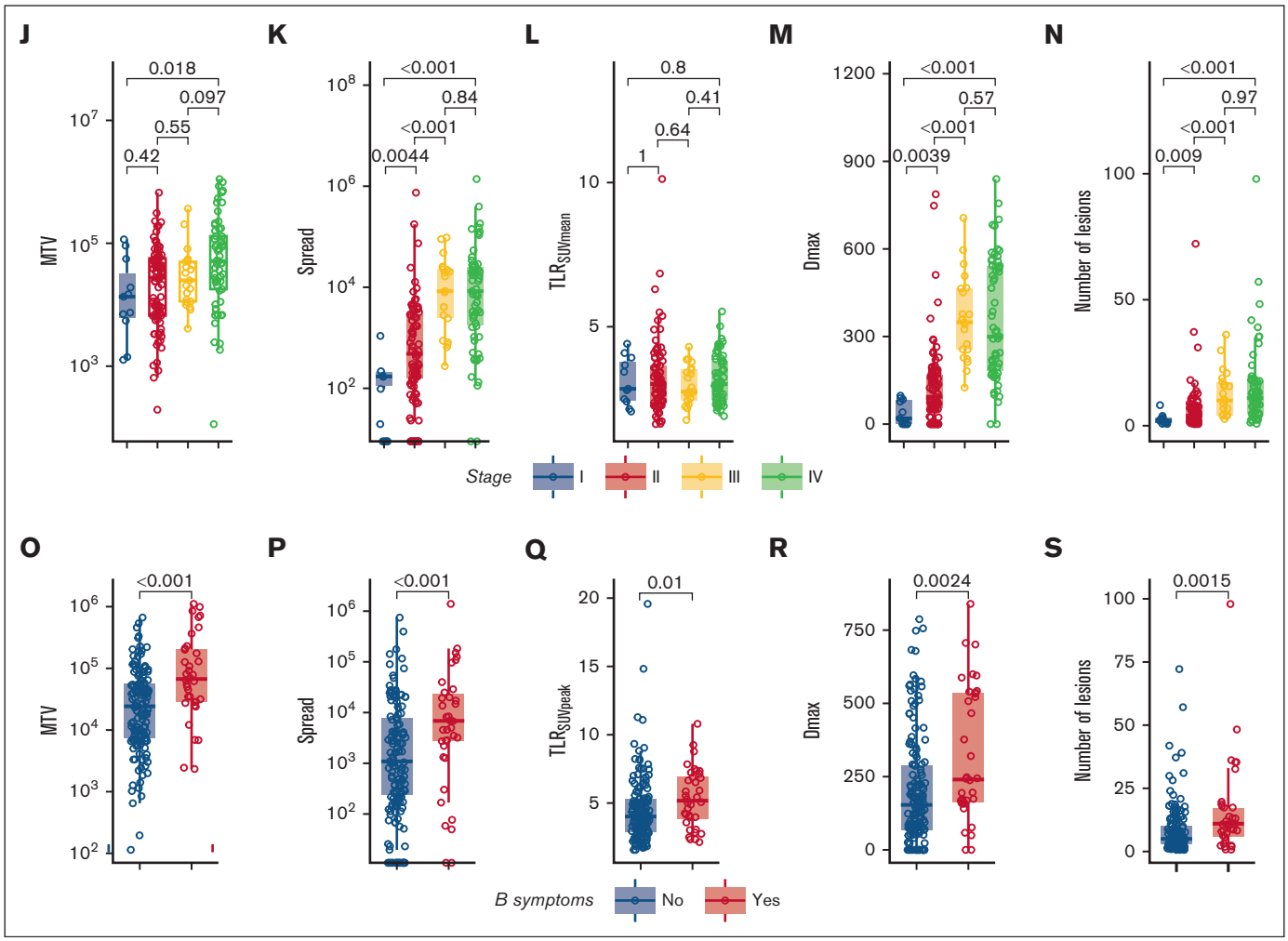


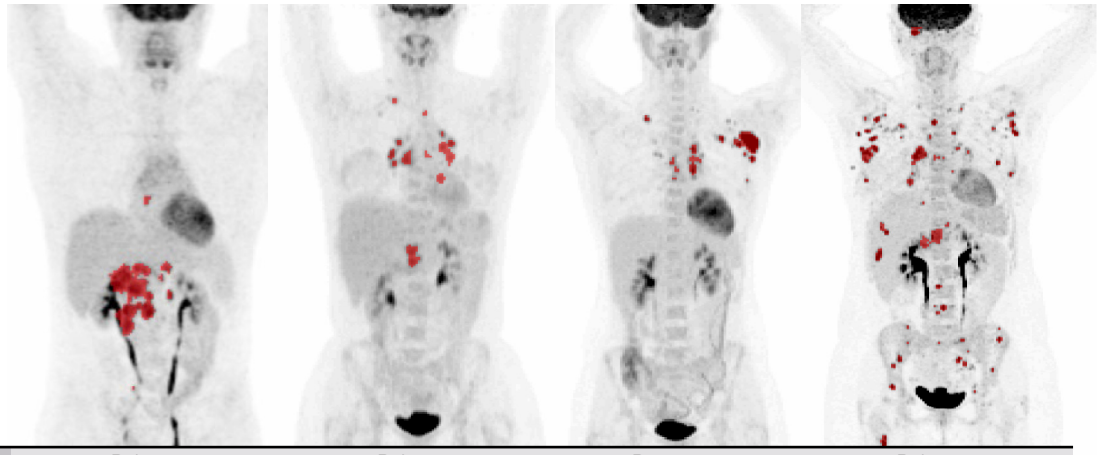
Figure 3 (continued) radiomics features. Asterisks indicate significance values (*$P < .05$; **$P < .01$; ***$P < .001$). (J-N) Boxplots of log-transformed radiomics features stratified for Ann Arbor stage. (O-S) Boxplots of log-transformed radiomics features stratified for the presence of B symptoms at baseline. D_{max}, the largest distance between 2 lesions in mm; $DSUV_{peak}$Bulk, difference of SUVpeak between the largest lesion and the lesion with the lowest SUVpeak in g/mL; Dvol, the difference in volume between the largest and the smallest lesion in mL; NumberLesions, number of lesions; $TLR_{SUVmean}$, tumor-to-liver ratio of mean SUV corrected for the liver SUV_{mean}; $TLR_{SUVpeak}$, ratio of SUV_{peak} corrected for the liver SUV_{mean}.

Several studies have developed prognostic models based on clinical characteristics and pre-ASCT response assessment to predict post-ASCT outcomes, but models for risk profiling at baseline, before starting second-line treatment, are scarce.[8,39] A baseline model has the advantage of being able to preselect patients with high-risk disease and change therapy upfront, preventing these patients from not responding to salvage chemotherapy while being at risk of toxicity. The PPV was 61.5% in both the training and validation cohorts, which is similar or slightly higher than the PPV of pre-ASCT response assessment by PET, as described in literature (PPV ranges between 40% and 60% for individual studies), but with our model, this prediction can already be done at baseline.[9,10,30,40] Therefore, it is worthwhile to explore the model's applicability for changing the salvage therapy in patients with R/R cHL with high-risk disease, such as treating these patients, who are most likely to be chemotherapy resistant, with checkpoint inhibitors.[41] In addition, our model showed a high NPV in both the training and validation cohorts, which means that the model is also suitable for selecting patients with a low risk of progression. This could be used to guide the selection of patients who can potentially be cured by replacing the ASCT for a less toxic consolidation with checkpoint inhibitors, as is currently being evaluated in several studies.[42]

A limitation of our analysis is that 32 PETs (16% of total cohort) had to be excluded from the analysis because of inefficient quality of the PETs or because the PET format was not compatible with our analysis software. These scans mainly originated from the ICE study,[30] which enrolled patients between 2007 and 2010, when the use of PETs was just emerging in clinical practice. Because the quality of PETs has much improved over the years, it is expected that the percentage of excluded PETs will be much lower in future trials.

Not all patient characteristics were balanced between the cohorts, which could have influenced the performance of the model on the validation cohort. Because the model was trained using the training cohort, the validation cohort showed lower AUC values. However,

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024



	A	B	C	D
R/R status	Relapse	Relapse	Refractory	Relapse
B symptoms	No	No	Yes	No
Stage	IV	III	IV	IV
$TLR_{SUVmean}$	3.06	2.43	4.59	4.48
MTV (mL)	81.6	11.0	23.9	47.3
Spread (cm)	194	411	677	39576
Model score	**0.03**	**0.05**	**0.88**	**0.46**
Outcome	**No PD**	**No PD**	**PD 3m after ASCT**	**PD 12m after ASCT**

Figure 4. Examples of maximum intensity projections of baseline PETs in 4 different patients with R/R cHL. The model score was calculated based on the prognostic model using clinical and radiomics features. The outcome represents the clinical outcome of the patient. (A-B) Patients with a low prediction score with low risk of progressive disease. (C-D) Patients with a high prediction score with a high risk of progressive disease. m, months; PD, progressive disease.

the cross-validated AUC of the training cohort closely resembled the AUC of the validation cohort, and the PPV was similar between the cohorts. This indicates that in both the training and validation cohorts, patients with high-risk disease were well identified. In addition, the ICE study had a higher number of events, possibly because this study did not treat patients with BV and the study was conducted ~10 years before the BV-ICE and BV-DHAP studies, so advances in supportive care could have improved over time. In the 3 patient cohorts that we used in this analysis, all patients were intended to receive salvage chemotherapy followed by ASCT, but the salvage chemotherapy schedules were different. Many different salvage regimens are used in R/R cHL across different countries but are generally comparable in terms of efficacy. Therefore, it is also useful to validate the model with different treatment regimens so that it can be extrapolated to other salvage regimens.

We used 3 different semiautomatic segmentation methods, which we have investigated earlier.[26] In our previous study, results with the SUV4.0 and SUV2.5 methods highly correlated and with these methods, there was the lowest need for manual adaptation during segmentation.[26] In the current analysis, the SUV4.0 method yielded the highest validated AUC score for the prognostic model and was again also the least time-consuming method. The combimethod (SUV4.0 + SUV2.5) that we tested did not improve the prognostic value of radiomics features. A possible explanation for this may be that low FDG-avid lesions (SUV < 4.0) are reactive to the lymphoma and do not substantially contribute to the disease characteristics and, therefore, have no influence on the prognostic capabilities of the radiomics features. In this study, we have confirmed the findings of our previous analysis in a larger cohort of patients and propose to use SUV4.0 as a standard segmentation method for cHL at baseline assessment.[26]

Other studies investigating quantitative PET features in cHL mainly focused on only MTV and were performed in single cohorts without validation in external cohorts.[9,11-16] Besides, most studies used a cutoff for MTV instead of the continuous variable in a logistic regression. In our model, MTV was not the highest contributing factor; therefore, we think that combining MTV with other quantitative PET features, for example, intensity and dissemination features, is important because this enables capturing differences between patients with localized bulky disease and those with disseminated disease.

Other biomarkers, such as circulating tumor DNA and thymus and activation regulated chemokine (TARC) have been shown to correlate with MTV.[43] Circulating tumor DNA seems a promising biomarker for detecting minimal residual disease, but its prognostic value at baseline is modest and comparable with that in studies that investigated MTV as a single biomarker.[44] We previously published an analysis of TARC levels in 65 patients with R/R cHL (who are also included in this analysis), in which we demonstrated that TARC has a high prognostic value after 1 cycle of chemotherapy, but it provides no prognostic value at baseline.[10] Combination of TARC and pre-ASCT SUV_{peak} increased the accuracy of predicting progression; therefore, combining biomarkers could possibly enhance the prognostic capacities of biomarker models.

PET-CT is already being performed as part of standard clinical practice in most countries.[37] We showed earlier that semiautomatic segmentation using the SUV4.0 method requires the

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

least manual adaptation by a nuclear medicine physician and is, thus, less observer dependent. Therefore, quantitative analysis of PETs can be used in clinical practice at low extra costs and will probably not be very time consuming. With upcoming technological advances, such as automated segmentation, it is expected that PET radiomics analysis can be performed much easier in the future.[45] Our model consists of robust quantitative PET features, which prevents a high variability of features between different PET scanners, hospitals, and observers. Therefore, quantitative PET analysis provides a promising method for prognostication, which is feasible to be implemented in prospective baseline risk-adapted clinical trials.

Acknowledgments

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

This work was financially supported by Dutch Foundation of hemato-oncological research (http://www.steunhematologie.nl/), which is a nonprofit donation fund of Amsterdam UMC.

The sponsors for this work had no role in gathering, analyzing, or interpreting the data and outcomes described in the manuscript.

Authorship

Contribution: J.D., R.B., and J.M.Z. designed the study; J.D. performed the PET segmentation under supervision of G.J.C.Z. and H.S.; G.J.C.Z. and H.S. reviewed the staging and response assessment of the PETs; J.D. performed the statistical analysis and drafted the manuscript, with contributions from all authors; and all authors collected and interpreted the data, and read, commented on, and approved the final version of the manuscript.

Conflict-of-interest disclosure: R.B. serves as a scientific adviser and chair of the EARL accreditation program. M.J.K. declares consultancy for Bristol Myers Squibb/Celgene, Kite/Gilead, Miltenyi Biotech, Novartis, and Takeda; receives honoraria from Kite/Gilead, Novartis, and Roche; and receives research funding from Kite/Gilead and Takeda. C.H.M. declares research support from Seagen. A.J.M. declares consultancy for Takeda, Imbrium Therapeutics, Janpix, Merck, and Seagen, and receives research funding from Incyte, Merck, Seagen, ADC Therapeutics, BeiGene, Miragen, and Bristol Myers Squibb. J.M.Z. receives research funding from Takeda and Roche, and declares consultancy for Karyopharm. The remaining authors declare no competing financial interests.

ORCID profiles: J.D., 0000-0001-9364-2501; G.J.C.Z., 0000-0002-9571-9362; H.S., 0000-0002-5170-4185; M.J.K., 0000-0002-8904-3802; A.J.M., 0000-0002-3408-050X; C.H.M., 0000-0002-9189-3152; J.J.E., 0000-0002-6094-0016; M.W.H., 0000-0002-3889-0921; R.B., 0000-0002-0313-5686; J.M.Z., 0000-0003-1074-5922.

Correspondence: Julia Driessen, Department of Hematology, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands; email: j.driessen@amsterdamumc.nl.

References

1. Driessen J, Visser O, Zijlstra JM, et al. Primary therapy and relative survival in classical Hodgkin lymphoma: a nationwide population-based study in the Netherlands, 1989-2017. *Leukemia*. 2021;35(2):494-505.
2. Myers RM, Hill BT, Shaw BE, et al. Long-term outcomes among 2-year survivors of autologous hematopoietic cell transplantation for Hodgkin and diffuse large b-cell lymphoma. *Cancer*. 2018;124(4):816-825.
3. von Tresckow B, Muller H, Eichenauer DA, et al. Outcome and risk factors of patients with Hodgkin lymphoma who relapse or progress after autologous stem cell transplant. *Leuk Lymphoma*. 2014;55(8):1922-1924.
4. Driessen J, Tonino SH, Moskowitz AJ, Kersten MJ. How to choose first salvage therapy in Hodgkin lymphoma: traditional chemotherapy vs novel agents. *Hematology Am Soc Hematol Educ Program*. 2021;2021(1):240-246.
5. André MPE, Girinsky T, Federico M, et al. Early positron emission tomography response-adapted treatment in stage I and II Hodgkin lymphoma: final results of the randomized EORTC/LYSA/FIL H10 trial. *J Clin Oncol*. 2017;35(16):1786-1794.
6. Radford J, Illidge T, Counsell N, et al. Results of a trial of PET-directed therapy for early-stage Hodgkin's lymphoma. *N Engl J Med*. 2015;372(17):1598-1607.
7. Borchmann P, Goergen H, Kobe C, et al. PET-guided treatment in patients with advanced-stage Hodgkin's lymphoma (HD18): final results of an open-label, international, randomised phase 3 trial by the German Hodgkin Study Group. *Lancet*. 2017;390(10114):2790-2802.
8. Bröckelmann PJ, Müller H, Casasnovas O, et al. Risk factors and a prognostic score for survival after autologous stem-cell transplantation for relapsed or refractory Hodgkin lymphoma. *Ann Oncol*. 2017;28(6):1352-1358.
9. Moskowitz AJ, Schoder H, Gavane S, et al. Prognostic significance of baseline metabolic tumor volume in relapsed and refractory Hodgkin lymphoma. *Blood*. 2017;130(20):2196-2203.
10. Driessen J, Kersten MJ, Visser L, et al. Prognostic value of TARC and quantitative PET parameters in relapsed or refractory Hodgkin lymphoma patients treated with brentuximab vedotin and DHAP. *Leukemia*. 2022;36(12):2853-2862.
11. 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(13):1456-1463.
12. 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(12):1656-1661.

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

13. Eisazadeh R, Mirshahvalad SA. (18)F-FDG PET/CT prognostic role in predicting response to salvage therapy in relapsed/refractory Hodgkin's lymphoma. *Clin Imaging*. 2022;92:25-31.
14. van Heek L, Stuka C, Kaul H, et al. Predictive value of baseline metabolic tumor volume in early-stage favorable Hodgkin lymphoma - data from the prospective, multicenter phase III HD16 trial. *BMC Cancer*. 2022;22(1):672.
15. Rossi C, André M, Dupuis J, et al. High-risk stage IIB Hodgkin lymphoma treated in the H10 and AHL2011 trials: total metabolic tumor volume is a useful risk factor to stratify patients at baseline. *Haematologica*. 2022;107(12):2897-2904.
16. Milgrom SA, Kim J, Chirindel A, et al. Prognostic value of baseline metabolic tumor volume in children and adolescents with intermediate-risk Hodgkin lymphoma treated with chemo-radiation therapy: FDG-PET parameter analysis in a subgroup from COG AHOD0031. *Pediatr Blood Cancer*. 2021; 68(9):e29212.
17. Cottereau AS, Nioche C, Dirand AS, et al. (18)F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. *J Nucl Med*. 2020;61(1):40-45.
18. Durmo R, Donati B, Rebaud L, et al. Prognostic value of lesion dissemination in doxorubicin, bleomycin, vinblastine, and dacarbazine-treated, interimPET-negative classical Hodgkin lymphoma patients: a radio-genomic study. *Hematol Oncol*. 2022;40(4):645-657.
19. van Helden EJ, Vacher YJL, van Wieringen WN, et al. Radiomics analysis of pre-treatment [(18)F]FDG PET/CT for patients with metastatic colorectal cancer undergoing palliative systemic treatment. *Eur J Nucl Med Mol Imaging*. 2018;45(13):2307-2317.
20. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017; 14(12):749-762.
21. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446.
22. Hsu CY, Doubrovin M, Hua CH, et al. Radiomics features differentiate between normal and tumoral high-FDG uptake. *Sci Rep*. 2018;8(1):3913.
23. Milgrom SA, Elhalawani H, Lee J, et al. A PET radiomics model to predict refractory mediastinal Hodgkin lymphoma. *Sci Rep*. 2019;9(1):1322.
24. Lue KH, Wu YF, Liu SH, et al. Intratumor heterogeneity assessed by (18)F-FDG PET/CT predicts treatment response and survival outcomes in patients with Hodgkin lymphoma. *Acad Radiol*. 2020;27(8):e183-e192.
25. 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(2):665-678.
26. Driessen J, Zwezerijnen GJ, Schöder H, et al. 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. *J Nucl Med*. 2022;63(9):1424-1430.
27. Chan FC, Mottok A, Gerrie AS, et al. Prognostic model to predict post-autologous stem-cell transplantation outcomes in classical Hodgkin lymphoma. *J Clin Oncol*. 2017;35(32):3722-3733.
28. 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(4):1129-1137.
29. Moskowitz AJ, Schöder H, Yahalom J, et al. PET-adapted sequential salvage therapy with brentuximab vedotin followed by augmented ifosamide, carboplatin, and etoposide for patients with relapsed and refractory Hodgkin's lymphoma: a non-randomised, open-label, single-centre, phase 2 study. *Lancet Oncol*. 2015;16(3):284-292.
30. 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(7):1665-1670.
31. Boellaard R, O'Doherty MJ, Weber WA, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging*. 2010;37(1):181-200.
32. 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(2):328-354.
33. Boellaard R. Quantitative oncology molecular analysis suite: ACCURATE. *J Nucl Med*. 2018;59(suppl 1):1753.
34. Pfaehler E, Zwanenburg A, de Jong JR, Boellaard R. An open source and easy to use radiomics calculator tool. *PLoS One*. 2019;14(2):e0212223.
35. Barrington SF, Kluge R. FDG PET for therapy monitoring in Hodgkin and non-Hodgkin lymphomas. *Eur J Nucl Med Mol Imaging*. 2017;44(suppl 1): 97-110.
36. Boktor RR, Walker G, Stacey R, Gledhill S, Pitman AG. Reference range for intrapatient variability in blood-pool and liver SUV for 18F-FDG PET. *J Nucl Med*. 2013;54(5):677-682.
37. 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(27):3059-3068.
38. Demler OV, Pencina MJ, D'Agostino RB, Sr. Misuse of DeLong test to compare AUCs for nested models. *Stat Med*. 2012;31(23):2577-2587.
39. Hahn T, McCarthy PL, Carreras J, et al. Simplified validated prognostic model for progression-free survival after autologous transplantation for hodgkin lymphoma. *Biol Blood Marrow Transplant*. 2013;19(12):1740-1744.
40. Herrera AF, Palmer J, Martin P, et al. Autologous stem-cell transplantation after second-line brentuximab vedotin in relapsed or refractory Hodgkin lymphoma. *Ann Oncol*. 2018;29(3):724-730.

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024

41. Moskowitz AJ, Shah G, Schöder H, et al. Phase II trial of pembrolizumab plus gemcitabine, vinorelbine, and liposomal doxorubicin as second-line therapy for relapsed or refractory classical Hodgkin lymphoma. *J Clin Oncol*. 2021;39(28):3109-3117.

42. Moskowitz AJ. Do all patients with primary refractory/first relapse of HL need autologous stem cell transplant? *Hematology Am Soc Hematol Educ Program*. 2022;2022(1):699-705.

43. Decazes P, Camus V, Bohers E, et al. Correlations between baseline (18)F-FDG PET tumour parameters and circulating DNA in diffuse large B cell lymphoma and Hodgkin lymphoma. *EJNMMI Res*. 2020;10(1):120.

44. Sobesky S, Mammadova L, Cirillo M, et al. In-depth cell-free DNA sequencing reveals genomic landscape of Hodgkin's lymphoma and facilitates ultrasensitive residual disease detection. *Med*. 2021;2(10):1171-1193.e11.

45. Pfaehler E, Mesotten L, Kramer G, et al. Repeatability of two semi-automatic artificial intelligence approaches for tumor segmentation in PET. *EJNMMI Res*. 2021;11(1):4.

Downloaded from http://ashpublications.net/bloodadvances/article-pdf/7/21/6732/2089291/blooda_adv-2023-010404-main.pdf by guest on 05 June 2024