

CLINICAL TRIALS AND OBSERVATIONS

Reclassifying patients with early-stage Hodgkin lymphoma based on functional radiographic markers at presentation

Mani Akhtari,^{1,2} Sarah A. Milgrom,¹ Chelsea C. Pinnix,¹ Jay P. Reddy,¹ Wenli Dong,³ Grace L. Smith,¹ Osama Mawlawi,⁴ Zeinab Abou Yehia,¹ Jillian Gunther,¹ Eleanor M. Osborne,¹ Therese Y. Andraos,¹ Christine F. Wogan,¹ Eric Rohren,⁵ Naveen Garg,⁶ Hubert Chuang,⁷ Joseph D. Khoury,⁸ Yasuhiro Oki,⁹ Michelle Fanale,⁹ and Bouthaina S. Dabaja¹

¹Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX; ²Department of Radiation Oncology, The University of Texas Medical Branch Hospitals, Galveston, TX; ³Department of Biostatistics and ⁴Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX; ⁵Department of Radiology, Baylor College of Medicine, Houston, TX; and ⁶Department of Diagnostic Radiology, ⁷Department of Nuclear Medicine, ⁸Department of Hematopathology, and ⁹Department of Lymphoma/Myeloma, The University of Texas MD Anderson Cancer Center, Houston, TX

KEY POINTS

- Radiographic parameters obtained from the initial PET-CT correlate strongly with survival outcomes in early-stage HL.
- Early-stage unfavorable HL patients can be subdivided into low- and high-risk categories based on these radiographic parameters.

The presence of bulky disease in Hodgkin lymphoma (HL), traditionally defined with a 1-dimensional measurement, can change a patient's risk grouping and thus the treatment approach. We hypothesized that 3-dimensional measurements of disease burden obtained from baseline ¹⁸F-fluorodeoxyglucose positron emission tomography-computed tomography (PET-CT) scans, such as metabolic tumor volume (MTV) and total lesion glycolysis (TLG), would more accurately risk-stratify patients. To test this hypothesis, we reviewed pretreatment PET-CT scans of patients with stage I-II HL treated at our institution between 2003 and 2013. Disease was delineated on prechemotherapy PET-CT scans by 2 methods: (1) manual contouring and (2) subthresholding of these contours to give the tumor volume with standardized uptake value ≥ 2.5 . MTV and TLG were extracted from the threshold volumes (MTV_t, TLG_t) and from the manually contoured soft-tissue volumes. At a median follow-up of 4.96 years for the 267 patients evaluated, 27 patients were diagnosed with relapsed or refractory disease and 12 died. Both MTV_t and TLG_t were highly correlated with freedom from progression and were dichotomized with 80th percentile cutoff values of 268 and 1703, respectively. Consideration of MTV and TLG enabled restratification of

early unfavorable HL patients as having low- and high-risk disease. We conclude that MTV and TLG provide a potential measure of tumor burden to aid in risk stratification of early unfavorable HL patients. (*Blood*. 2018;131(1):84-94)

Introduction

Hodgkin lymphoma (HL) is highly curable. A current research focus is selective deescalation of therapy to reduce treatment-related morbidity while maintaining excellent disease control.¹⁻³ Up-front risk-stratification may be used to guide therapy and determine when treatment deescalation is appropriate.¹⁻⁴ Several groups use slightly different classification systems (Table 1), but, in general, patients are divided into 3 categories: early-stage favorable (ESF), early-stage unfavorable (ESU), and advanced. A common risk factor in all groupings, which results in classification as ESU as opposed to ESF, is the presence of bulky disease. Some definitions of bulky disease have included a mediastinal mass greater than one-third of the maximum intrathoracic diameter or any mass > 10 cm.⁵ One potential shortcoming of these measures however, is the quantification of disease burden based on a 1-dimensional measurement.

It has been known for decades that tumor burden is the most important prognostic factor in early-stage HL.⁶ Although the 2-dimensional (2D) measurement of bulky disease has been

possible for the past few decades, recent advances in functional imaging have made it possible to assess bulk much more accurately by measuring the total metabolic disease burden in 3 dimensions. Using ¹⁸F-fluorodeoxyglucose (¹⁸FDG) positron emission tomography-computed tomography (PET-CT) tumor bulk can be assessed by metabolic tumor volume (MTV) and total lesion glycolysis (TLG).^{7,8} MTV represents the total volumetric sum of all areas of disease; TLG represents the volumetric sum adjusted for standardized uptake value (SUV) and is defined as MTV \times the average SUV. We undertook this study to evaluate the prognostic significance of up-front PET-CT characteristics, specifically MTV and TLG, in early-stage HL patients. Our aim was to evaluate whether these 2 PET-CT markers of total disease burden could be used to further risk-stratify early-stage HL patients.

Methods

Inclusion criteria

After approval by our institutional review board, the records of all patients with a diagnosis of HL treated at our institution between

Table 1. Common criteria for disease stage groupings in Hodgkin lymphoma

	Early-stage favorable	Early-stage unfavorable	IIB-advanced
EORTC ⁴	Stage I or II with no risk factors	Stage I or II with any risk factors	Stage III or IV
GHSG ^{1,2}	Stage I or II with no risk factors	Stage IA or IB and stage IIA with ≥ 1 risk factors Stage IIB with ≥ 1 risk factors, excluding those with bulky disease or extranodal extension	Stage III or IV Stage IIB if with bulky disease or extranodal extension
NCIC ¹⁸	Stage IA or IIA with no risk factors	Stage I or II with any risk factors	Stage III or IV
NCCN ¹⁹	Stage IA or IIA with no risk factors	Stage I or II with any risk factors	Stage III or IV

The definition of involved sites is different for each grouping classification. EORTC defines bulky disease as a mediastinal mass ratio (maximum width of mass/maximum intrathoracic diameter) of >0.35 at T5-T6. EORTC risk factors include age ≥ 50 , bulky disease, >3 involved sites, ESR >50 or >30 if B-symptoms are present. GHSG defines bulky disease as a mediastinal mass ratio of >0.33 . GHSG risk factors include >2 involved sites, bulky disease, extranodal extension, ESR >50 or >30 if B-symptoms are present. NCIC defines bulky disease as a mediastinal mass ratio of >0.33 or a mass >10 cm. NCIC risk factors include age ≥ 40 , bulky disease, B-symptoms, ESR >50 , and >3 involved sites. NCCN defines bulky disease as mediastinal mass ratio of >0.33 or a mass >10 cm. NCCN risk factors include bulky disease, extranodal extension, ESR >50 , or >3 involved sites.

EORTC, European Organization for Research and Treatment of Cancer; ESR, erythrocyte sedimentation rate; NCCN, National Comprehensive Cancer Network; NCIC, National Cancer Institutes of Canada.

2003 through 2013 were retrospectively reviewed. Patients with Ann Arbor stage I or II disease, who were 18 years or older at the time of diagnosis, and who had a fusible initial PET-CT were included in the study. Because nodular lymphocyte-predominant HL is traditionally managed differently, all histologic subtypes of HL except for nodular lymphocyte-predominant HL were included. All patients with follow-up time ≤ 6 months were excluded from analysis unless they experienced progression or death, in which case they were counted as an event.

Patient, disease, and treatment characteristics

Baseline patient characteristics were evaluated. Bulky disease was defined as any nodal mass or conglomerate >10 cm in the axial, sagittal, or coronal dimensions. Disease was staged according to the Ann Arbor system and then subdivided into ESF, ESU, or advanced based on the German Hodgkin Study Group (GHSG) risk groupings.^{1,2} Per GHSG groupings, stage IIB patients with bulky disease or extranodal extension were classified as advanced. Because our cohort only included stage I and II patients, in the remainder of our results, advanced refers strictly to IIB bulky patients and will be referred to as IIB-advanced throughout this manuscript. Treatment-related information was recorded. Radiation therapy was designated as consolidative for those patients with a complete response to initial chemotherapy, as determined by the treating clinicians at that time.

Initial PET-CTs

PET-CT images for patients from January 2003 to December 2013 were analyzed. PET data, when acquired at our institution, were in 2D mode before January 2008 and in 3-dimensional (3D) mode after that date. ¹⁸F-DG PET-CTs obtained at our institution were acquired on 1 of 4 scanners: a DST machine, 2 DRX machines, or a DSTE machine (all from GE Healthcare, Milwaukee, WI). The corresponding CT scanners were 8-slice (DST model PET scanner), 16-slice or 64-slice (DRX model), or 64-slice machines (DSTE model). All PET-CT scanners at our institution used the same DISCOVERY platform by GE. An intravenous FDG injection of 555 to 629 MBq (15-17 mCi) or of 333 to 407 MBq (9-11 mCi) was administered for 2D and 3D imaging, respectively, and emission scans were acquired at 3 minutes per field of view. The injection-to-scan time of all patients was a median of 70 minutes

and an average of 75 minutes with a standard deviation of 17 minutes. PET images were reconstructed with standard vendor-provided reconstruction algorithms. Noncontrast-enhanced CT images, from the base of the skull to the mid-thigh, were acquired with the scanner in helical mode at a 3.75-mm slice thickness. All CT scans obtained at our institution were of diagnostic quality.

The PET-CT scanners at our institution are subject to a rigorous quality assurance/quality control program that entails daily checks for coincidences and single events mean and variance in addition to dead time, timing resolution, energy, and photo-multiplier tube gains on all detector responses for each scanner system. We also perform full scanner calibration and normalization on a quarterly basis along with American College of Radiology testing to ensure accurate scanner quantification. Annual testing is also performed based on the National Electrical Manufacturers Association NU2 standard for assessing resolution, sensitivity, count rate, scatter fraction, image quality, and accuracy. Finally, reconstruction parameters are optimized to ensure harmonization of SUV measurements between scanners.

Radiographic analysis

After image reconstruction, PET-CT images were transferred to MIM software, version 6.4.9 (MIM Software Inc, Cleveland, OH), and fused for further analysis. All SUV measurements reported in this work are based on patient body weight. Because no universal consensus has been reached on how to define MTV, we measured MTV on the initial PET-CT scans using a threshold method restricted to areas of disease with SUV ≥ 2.5 (MTV extracted from threshold volumes [MTV_t]).⁹ To account for areas of tumor that might not have significant uptake because of necrosis or other causes, we devised the soft-tissue method, in which the soft-tissue nodes or masses showing any SUV uptake were contoured and the 3D volume in cubic centimeters was designated as the MTV_{st}. TLG extracted from threshold volumes (TLG_t) or TLG manually contoured soft-tissue volumes (TLG_{st}) was calculated as mean SUV in the contoured regions \times the corresponding MTV. Representative contours from both methods of delineation are presented in Figure 1. The diameter of the longest nodal mass or conglomerate was measured for each patient in the axial, sagittal, and coronal dimensions.

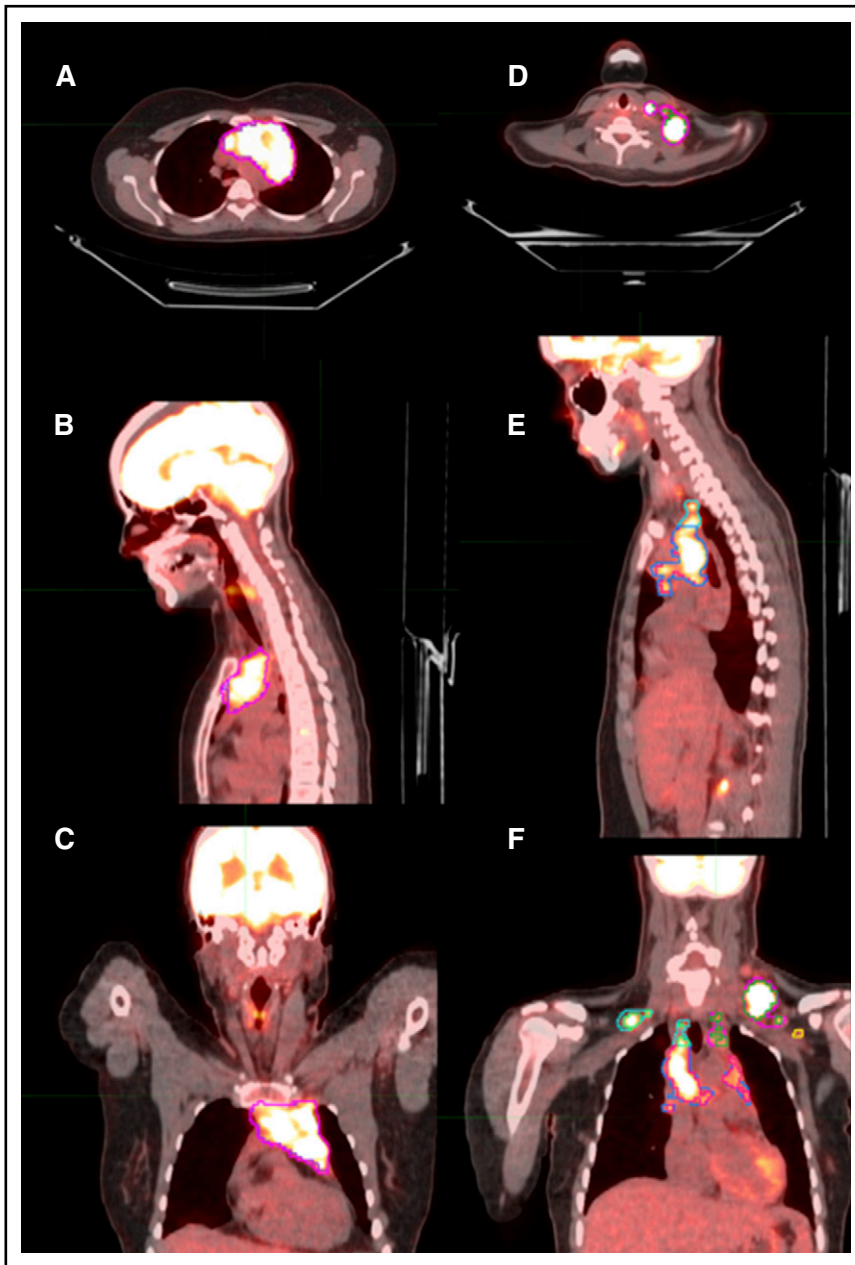


Figure 1. Representative delineation of MTV based on both the MTV_{st} and MTV_t contouring methods.

Axial, sagittal, and coronal sections might not show the same node but rather different regions of disease in each anatomical section. (A-C) Axial, sagittal, and coronal scans of mediastinal disease contoured based on MTV_{st} (magenta) and MTV_t (blue) methods. (D) Axial view of left cervical neck disease contoured based on MTV_{st} (magenta) and MTV_t (green) methods. (E) Sagittal view of left cervical neck disease contoured based on MTV_{st} (magenta) and MTV_t (green) methods and mediastinal disease contoured based on MTV_{st} (blue) and MTV_t (pink) methods. (F) Coronal view of left cervical neck disease contoured based on MTV_{st} (magenta) and MTV_t (green) methods, mediastinal disease contoured based on MTV_{st} (blue) and MTV_t (pink) methods, left axillary disease contoured based on MTV_{st} (yellow) and MTV_t (brown) methods, and right cervical neck disease contoured based on MTV_{st} (light blue) and MTV_t (light green) methods.

Data management

Study data were collected and managed by using Research Electronic Data Capture (REDCap) tools hosted at <http://redcap.mdanderson.org>.¹⁰ (REDCap is a secure, Web-based application designed to support data capture for research studies by providing (1) an intuitive interface for validated data entry, (2) audit trails for tracking data manipulation and export procedures, (3) automated export procedures for seamless data downloads to common statistical packages, and (4) procedures for importing data from external sources.)

Outcomes

The primary clinical outcome was freedom from progression (FFP), which was defined as the time from diagnosis to the time with relapsed or refractory disease. Cases in which persistent disease was identified during or within 90 days of completion of

up-front therapy were deemed refractory; disease that returned >3 months after up-front therapy was classified as relapsed. Patients who did not experience an event (refractory or relapsed disease) were censored at the date of the last follow-up or the date of death from other causes. Overall survival (OS) was defined as the time from diagnosis to death from any cause. Patients who did not experience an event were censored at the date of last known follow-up.

Statistical analysis

Categorical variables are reported as frequencies and percentages; continuous data are summarized as mean, median, and range. Both χ^2 and Fisher's exact tests were used to evaluate associations between categorical variables and study group. Wilcoxon's rank-sum test was used to compare the distributions of continuous variables (such as MTV and TLG) between the

Table 2. Patient and treatment characteristics

	No. of patients (%)	Median (range)
Sex		
Male	148 (55.4)	
Female	119 (44.6)	
GHSg disease classification		
Early favorable	74 (27.7)	
Early unfavorable	178 (66.7)	
IIB-advanced	15 (5.6)	
Ann Arbor disease stage		
IA	32 (12)	
IB	10 (3.7)	
IAE	1 (0.4)	
IIA	162 (60.7)	
IIB	52 (19.5)	
IIBE	2 (0.7)	
IIAE	8 (3.0)	
B-symptoms		
Present	66 (24.7)	
Absent	201 (75.3)	
ESR		
Normal	61 (22.8)	
Elevated	31 (11.6)	
Unknown	175 (65.5)	
Bulky disease		
Present	74 (27.7)	
Absent	193 (72.3)	
Extranodal disease		
Absent	254 (95.1)	
Present	13 (4.9)	
Chemotherapy regimens		
ABVD	239 (89.5)	
Other	28 (10.5)	
No. of chemotherapy cycles		
0	1 (0.4)	
2	18 (6.7)	
3	5 (1.9)	
4	108 (40.4)	
5	10 (3.7)	
Received consolidation RT		
Yes	187 (70.0)	
No	63 (23.6)	
NA	17 (6.4)	
Age at diagnosis, y	267	31.96 (18-95.4)
ESR, mm/h	92	29.5 (3-107)
Radiation dose (Gy)	183	30.6 (20-42)
No. of involved Ann Arbor sites	267	3 (1-10)

Other chemotherapy regimens besides ABVD included Adriamycin, hydroxydaunorubicin, and bleomycin or rituximab-ABVD.

ABVD, doxorubicin, bleomycin, vinblastine, and dacarbazine; NA, not available; RT, radiation therapy.

2 study groups. Kruskal-Wallis test was used to compare the distributions of continuous variables among the 3 GHSg subgroups (ESF, ESU, and IIB-advanced). Kaplan-Meier curves were produced according to the prognostic factors of interest (GHSg and categorized MTV_t and TLG). The log-rank test was used to test differences between the prognostic-factor groups. Univariate Cox proportional hazard models were used to determine the effects of potential prognostic factors on survival distributions (FFP and OS). Multivariable Cox proportional hazard models were used to examine the effect of MTV_t and TLG_t on FFP after adjusting for GHSg. Variable selection for the multivariable analysis was based on clinical interest and on the results from univariate analysis, with selection of covariates that were not collinear or minimized overfitting, and were based on the number of events. The 80th percentile values of MTV_t and TLG_t values were then used to dichotomize the continuous MTV_t and TLG_t variables into the 2-level categorical variables (high vs low).

Harrell's concordance (C-) index was used to measure the performance of the survival models.¹¹ The C-index can be interpreted as the probability of concordance between the predicted and observed survival times. A C-index of 1 indicates perfect prediction accuracy; a C-index of 0.5 is as good as a random predictor. To determine whether MTV or TLG added predictive information beyond GHSg alone, we used the *rcorr.cens* function and *U* statistics to test whether the difference in statistical predictive accuracy between the Cox regression models was significant. The biased-corrected C-index was calculated using a bootstrap internal validation procedure with 500 repeats. All tests were 2-sided. *P* < .05 indicates statistical significance. All analyses were conducted using SAS 9.3 (SAS, Cary, NC), S-Plus 8.0 (TIBCO Software Inc., Palo Alto, CA), and R 2.14.2 software (R Foundation).

Results

Patient, disease, and treatment characteristics

A total of 267 patients were identified who met the inclusion criteria; their baseline characteristics and treatment details are listed in Table 2. The median age at diagnosis was 32 (range, 18-95) years. Among all the qualifying patients, 178 (67%) were classified as ESU, 74 (28%) as ESF, and 15 (6%) as IIB-advanced. Forty-three patients (16%) were classified as having stage I disease and 224 (84%) as stage II. Sixty-six patients (25%) had B-symptoms at initial presentation, 61 (23%) had extranodal extension, and 74 (28%) had bulky disease. All but 1 patient, who was deemed unable to tolerate systemic therapy, received at least 2 cycles of chemotherapy. Most patients, 239 (89%) received ABVD. Seventeen patients (6%) with either refractory primary disease or disease progression received salvage treatment; thus, consolidation radiation therapy was not given. Among the remaining patients (*n* = 250), all of whom had had a complete response to chemotherapy, 187 (75%) received consolidative radiation therapy. The median dose prescribed to those who received radiation therapy was 30.6 Gy (range, 20-42 Gy).

Radiographic parameters

Means, medians, and ranges of the radiographic parameters measured are reported in Table 3. A total of 16.7% of the scans

Table 3. Radiographic parameters

	Mean	Median	Range
Total MTV _{st}	252.4	179.5	1.15-2 420.94
Total TLG _{st}	1284.2	1489.5	5.9-10 490.9
Total MTV _t	190.2	118.7	0-1 822.5
Total TLG _t	1195.7	733.0	1.95-9 937.4
Longest axial diameter of disease, cm	5.8	5.7	1.2-14.0
Longest sagittal diameter of disease, cm	7.1	6.9	1.1-18.9
Longest coronal diameter of disease, cm	6.1	5.3	1.3-17.7
Maximum SUV	13.2	12.6	3.2-50.9

were performed outside of our institution. In an effort to ensure these scans had similar quantitative performance to the studies done at our institution, we measured the liver mean and maximum SUV in a representative sample contour. The liver SUV measurements were similar for internal and external PET-CT scans.

Clinical outcomes

The 5-year OS rate was 95.5% (95% confidence interval [CI], 91.9-98.0) and the 5-year FFP rate was 90% (95% CI, 86.1-93.5). The median follow-up time was 4.96 years (range, 1.03-12.15 years) for living patients. Among the 267 patients evaluated, there was a total of 27 events: 10 patients were diagnosed with relapsed disease and with 17 with refractory disease.

We set out to identify patient- or treatment-related characteristics in addition to radiographic parameters that were associated with FFP. There was a high degree of correlation between the 2 MTV and TLG contouring methods and, given the greater objectivity as well as more prevalent use of the threshold method, we used MTV_t and TLG_t for the analysis. On univariate analysis (Table 4), factors associated with worse FFP were GHSG classification (IIB-advanced vs ESF: hazard ratio [HR], 7.56, $P = .008$; ESU vs ESF: HR, 2.89, $P = .086$), not receiving consolidation RT (HR, 4.71, $P = .016$), total MTV (for every 100-unit increase in MTV_t: HR, 1.72, $P < .0005$), total TLG (for every 500-unit increase in TLG_t: HR, 1.13, $P < .005$), and axial (HR, 1.17, $P = .032$), sagittal (HR, 1.11, $P = 0.047$), or coronal diameter (HR, 1.16, $P < .005$) of the longest node or nodal conglomerate. On multivariable Cox proportional hazard model, after adjusting for GHSG classification, total MTV_t (for every 100-unit increase: HR, 1.14; 95% CI 1.02-1.26; $P = .016$) and total TLG_t (for every 500-unit increase: HR, 1.096; 95% CI, 1.00-1.20; $P = .047$) were strongly associated with FFP. Because the GHSG classification has been used in numerous randomized clinical trials to assess the risk of treatment failure in patients with HL, we next assessed whether

adding MTV and TLG improved the predictive accuracy for FFP. Cox regression models revealed better statistical predictive accuracy for FFP when total MTV_t (bias-corrected C-index for GHSG + MTV_t, 0.6, $P = .056$) or total TLG_t (bias-corrected C-index for GHSG + TLG_t, 0.67, $P = .069$) were added to the model compared with the GHSG classification alone (bias-corrected C-index, 0.61). C-indexes comparing GHSG + MTV_t vs GHSG + TLG_t were not statistically different ($P = .603$), showing that both functional parameters add a similar level of predictive ability for FFP.

Tumor burden and clinical outcomes

Given the added value of MTV and TLG to the predictive ability of GHSG, we then set out to investigate the predictive value of these parameters independently. We used 80th percentile values as the cutoffs to dichotomize continuous MTV_t and TLG_t and assigned all patients into high-MTV_t (≥ 268) or low-MTV_t (< 268) subgroups and high-TLG_t (≥ 1703) and low-TLG_t (< 1703) subgroups. Fifty-three patients fell under the high-MTV_t/TLG_t categories and 214 under the low-MTV_t/TLG_t categories. Patients with IIB-advanced disease were more likely to have high MTV_t ($P < .001$), high TLG_t ($P < .001$), bulky disease ($P < .001$), higher RT doses ($P < .001$), and larger axial, sagittal, and coronal lengths of the longest tumor mass ($P < .001$).

Next, to explore differences between patient subgroups based on the 80th percentile cutoff values, univariate analysis showed that patients with high MTV_t or high TLG_t had worse FFP than did those patients with low MTV_t (HR, 3.09; 95% CI, 1.43-6.65; $P = .004$) and low TLG_t (HR, 3.65; 95% CI, 1.71-7.79; $P = .001$). In multivariable Cox models including GHSG and total MTV_t or TLG_t, GHSG grouping was not significantly associated with outcome, but total MTV_t categorization of high vs low (HR, 2.20; 95% CI, 0.92-5.25; $P = .076$) and TLG_t categorization of high vs low (HR, 2.822; 95% CI, 1.23-6.48; $P = .014$) correlated with worse FFP.

We then examined FFP (Figure 2) and OS (Figure 3) according to high vs low MTV_t and high vs low TLG_t for each GHSG subgrouping by using Kaplan-Meier plots. Among patients with ESU, FFP was significantly worse for those with high MTV_t ($P = .008$) or high TLG_t ($P = .001$) than for those with low MTV_t and TLG_t (Figure 2B-C). In this same cohort of ESU patients, worse OS may also have been present in those with high MTV_t ($P = .089$) and high TLG_t ($P = .087$) (Figure 3B-C).

When we compared the FFP stratified by GHSG in combination with MTV_t or TLG_t volume, there was a clear difference between FFP of ESU low MTV_t patients compared with ESU high MTV_t patients ($P < .001$), whereas there was no difference between ESF and ESU low MTV_t patients ($P = .3$) or high MTV_t and IIB-advanced patients ($P = .7$). The same statistical significance applied when we compared FFP of ESU low TLG_t with those of ESU high TLG_t ($P < .001$) (Figure 4A), but no difference when comparing ESF and ESU low TLG_t or ESU high TLG_t ($P = .4$) and IIB-advanced patients ($P = .9$) (Figure 4B). We then decided to compare the FFP in all GHSG groups between the low MTV_t and high MTV_t groups. The high MTV_t patients did significantly worse when compared with low MTV_t patients ($P < .0001$) (Figure 5A). Patients with high TLG_t also had a

Table 4. Univariate Cox proportional hazards model for predicting event-free survival

Potential prognostic factor	HR (lower limit-upper limit)	P
Categorical variables		
Sex		
Female vs male	0.84 (0.3941-1.7839)	.6473
Consolidation RT		
No vs yes	4.71 (1.3281-16.6805)	.0164
GHSG classification		
IIB-advanced vs early-favorable	7.56 (1.6917-33.7941)	.0081
Early-unfavorable vs early-favorable	2.89 (0.8589-9.7276)	.0865
Ann Arbor disease stage		
I vs II	0.19 (0.0260-1.4101)	.1046
B-symptoms		
No vs yes	0.53 (0.2418-1.1540)	.1095
Bulky disease		
No vs yes	0.53 (0.2447-1.1361)	.1022
Extranodal disease		
No vs yes	1.31 (0.1774-9.6295)	.7927
Chemotherapy regimen		
ABVD vs other	0.93 (0.2792-3.0789)	.9017
Continuous variables		
Age at diagnosis	1.01 (0.9823-1.0305)	.6174
ESR	0.99 (0.9763-1.0242)	.9966
Radiation dose	1.32 (1.0929-1.5910)	.0039
Total MTV _{st}	1.00 (1.0006-1.0019)	.0002
Total TLG _{st}	1.00 (1.0001-1.0004)	.0004
Total MTV _t	1.00 (1.0007-1.0025)	.0004
Total MTV _t 100-unit increase	1.17 (1.0735-1.2793)	.0004
Total TLG _t	1.00 (1.0001-1.0004)	.0011
Total TLG _t 500-unit increase	1.13 (1.0514-1.2228)	.0011
Longest axial diameter of disease	1.17 (1.0139-1.3604)	.0320
Longest sagittal diameter of disease	1.11 (1.0013-1.2341)	.0471
Longest coronal diameter of disease	1.17 (1.0618-1.2872)	.0015
Maximum SUV	1.02 (0.9633-1.0781)	.5100

significantly worse FFP compared with those with low TLG_t (Figure 5B).

Discussion

We showed that functional 3D measurements of tumor burden, namely MTV and TLG, can be valuable tools in predicting FFP and OS outcomes in patients with HL. Total MTV_t and TLG_t correlated significantly with FFP. Further, GHSG groupings had better predictive value when MTV_t and TLG_t were incorporated into the model. Patients with MTV_t >268 and TLG_t >1703 in our group had worse FFP rates, shorter FFP times, and were more likely to have bulky disease and IIB-advanced stage disease. Our results closely correlated with previous work by Kanoun et al, who had shown that all methods of MTV determination correlated with outcome in their cohort of 59 HL patients stage I-IV and had

established a MTV_t cutoff of 432 with the SUV >2.5 method.¹²

We have also demonstrated that even when the GHSG classification alone could not significantly predict outcome in our multivariate model, high MTV or TLG correlated with worse FFP. Upon examining the various GHSG subgroups divided according to MTV, we were able to distinguish 2 separate groups in the ESU category: FFP was significantly worse for patients with ESU disease who had high MTV or TLG than for patients with ESU disease and low MTV or TLG. This finding allows us to substratify patients with ESU into 2 distinct categories: early-stage low-risk unfavorable and early-stage high-risk unfavorable. This substratification opens the door for further classifying ESU patients into those who might benefit from treatment escalation (high-risk unfavorable group) and those who have excellent outcomes based on our current treatment paradigms (low-risk group). It is, however, too early to interpret these data as need of treatment deescalation in the ESU low-risk group even though patients' outcomes are nearly as excellent as ESF because they received more intensive treatment, such as higher doses of radiation therapy. If those patients were treated with less aggressive regimens in line with ESF patients they might not do as well, which could be a topic for future prospective trials.

HL continues to present a therapeutic challenge. Although most cases of early-stage disease are highly curable with excellent outcomes, a smaller subset of patients continues to experience treatment failure despite combined modality therapy. Even in the seminal GHSG HD10 trial, patients with ESF disease who received the most extensive therapy (ABVD ×4, 30 Gy) had a failure rate of 11.6%.² Numerous studies have been performed to date with the aim of identifying patients at highest risk of relapse in the hopes that such patients can experience improved outcomes with escalation of therapy, whereas others can safely undergo deescalation of therapy. In addition to the trials resulting in the clinical groupings in Table 1, other trials have attempted to incorporate findings from functional imaging such as PET-CT into the treatment paradigm. Importantly, most such studies have focused on the role of early/interim PET-CT as a marker of overall response and future outcomes. The H10 trial⁴ randomized patients with stage I/II HL who had negative early PET scans to receive 1 or 2 additional cycles of chemotherapy in lieu of radiation therapy; this trial was closed early because of 17 additional events in the chemotherapy-only arm. Similar results were seen in the United Kingdom National Cancer Research Institute's UK PET Scan in Planning Treatment in Patients Undergoing Combination Chemotherapy For Stage IA or Stage IIA Hodgkin Lymphoma trial,¹³ with the chemotherapy-only arm experiencing a 4% detriment in progression-free survival, which resulted in failure to show noninferiority of chemotherapy-only based on a negative PET scan. Moreover, patients who experience relapse or failure after initial treatment can undergo salvage chemotherapy, stem cell transplantation, and possibly even higher doses of radiation therapy. Therefore, additional tools that will allow us to accurately determine which patients can safely undergo deescalation of therapy can make significant contributions to the management of HL.

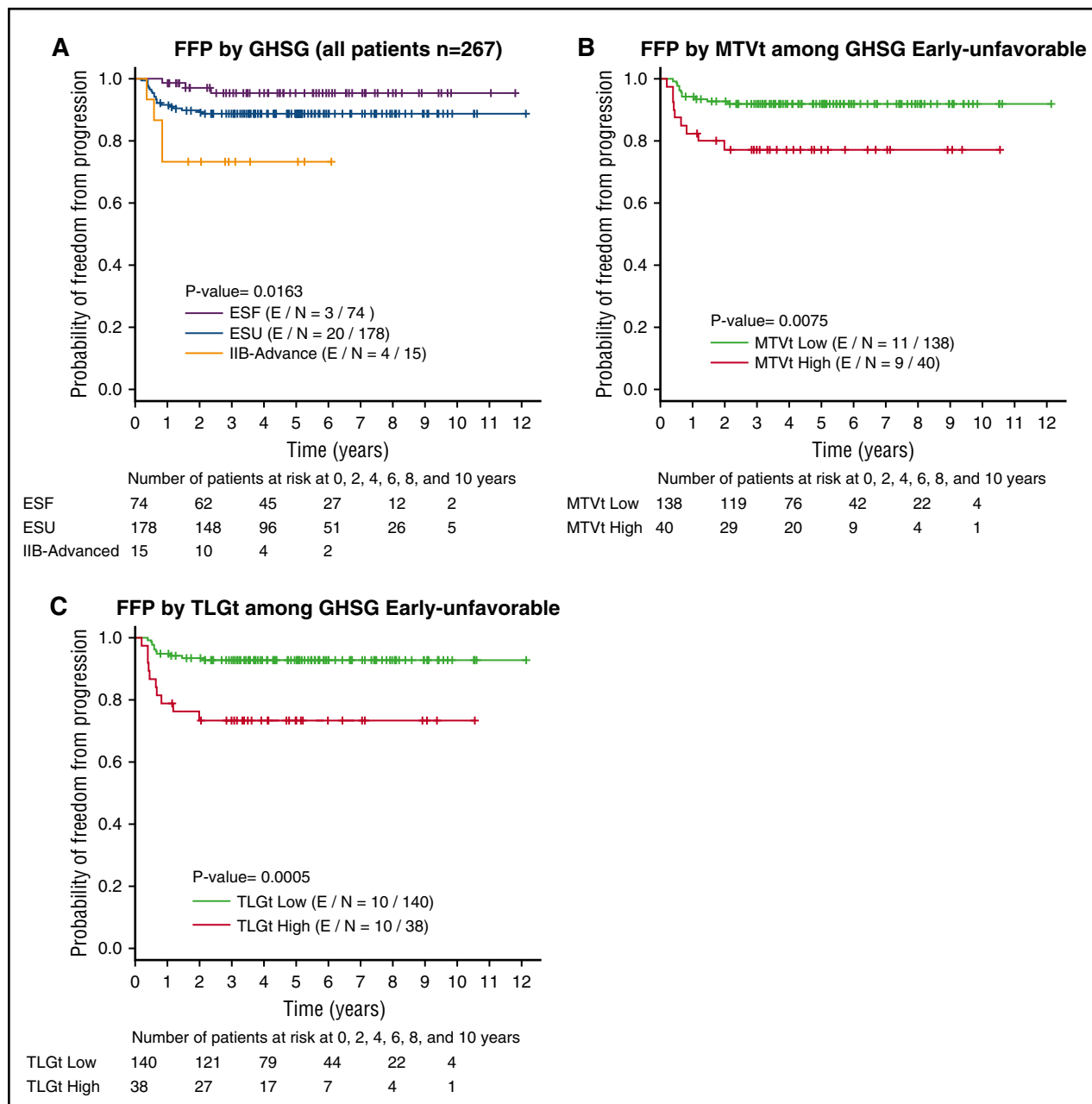


Figure 2. Kaplan-Meier plots for FFP survival. (A) All patients according to GHSG groupings; (B) MTV_t-low vs MTV_t-high among patients with ESU disease; (C) TLG_t-low vs TLG_t-high among patients with ESU disease. E/N, events/total N.

In these 2 trials, a negative interim PET-CT did not identify early-stage HL patients for whom RT could be safely omitted without an increased risk of progression; therefore, we set out to examine any possible role the initial PET-CT might have in predicting outcome. We have now shown that functional 3D measurement of tumor burden, namely MTV and TLG, obtained from the initial PET-CT can be valuable tools in predicting FFP and OS.

Our study had several limitations that must be addressed in future studies. The retrospective nature of our investigation and the small number of events limited our analyses. Our cohort comprised mostly ESU patients, with small patient populations with ESF and IIB-advanced disease. Additionally,

the cutoff values for MTV and TLG were obtained from a single institutional dataset and require external validation before they can be used for clinical decision-making. Validation of these parameters in larger, multi-institutional cohorts will allow more accurate determination of values that can be used clinically. Another limitation of our study is that not all PET imaging was acquired using the same scanner. The use of various scanners might influence the data and the overall results. The majority of scans (84%) were performed at our institution, where scanners are assessed by regular, rigorous quality assurance and are optimized for minimal scanner-to-scanner variation. The remaining scans (16%) were performed at outside institutions; therefore, there may be more variability in image acquisition and processing. To

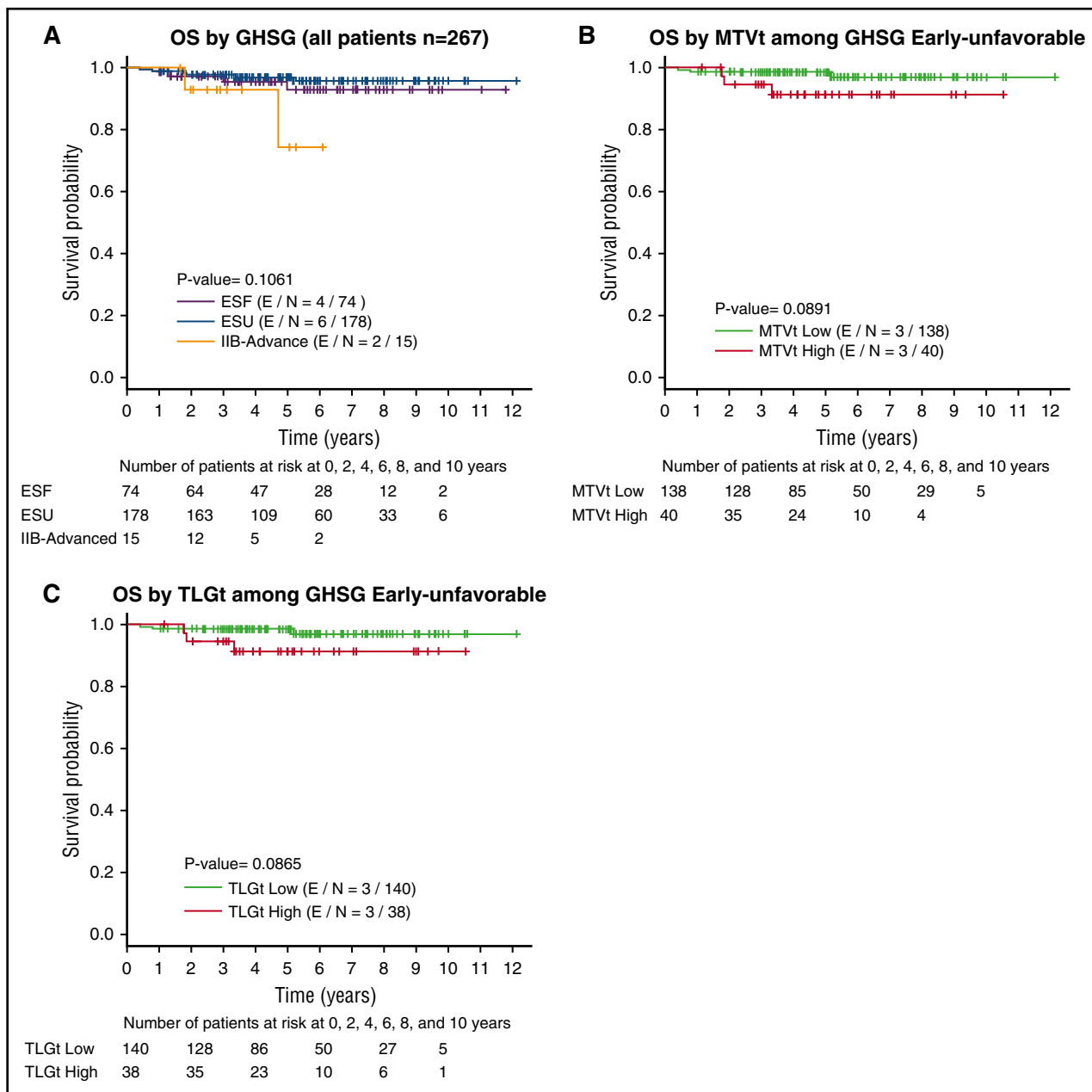


Figure 3. Kaplan-Meier plots for OS. (A) All patients according to GHSG grouping; (B) for MTV_t-low vs MTV_t-high among patients with ESU disease; (C) for TLG_t-low vs MTV_t-high among patients with ESU disease.

assess for major differences between the internal and external scans, we evaluated the PET reconstruction parameters and measured the SUV of the liver. We did not determine significant differences; thus, we do not think that the inclusion of a minority of scans performed outside of our institution significantly affected our results. Nonetheless, validation of these findings, with PET scans obtained using a standardized approach, is recommended. Another potential limitation of this study is the pooling of results from PET studies acquired in 2D and 3D modes. This was primarily the result of the scans from 2003 to 2013, in which the PET technology evolved from 2D to 3D. Although, theoretically, PET SUV measurements should not be affected by the acquisition modes (2D vs 3D), particularly if scatter correction is accurately accounted for

(which is the case in this study because all 3D PET scans were reconstructed using model based scatter correction techniques), there is still the potential for the maximum SUV measurement to be affected because it relies on a single pixel value.

A separate issue is the contouring method for MTV and TLG. To date, no standard technique has been agreed upon, although numerous limits have been proposed for defining metabolically active tumors.^{8,14,15} Most of the techniques in current use involve the threshold method, in which an SUV above a certain threshold constitutes active disease. However, no consensus has been reached as to what the exact threshold should be; some studies used different cutoff

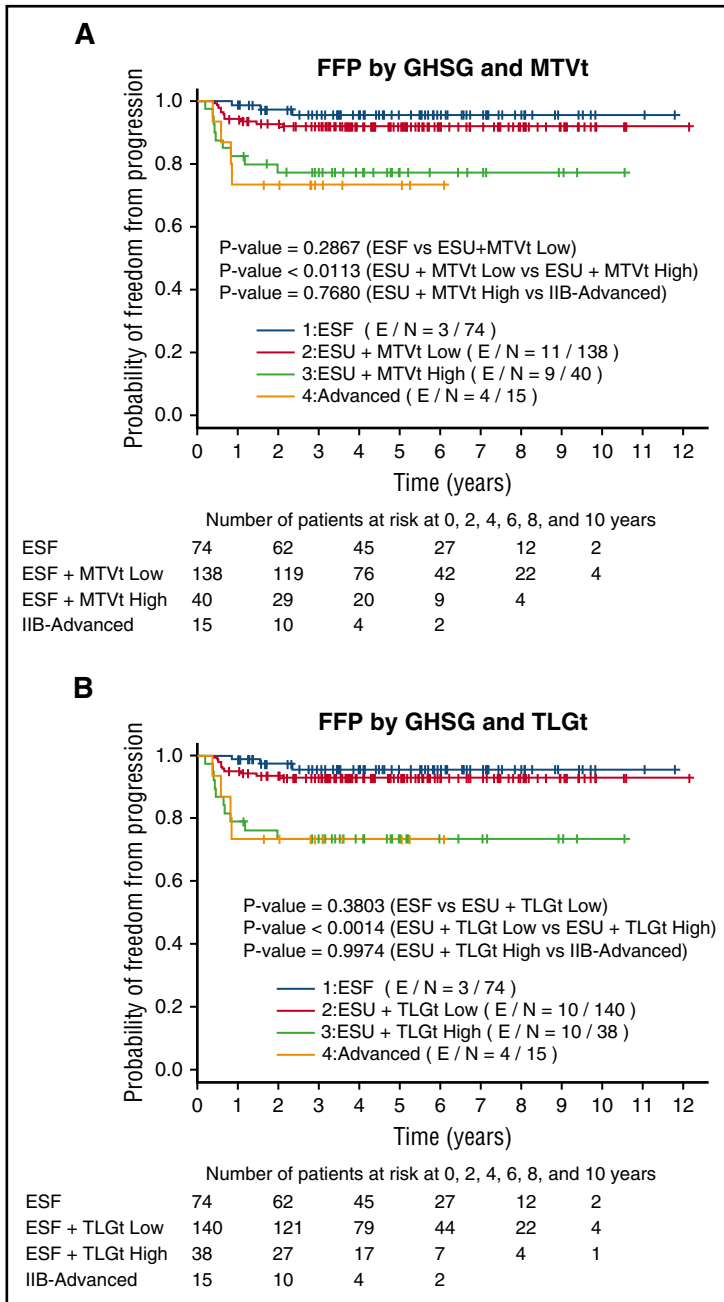


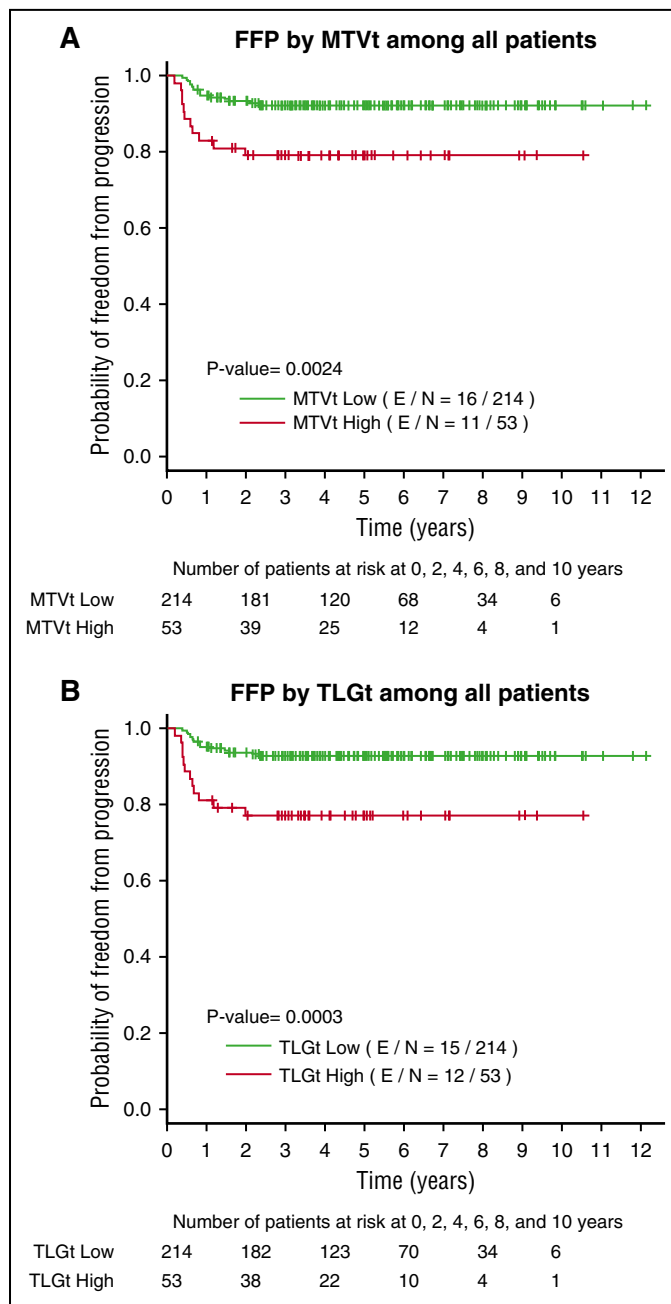
Figure 4. FFP graphed for GHSG ESF, ESU low MTV_t and high MTV_t, and IIB-advanced and GHSG ESF, ESU low TLG_t and high TLG_t, and IIB-advanced. (A) FFP between ESU low MTV_t and high MTV_t ($P < .02$) as well as low TLG_t and high TLG_t ($P < .002$) differs significantly, whereas there is no difference between ESF and ESU low MTV_t ($P = .3$) or ESU high MTV_t and IIB-advanced ($P = .8$). (B) There is also no difference between ESF and ESU low TLG_t ($P = .4$) and ESU high TLG_t and IIB-advanced ($P = 1$).

values for SUV, such as 2.5,¹⁶ a threshold of 41% of the maximum SUV,¹⁷ or background activity thresholds from the liver or the mediastinal blood pool. Because none of these methods has proven consistently superior to others methods, we chose to use the common SUV threshold of 2.5 and obtained a second set of values by contouring all of the soft-tissue components of the disease regardless of the SUV. We chose this second method because many large HL tumors can have necrotic areas in which the SUV is <2.5. Both methods correlated well with outcomes, and thus for further subanalyses we used the threshold method (SUV >2.5) results because it is more objective than manual contouring and has been used more frequently in other studies. Our cutoff values require validation in an external data set. Notably, however, our cutoff for MTV is close

to the only other known predictive cutoff of MTV_t in HL: 431.¹²

In conclusion, our findings, from 1 of the largest single-institution HL databases in the modern era of PET-CT, have shown that MTV and TLG, 2 measures of functional imaging available from baseline PET-CT scans, can aid in predicting which patients with early-stage HL will have worse outcomes by adding measurements that were not previously available for categorizing patients with HL. Most important, we have shown that not all cases of ESU HL are the same. In fact, 2 distinct categories can be discerned by the MTV or TLG: low and high disease burdens. Future studies will be needed to confirm these findings, validate our cutoff thresholds for MTV and TLG, and assess the clinical relevance of more accurately risk-stratifying ESU HL patients.

Figure 5. FFP survival graphed for patients in all groups. (A) Low MTV_t and high MTV_t and (B) low TLG_t and high TLG_t. FFP is significantly worse among those with high MTV_t compared with low MTV_t ($P = .002$) and those with high TLG_t compared with low TLG_t ($P = .0003$).



Acknowledgments

This work was supported in part by the National Institutes of Health National Cancer Institute, Cancer Center Support (Core) (grant CA 016672) to The University of Texas MD Anderson Cancer Center. No other funding was received for design, completion, or analysis of this study.

Authorship

Contribution: M.A., S.A.M., J.P.R., C.C.P., and B.S.D. designed the research, collected the data, and wrote the paper; W.D. analyzed the data and wrote the paper; G.L.S. designed the research and wrote the paper; O.M. designed the research and collected the data; Z.A.Y. J.G., E.M.O., and T.Y.A. collected the data; C.F.W. wrote the paper; and E.R., N.G., H.C., J.D.K., Y.O., and M.F. designed the research.

Conflict-of-interest disclosure: N.G. is owner of Garglet LLC, a medical informatics software company. The remaining authors declare no competing financial interests.

Correspondence: Bouthaina S. Dabaja, Department of Radiation Oncology, Unit 97, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030; e-mail: bdabaja@mdanderson.org.

Footnotes

Submitted 3 April 2017; accepted 3 October 2017. Prepublished online as *Blood* First Edition paper, 16 October 2017; DOI 10.1182/blood-2017-04-773838.

There is a *Blood* Commentary on this article in this issue.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

REFERENCES

- Eich HT, Diehl V, Gorgen H, et al. Intensified chemotherapy and dose-reduced involved-field radiotherapy in patients with early unfavorable Hodgkin's lymphoma: final analysis of the German Hodgkin Study Group HD11 trial. *J Clin Oncol*. 2010;28(27):4199-4206.
- Engert A, Plutschow A, Eich HT, et al. Reduced treatment intensity in patients with early-stage Hodgkin's lymphoma. *N Engl J Med*. 2010;363(7):640-652.
- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA Cancer J Clin*. 2015;65(1):5-29.
- Andre 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.
- Lister TA, Crowther D, Sutcliffe SB, et al. Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. *J Clin Oncol*. 1989;7(11):1630-1636.
- Specht L, Nordentoft AM, Cold S, et al. Tumor burden as the most prognostic factor in early stage Hodgkin's disease. *Cancer*. 1988;61(8):1719-1727.
- Berkowitz A, Basu S, Srinivas S, Sankaran S, Schuster S, Alavi A. Determination of whole-body metabolic burden as a quantitative measure of disease activity in lymphoma: a novel approach with fluorodeoxyglucose-PET. *Nucl Med Commun*. 2008;29(6):521-526.
- Kim TM, Paeng JC, Chun IK, et al. Total lesion glycolysis in positron emission tomography is a better predictor of outcome than the International Prognostic Index for patients with diffuse large B cell lymphoma. *Cancer*. 2013;119(6):1195-1202.
- Freudenberg LS, Antoch G, Schutt P, et al. FDG-PET/CT in re-staging of patients with lymphoma. *Eur J Nucl Med Mol Imaging*. 2004;31(3):325-329.
- Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381.
- Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.
- 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(10):e0140830.
- 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.
- Galicchio R, Mansueto G, Simeon V, et al. F-18 FDG PET/CT quantization parameters as predictors of outcome in patients with diffuse large B-cell lymphoma. *Eur J Haematol*. 2014;92(5):382-389.
- Hussien AE, Furth C, Schonberger S, et al. FDG-PET response prediction in pediatric Hodgkin's lymphoma: impact of metabolically defined tumor volumes and individualized SUV measurements on the positive predictive Value. *Cancers (Basel)*. 2015;7(1):287-304.
- Hyun SH, Choi JY, Kim K, et al. Volume-based parameters of (18)F-fluorodeoxyglucose positron emission tomography/computed tomography improve outcome prediction in early-stage non-small cell lung cancer after surgical resection. *Ann Surg*. 2013;257(2):364-370.
- Lee P, Weerasuriya DK, Lavori PW, et al. Metabolic tumor burden predicts for disease progression and death in lung cancer. *Int J Radiat Oncol Biol Phys*. 2007;69(2):328-333.
- Meyer RM, Gospodarowicz MK, Connors JM, et al; Eastern Cooperative Oncology Group. Randomized comparison of ABVD chemotherapy with a strategy that includes radiation therapy in patients with limited-stage Hodgkin's lymphoma: National Cancer Institute of Canada Clinical Trials Group and the Eastern Cooperative Oncology Group. *J Clin Oncol*. 2005;23(21):4634-4642.
- National Comprehensive Cancer Network. NCCN guidelines on Hodgkin lymphoma. 2016 6/21/2016; 3.2016. https://www.nccn.org/professionals/physician_gls/pdf/hodgkins.pdf. Accessed 3 June 2016.