Interobserver Agreement of Interim and End-of-Treatment ^{18}F-FDG PET/CT in Diffuse Large B-Cell Lymphoma: Impact on Clinical Practice and Trials

Coreline N. Burggraaff[1], Alexander C. Cornelisse[1], Otto S. Hoekstra[2], Pieternella J. Lugtenburg[3], Bart De Keizer[4], Anne I.J. Arens[5], Filiz Celik[6], Julia E. Huijbregts[7], Henrica C.W. De Vet[8], and Josée M. Zijlstra[1] on behalf of the HOVON Imaging Working Group

[1]Department of Hematology, VU University Medical Center, Cancer Center Amsterdam, Amsterdam, The Netherlands; [2]Department of Radiology and Nuclear Medicine, VU University Medical Center, Cancer Center Amsterdam, Amsterdam, The Netherlands; [3]Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands; [4]Department of Radiology and Nuclear Medicine, University Medical Center Utrecht, Utrecht, The Netherlands; [5]Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands; [6]Department of Radiology and Nuclear Medicine, Deventer Ziekenhuis, Deventer, The Netherlands; [7]Department of Radiology and Nuclear Medicine, Gelre Ziekenhuis, Apeldoorn, The Netherlands; and [8]Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

We aimed to assess the interobserver agreement of interim PET (I-PET) and end-of-treatment PET (EoT-PET) using the Deauville score (DS) in first-line diffuse large B-cell lymphoma (DLBCL) patients. **Methods:** I-PET and EoT-PET scans of DLBCL patients were performed in the HOVON84 study (2007–2012), an international multicenter randomized controlled trial. Patients received R-CHOP14 and were randomized to receive rituximab intensification in the first 4 cycles or not. I-PET was performed after 4 cycles (for observational purposes), and EoT-PET after 6 or 8 cycles. Two independent central reviewers retrospectively scored all scans according to the DS system, masked to clinical outcomes. Results were dichotomized as negative (DS of 1–3) or positive (DS of 4–5). Besides percentage overall agreement (OA), we calculated agreement for positive and negative scores, expressed as positive agreement (PA) and negative agreement (NA), respectively. **Results:** 465 I-PET and 457 EoT-PET scans were centrally reviewed; baseline ^{18}F-FDG PET or PET/CT was available in 75%–77%, and CT in the remaining cases. Percentage OA for I-PET and EoT-PET were 87.7% and 91.7% ($P = 0.049$), with NA of 92.0% and 95.0% ($P = 0.091$), and PA of 73.7% and 76.3% ($P = 0.656$), respectively. **Conclusion:** Interobserver agreement using DS in DLBCL patients in I-PET and EoT-PET yields high OA and NA. The lower PA suggests that EoT-PET/CT treatment evaluation in daily practice and I-PET–adapted trials may benefit from dual reads and central review, respectively.

Key Words: observer variation; positron emission tomography; DLBCL; Deauville score; Lugano criteria

J Nucl Med 2018; 59:1831–1836
DOI: 10.2967/jnumed.118.210807

Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of malignant lymphoma, accounting for 30%–40% of non-Hodgkin lymphomas (*1*). Current international guidelines (*2,3*) recommend ^{18}F-FDG PET before therapy in typically ^{18}F-FDG–avid lymphoma types—for example, Hodgkin lymphoma and DLBCL (*4*)—and to apply the Lugano response classification based on the Deauville score (DS) on a 5-point scale at the end of treatment. Application of ^{18}F-FDG PET during therapy (interim-PET, or I-PET) allows PET-guided patient management, with success in Hodgkin lymphoma (*5–8*). In DLBCL, the value of I-PET is less clear (*9*): most I-PET–adapted trials in DLBCL did not demonstrate a strategy that overcomes treatment resistance (*10*), except for a phase II study with intensification after rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) at a 14-d cycle (R-CHOP14) in I-PET–positive patients to R-ICE and Z-BEAM autologous stem cell transplantation (*11*). Therefore, I-PET is currently not used in clinical practice. An important prerequisite for these PET-guided studies is a consistent classification of the I-PET scans into a positive or negative category.

Similar to other disciplines, observer variation is the Achilles' heel of radiology (*12*). In the DS scoring system, the ^{18}F-FDG uptake in potentially malignant tissue is rated versus normal ^{18}F-FDG distribution in mediastinal blood pool and liver. Such a semiquantitative approach is less prone to observer variation than visual readings purely based on perception, knowledge, experience, and pattern recognition (*12*), possibly influenced by optical illusion effects (*13*). There are few studies on interobserver agreement of DS in PET scans in DLBCL patients treated with rituximab-containing chemotherapy, reporting a 0.4–0.8 range of κ-values for I-PET (*14–16*) and 0.5 for DS in end-of-treatment PET (EoT-PET) (*15*).

In clinical practice and trials, it is essential to know the specific agreement, that is, the absolute probability of obtaining the same test result by different reviewers rating the same scan. In cases of I-PET–driven treatment escalation on a positive I-PET scan or I-PET–driven treatment deescalation on a negative I-PET scan, observer variation–driven misclassification might compromise the

Received Mar. 5, 2018; revision accepted Apr. 24, 2018.
For correspondence or reprints contact: Josée M. Zijlstra, Department of Hematology, VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands.
E-mail: j.zijlstra@vumc.nl
Published online May 4, 2018.

results of clinical trials or induce overtreatment or undertreatment, respectively.

Our primary objective was to assess the interobserver agreement of I-PET and EoT-PET using DS in a large randomized clinical trial in DLBCL patients. Secondary objectives were to identify potential sources of observer variation (timing of PET—that is, I-PET vs. EoT-PET; baseline imaging modality—CT vs. PET or PET/CT, and the site of residual tracer uptake). The results are reported in accordance with Guidelines for Reporting Reliability and Agreement studies (*17*).

MATERIALS AND METHODS

Study Population

I-PET and EoT-PET scans were collected from the HOVON84 study, an international multicenter phase 3 trial in DLBCL (EudraCT 2006-005174-42). Patients were enrolled from 69 hospitals in The Netherlands, Belgium, and Denmark between November 2007 and April 2012. The main inclusion criteria were newly diagnosed, histologically proven, CD-20–positive DLBCL patients with Ann Arbor stage II–IV, age 18–80 y, and World Health Organization performance status 0–2. Exclusion criteria were primary central nervous system, testicular, transformed indolent, and primary mediastinal B-cell lymphoma, as well as posttransplant lymphoproliferative disease. Patients were randomized between standard R-CHOP14 or R-CHOP14 with intensification of rituximab in the first 4 cycles (R2-CHOP14). Administration of granulocyte colony-stimulating factor was mandatory and served to oppose the neutropenic side effects of the R-CHOP14 scheme. Six milligrams of pegfilgrastim were injected subcutaneously on day 2 of each R-CHOP cycle. I-PET and EoT-PET were performed after 4 cycles of therapy for observational purposes only, and after 6 (patients aged $>$ 65 y) or 8 cycles (patients $\leq$ 60 y), respectively. Baseline PET was recommended but not mandatory. HOVON84 has been approved by the institutional review board, and all subjects signed an informed consent form for use of their data for scientific purposes.

Image Analysis

DS was used for central image review (*2*). Between 2013 and 2016, each I-PET and EoT-PET scan was read independently by 2 reviewers from a pool of 10, who randomly drew scans from the image warehouse. All PET and CT scans were anonymized and uploaded to a database server hosted by Keosys (Imagys), allowing reviewers to read the images in their own workspaces. Seven percent of the PET scans performed in the HOVON84 trial were done with dedicated PET scanners, but this analysis of interobserver agreement was limited to PET/CT examinations. Reviewers were experienced nuclear medicine physicians (>5 y of experience with response evaluation of lymphoma in academic or large peripheral hospitals), actively participating in the HOVON Imaging Working Group. They were masked to clinical follow-up and randomization arm. Reviewers had access to all baseline imaging data (electronic case records containing clinical and imaging staging information provided by local clinicians and image reviewers). For the trial, discrepancies between the 2 reviewers were adjudicated by a third reviewer.

Reviewers used an electronic case record with prespecified nodal localizations (specifying regions as Waldeyer's ring, cervical, supraclavicular, axillary, mediastinum, hilar, paraaortic, mesenteric, spleen, iliac, inguinal, and other) and extranodal locations (gastrointestinal, central nervous system, skin, liver, lung, pleural, skeletal, and other). Open text fields were available for explanation of difficulties in reading. Reviewers assigned a DS for individual nodal and extranodal localizations together with a final patient-based score (highest lesional DS). We analyzed the DS of I-PET and EoT-PET as ordinal as well as dichotomized scores (DS 1–3 considered negative, DS 4–5 positive) (*2*).

Statistical Analysis

We performed patient- and region-based analyses. Besides the percentage overall agreement (OA), we calculated the percentage specific agreement, separating positive agreement (PA) from negative agreement (NA). PA and NA were defined as the probability that, if one reviewer assigns a positive or negative score, respectively, a second reviewer scores positive or negative as well (*18*). The prevalence of positive scans was calculated as the sum of the number of scans in which both reviewers scored positive and half the scans with discrepancies divided by the total number of scans. We analyzed the following potential sources of observer variation: I-PET and EoT-PET; availability of a baseline PET, PET/CT, or CT scan for reference; and residual ^{18}F-FDG uptake in different nodal and extranodal localizations. Discrepancies in these specific sites were related to baseline lymphoma prevalence, to assess which localizations were most difficult to read. In addition, we checked the assumption that there was no difference in observer variation between the control and intervention arms. For comparison of the percentage OA, PA, and NA between groups, the χ^2 test was used. A P value of less than 0.05 was considered statistically significant. Statistical analyses were performed using SPSS Statistics (IBM, version 20).

RESULTS

Study Population

In total, 575 patients were eligible for the final analysis of the main trial (*19*). I-PET and EoT-PET evaluation was performed in 534 and 517 patients, respectively (Supplemental Figs. 1 and 2; supplemental materials are available at http://jnm.snmjournals.org). For 7 patients, no PET data were received from the hospitals, 38 I-PET and 34 EoT-PET scans were performed on a stand-alone PET scanner, and 11 I-PET and 7 EoT-PET scans were not accessible in DICOM format or contained incomplete series.

I-PET

Table 1 summarizes the results of the interobserver agreement for I-PET with dichotomized DS. We obtained 465 evaluable scan results, because in 13 I-PET scans one of the reviewers did not provide a DS rating. The median time interval of I-PET scanning after the last chemotherapy cycle was 11 d (interquartile range, 9–13 d). In 408 of 465 scans, the reviewers agreed on the final conclusion (negative or positive), yielding a percentage OA of 87.7% (95% confidence interval [95%CI], 84.7–90.8). The prevalence of positive I-PET scans was 23.3%. The NA, at 92% (95%CI, 89.1–95.0), was markedly higher than the PA, at 73.7% (95%CI, 65.0–82.5).

A baseline ^{18}F-FDG PET or PET/CT scan was available in 77% (n = 349 integrated PET/CT scan and n = 8 PET stand-alone with a separate CT scan), and diagnostic CT in the remaining cases (n = 108). Percentage OA, NA, and PA were not statistically significant between these groups (percentage OA, 88% and 87%, P = 0.799; NA, 92% and 92%, P = 0.947; PA, 75.7% and 65%; P = 0.347). Percentage OA was similar in both treatment arms (P = 0.606).

For ordinal DSs, the reviewers agreed in 214 of 465 cases (Supplemental Table 1), resulting in 46% exact agreement (95%CI, 41.4–50.7). Percentage agreement was 78.3% (95%CI, 74.4–82.1) when we allowed a 1-point difference—except for a discrepancy between scores 3 and 4—between the reviewers' scorings.

Table 2 presents the percentage OA for the specific nodal and extranodal localizations for dichotomized DS, related to the baseline prevalence. Gastrointestinal, Waldeyer's ring, skeletal, spleen, and mesenteric sites showed a relatively large number of discrepancies.

TABLE 1
Interobserver Agreement on Dichotomized DS, by Baseline Modality

	I-PET/CT				EoT-PET/CT			
Agreement	Total (n = 465)	Baseline CT only (n = 108)	Baseline ^{18}F-FDG PET or PET/CT (n = 357)	P*	Total (n = 457)	Baseline CT only (n = 114)	Baseline ^{18}F-FDG PET or PET/CT (n = 343)	P*
Positivity†	23.3	18.5	24.8		17.5	18.0	17.3	
Percentage OA	87.7 (84.7–90.8)	87.0 (80.2–93.8)	88.0 (84.4–91.5)	0.799	91.7 (89.0–94.3)	93.9 (89.0–98.7)	91.0 (87.8–94.1)	0.332
Percentage PA	73.7 (65.0–82.5)	65.0 (41.6–88.4)	75.7 (66.2–85.2)	0.347	76.3 (66.3–86.2)	82.9 (59.2–90.8)	73.9 (62.0–85.9)	0.486
Percentage NA	92.0 (89.1–95.0)	92.0 (85.8–98.3)	92.0 (88.6–95.4)	0.947	95.0 (92.6–97.3)	96.3 (89.5–97.9)	94.5 (91.7–97.4)	0.605

DS 1–3 = negative; DS 4–5 = positive.
*P values of χ^2 test refer to comparison of baseline CT vs. baseline PET or PET/CT.
†Prevalence of positive scans was calculated as sum of number of scans in which both reviewers scored positively and half of scans with discrepancies divided by total number of scans.
Data are percentages, with 95%CIs in parentheses.

An example of a discrepancy in the interim assessment of a mesenteric bulky lesion is shown in Figure 1.

EoT-PET

Because in 10 EoT-PET scans one reviewer, and in 2 scans both, did not give a final conclusion, 457 scans were evaluable (Table 1). The median interval of EoT-PET scanning after the last chemotherapy cycle was 31 d (interquartile range, 22.5–48). The prevalence of positive EoT-PET scans was 17.5%. In 419 of 457 scans, the reviewers agreed on the final conclusion (negative or positive), yielding a percentage OA of 91.7% (95%CI, 89.0–94.3), a PA of 76.3% (95%CI, 66.3–86.2), and an NA of 95% (95%CI, 92.6–97.3).

Baseline ^{18}F-FDG PET or PET/CT was available in 75% (n = 333 integrated PET/CT, and n = 10 PET stand-alone with a separate CT scan), and diagnostic CT was available in the remaining cases (n = 114). Percentage OA, NA, and PA did not significantly differ between these groups (percentage OA, 91% and 93.3%, P = 0.332; NA, 94.5% and 96.3%, P = 0.605; PA, 73.9% and 82.9%; P = 0.486). Percentage OA for R2-CHOP14 compared with R-CHOP14 was 93.9 versus 89.4% (P = 0.082).

For ordinal DSs, the reviewers agreed in 220 of 457 cases (Supplemental Table 2), resulting in a percentage of exact agreement of 48.1% (95%CI, 43.4–52.8). Percentage agreement was 83.4% (95%CI, 79.8–86.9) when we allowed a 1-point difference—except for a discrepancy between scores 3 and 4—between the reviewers' scorings.

Supplemental Table 3 presents the percentage OA of the specific nodal and extranodal locations for dichotomized DS, related to the baseline prevalence. Gastrointestinal, skeletal, and mesenteric sites relatively showed the greatest number of discrepancies. Observer variation at EoT-PET in spleen and Waldeyer's ring was less than at I-PET.

Comparison I-PET and EoT-PET Interobserver Agreement

PA did not significantly differ between I-PET and EoT-PET assessments (P = 0.656), but percentage OA was lower for I-PET (87.7% vs. 91.7%, respectively, P = 0.049), and there was a trend toward lower NA (92.0% vs. 95.0%, respectively, P = 0.091).

DISCUSSION

Our study presents the interobserver agreement of DS results for I-PET and EoT-PET from a central review of a large multicenter randomized clinical trial in DLBCL. We found high percentages of OA (88%–92%) and NA (92%–95%) for both I-PET and EoT-PET using a DS of at least 4 for test positivity, at a lower (74%–76%) PA.

Most studies on interobserver agreement primarily report Cohen's κ and some present percentage of OA in addition. Cohen's κ is a relative measure, and the values are low in relatively low-prevalence situations (e.g., of residual lymphoma sites). Therefore, we report percentage OA, which is independent of differences in prevalence. In addition, we report specific agreement measures, which reflect the absolute probability that another reviewer gives the same conclusion as a colleague, specified for positive and negative test results (*18*). In other words: 74% PA implies that if one reviewer rates an I-PET scan as positive, the probability that another reviewer will provide the same result is 74%.

Similar studies presented different agreement measures (*14–16*). Itti et al. (*14*) (n = 114, 3 readers) reported pairwise Cohen's κ-values (0.53–0.80) for I-PET after 2 cycles of R-CHOP or R-ACVBP in a retrospective cohort but did not report specific agreement measures. From their presented data, we calculated OA of 77%–90% between observer pairs, subdivided into a NA of 81%–91% and PA of 72%–89%. Horning et al. (*16*) (n = 38 patients, 3 readers) reported a Fleiss' κ of 0.50 and percentage OA of 71% for I-PET after 3 cycles of R-CHOP, but specific agreement measures could not be extracted. Han et al. (*15*) reported κ-values of 0.41–0.52 and OAs of 82%–88% for I-PET (n = 55, after 3 cycles of R-CHOP) and EoT-PET (n = 57), respectively, as assessed by 2 readers. NA and PA as extracted from their presented data were 89% and 50% for I-PET and 92% and 59% for EoT-PET, respectively.

TABLE 2
Interobserver Agreement on Specific Nodal and Extranodal Localizations on I-PET

Location	Number baseline positive	Number of discrepancies on I-PET	Agreement on negativity (absolute)	Agreement on positivity (absolute)	Percentage OA	Related to baseline prevalence
Nodal						
Paraaortic*	414	17	899	14	98.2	4.1%
Cervical*†	302	8	915	6	99.1	2.6%
Iliac*	272	6	917	7	99.4	2.2%
Supraclavicular*	228	6	920	4	99.4	2.6%
Axillary*	225	9	920	1	99.0	4.0%
Mediastinal†	212	12	445	6	97.4	5.7%
Inguinal*	210	3	926	1	99.7	1.4%
Mesenteric	189	16	433	16	96.6	8.5%
Hilar*†	147	7	918	3	99.2	4.8%
Spleen†	115	11	442	6	97.6	9.6%
Other	105	7	457	1	98.5	6.7%
Waldeyer†	53	8	456	0	98.3	15.1%
Extranodal						
Other extranodal†	124	17	436	8	96.3	13.7%
Skeletal†	95	12	447	4	97.4	12.6%
Gastrointestinal†	61	12	441	7	97.4	16.7%
Lung†	55	3	455	6	99.4	5.5%
Liver	37	3	461	1	99.4	8.1%
Pleura	25	1	464	0	99.8	4.0%
Skin	11	0	465	0	100.0	0.0%
Central nervous system	0	0	465	0	100.0	0.0%

*Right and left are summed and presented together.
†Totals not 465 or 930, because of missing values or localization scored as unclear.
Percentage OA = (number of agreement on positivity + number of agreement on negativity)/(number of discrepancies + number of agreement on positivity + number of agreement on negativity) × 100%; related to baseline prevalence = (number of discrepancies/number baseline positive) × 100%.

Taken together, it appears that NA was generally above 80% in all studies (probably at least partly related to the high prevalence of negative scans). However, PA seemed to have a wider range between studies. In our study, we found a Cohen's κ-value of 0.65 and 0.71 for I-PET and EoT-PET, respectively.

Our data suggest that I-PET is more difficult to assess than EoT-PET. We found that the percentage of OA was lower for I-PET than for EoT-PET. The trend toward a lower NA for I-PET than for EoT-PET, could (in part) be caused by the higher number of negative scans at the end of treatment. In the study from Han et al., agreement measures also seemed generally higher for EoT-PET than for I-PET (*15*). Treatment-related inflammation shortly after chemotherapy might hamper the identification of lymphoma-related ^{18}F-FDG uptake.

In addition, we explored observer agreement as a function of disease location. Related to initially involved sites, we found the lowest percentages of OA for mesenteric, gastrointestinal, and skeletal sites in I-PET and EoT-PET. In these tissues, the local background of ^{18}F-FDG varies between and within patients over time; uptake due to intercurrent inflammation and, for example (healing), pathologic fractures needs to be accounted for, and this is not always covered by the Lugano criteria. In I-PET, discrepancies in spleen and Waldeyer's ring were more common. The short interval between the I-PET exams after the previous R-CHOP14 course (*20*) and the recent administration of granulocyte colony-stimulating factor (*21*) in our study could cause false-positive uptake in these organs. In an intra- and interobserver agreement study of baseline PET/CT from a mix of lymphoma subtypes, the lowest weighted κ-values were observed in hilar nodes, infraclavicular nodes, and bowel (*22*). However, these sites were also those sites that were least frequently involved with lymphoma in their cohort, thus κ-values could have been low because of the low prevalence of these sites. For some specific nodal and extranodal localizations, only a few positive cases were identified; therefore, we decided to report on numbers of discrepancies and percentage OA only.

Baseline PET/CT provides more accurate staging than CT only (*2,3*) and serves as a reference to quantify tumor response (SUV, metabolically active tumor volume). In our study, in which baseline PET/CT was not mandatory according to prevailing guidelines at the start of the study (*23,24*), we found that observer variability in treatment evaluation was independent of the baseline imaging modality.

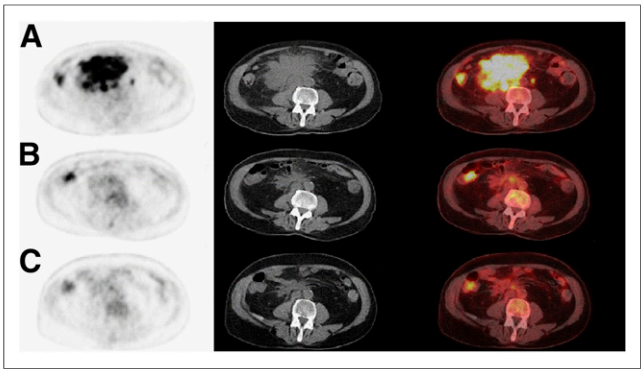

FIGURE 1. Example of discrepancy between reviewers' assessment of mesenteric lymph nodes on, from left to right, axial attenuation-corrected PET, low-dose CT, and fused PET/CT images. (A) Baseline ^{18}F-FDG PET/CT with mesenteric bulky mass. (B) I-PET/CT after 4 cycles of R-CHOP14. One reviewer scored scan negatively (DS 1) and the other reviewer scored DS 4 for residual uptake in mesenteric mass. (C) EoT-PET/CT after 6 cycles of R-CHOP14. Both reviewers scored scan negatively (DS 1 and DS 2, respectively).

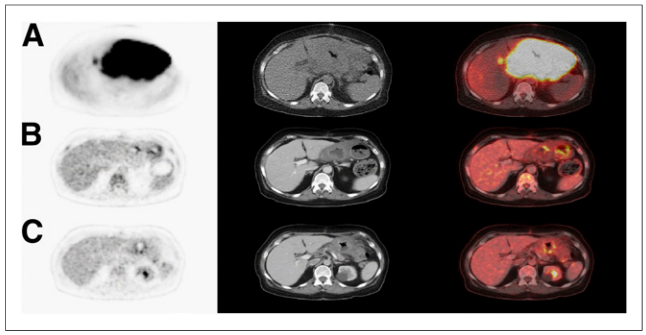

FIGURE 3. Example of discrepancy between reviewers' assessment of stomach on, from left to right, axial attenuation-corrected PET, low-dose CT, and fused PET/CT images. (A) Baseline ^{18}F-FDG PET/CT with clear localization of lymphoma in stomach. (B) I-PET/CT after 4 cycles of R-CHOP14. Reviewer 1 did not give final DS score and commented that stomach was "DS 4 but could be physiologic uptake." Reviewer 2 scored this scan negatively (DS 2). (C) EoT-PET/CT after 6 cycles of R-CHOP14. Reviewer 1 still commented on stomach but now scored negatively. Reviewer 2 again scored scan negatively (DS 2).

A strength of this study is that the I-PET and EoT-PET scans were assessed by 2 reviewers from a pool of 10, in contrast to previous studies with small, fixed numbers of reviewers. Scoring by a pool of reviewers represents the normal situation of ^{18}F-FDG PET/CT lymphoma response assessment in clinical practice. A limitation is that scans rated as unclear were excluded from our main analyses; in 13 I-PET and 10 EoT-PET scans one of the reviewers rated the final conclusion as unclear, specifying in the free-text section that they were not certain that the residual ^{18}F-FDG uptake was lymphoma-related. A similar conclusion was drawn by both observers in 2 EoT-PET scans. In analysis of a best-case (both reviewers agreed on negative or positive scores) and worst-case (discrepancy) scenario, we found that these results only slightly affected observer agreement: for I-PET and EoT-PET, percentages of OA were 88.1% and 91.9% in the best-case scenario, versus 85.4% and 89.7% in the worst-case scenario. In most of these unclear scans—13 of 13 I-PET scans and 6 of 12 EoT-PET scans—residual ^{18}F-FDG uptake in extranodal lymphoma sites caused the uncertainty, especially skeletal lesions in I-PET scans (perhaps because of enhanced bone marrow background uptake due to granulocyte colony-stimulating factor, or healing fractures or bone remodeling in previous lymphoma locations). Other reasons mentioned for unclear reads were missing baseline PET status, no contrast-enhanced CT scan available, inferior quality of a CT scan, a possible sarcoidlike response, or uncertainty about a nonresponding lesion while all other lesions responded (Figs. 2 and 3). Another limitation is the use of older-generation PET/CT systems (Supplemental Table 4), which could influence the generalizability of our results.

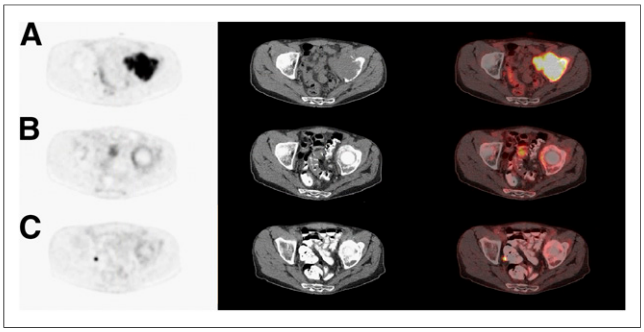

FIGURE 2. Example of discrepancy between reviewers' assessment of skeletal lesion on, from left to right, axial attenuation-corrected PET, low-dose CT, and fused PET/CT images. (A) Baseline ^{18}F-FDG PET/CT with skeletal lesion in left acetabulum. (B) I-PET/CT after 4 cycles of R-CHOP14 showing rim of uptake scored by one reviewer as DS 4 and by other reviewer as unclear. (C) EoT-PET/CT after 8 cycles of R-CHOP14 showing residual uptake scored by one reviewer as DS 4 and by other reviewer as unclear.

Our findings, especially the suboptimal PA, have implications for trial design and clinical practice. A PA of 74% at I-PET clearly emphasizes the need for central review procedures in clinical trials investigating intensified therapy in I-PET–positive patients. The 76% PA at EoT-PET reinforces the recommendation to discuss patients during multidisciplinary tumor board meetings, allowing for a second read of the test result, aiming for optimal patient management (e.g., confirmatory scan or biopsy).

Our data indicate that reviewers are especially uncertain in cases of extranodal lymphoma involvement, which is common in DLBCL patients, with baseline frequencies of up to 20% depending on the site (Table 2 and Supplemental Table 3). These results could be helpful in focusing the training of nuclear medicine physicians, such as by using the harmonization approach of Ceriani et al. (*25*).

During the last 10–15 y, ^{18}F-FDG PET/CT systems' quality continued to evolve, and guidelines therefore need to be updated on a regular basis (*26,27*).

CONCLUSION

Interobserver agreement among experienced nuclear medicine physicians using DS for I-PET and EoT-PET response assessment in DLBCL has high percentages of OA (88%–92%) and NA (92%–95%). The lower (74%–76%) PA suggests that the accuracy of EoT-PET/CT treatment evaluation in daily practice and I-PET–adapted trials may benefit from dual reads and central review, respectively.

DISCLOSURE

This work was financially supported by an Alpe d'HuZes/KWF fund, provided by the Dutch Cancer Society (VU 2012-5848). Pieternella Lugtenburg has received research funding from Roche, Servier, and Takeda and honoraria for advisory boards from Roche, Servier, Sandoz, BMS, Genmab, Takeda, and Celgene. No other potential conflict of interest relevant to this article was reported.

REFERENCES

1. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood.* 2016;127:2375–2390.
2. Barrington SF, Mikhaeel NG, Kostakoglu L, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the International Conference on Malignant Lymphomas Imaging Working Group. *J Clin Oncol.* 2014; 32:3048–3058.
3. Cheson BD, Fisher RI, Barrington SF, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. *J Clin Oncol.* 2014;32:3059–3068.
4. Weiler-Sagie M, Bushelev O, Epelbaum R, et al. ^{18}F-FDG avidity in lymphoma readdressed: a study of 766 patients. *J Nucl Med.* 2010;51:25–30.
5. 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.* 2018;390:2790–2802.
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;372:1598–1607.
7. 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;35:1786–1794.
8. Johnson P, Federico M, Kirkwood A, et al. Adapted treatment guided by interim PET-CT scan in advanced Hodgkin's lymphoma. *N Engl J Med.* 2016;374:2419–2429.
9. Terasawa T, Lau J, Bardet S, et al. Fluorine-18-fluorodeoxyglucose positron emission tomography for interim response assessment of advanced-stage Hodgkin's lymphoma and diffuse large B-cell lymphoma: a systematic review. *J Clin Oncol.* 2009;27:1906–1914.
10. Zijlstra JM, Burggraaff CN, Kersten MJ, Barrington SF; EHA Scientific Working Group on Lymphoma. FDG-PET as a biomarker for early response in diffuse large B-cell lymphoma as well as in Hodgkin lymphoma? Ready for implementation in clinical practice? *Haematologica* [editorial]. 2016;101:1279–1283.
11. Hertzberg M, Gandhi MK, Trotman J, et al.; Australasian Leukaemia Lymphoma Group (ALLG). Early treatment intensification with R-ICE and ^{90}Y-ibritumomab tiuxetan (Zevalin)-BEAM stem cell transplantation in patients with high-risk diffuse large B-cell lymphoma patients and positive interim PET after 4 cycles of R-CHOP-14. *Haematologica.* 2017;102:356–363.
12. Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. *Br J Radiol.* 1997;70:1085–1098.
13. Hasenclever D, Kurch L, Mauz-Körholz C, et al. qPET: a quantitative extension of the Deauville scale to assess response in interim FDG-PET scans in lymphoma. *Eur J Nucl Med Mol Imaging.* 2014;41:1301–1308.
14. Itti E, Meignan M, Berriolo-Riedinger A, et al. An international confirmatory study of the prognostic value of early PET/CT in diffuse large B-cell lymphoma: comparison between Deauville criteria and ΔSUVmax. *Eur J Nucl Med Mol Imaging.* 2013;40:1312–1320.
15. Han EJ, O JH, Yoon H, et al. FDG PET/CT response in diffuse large B-cell lymphoma: reader variability and association with clinical outcome. *Medicine (Baltimore).* 2016;95:e4983.
16. Horning SJ, Juweid ME, Schöder H, et al. Interim positron emission tomography scans in diffuse large B-cell lymphoma: an independent expert nuclear medicine evaluation of the Eastern Cooperative Oncology Group E3404 study. *Blood.* 2010;115:775–777.
17. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol.* 2011;64:96–106.
18. de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's κ. *BMJ.* 2013;346:f2125.
19. Lugtenburg PJ, de Nully Brown P, van der Holt B, et al. Randomized phase III study on the effect of early intensification of rituximab in combination with 2-weekly CHOP chemotherapy followed by rituximab or no maintenance in patients with diffuse large B-cell lymphoma: results from a HOVON-Nordic Lymphoma Group study [abstract]. *J Clin Oncol.* 2016;34(suppl):7504.
20. Hüttmann A, Müller S, Jöckel KH, Dührsen U. Pitfalls of interim positron emission tomography scanning in diffuse large B-cell lymphoma. *J Clin Oncol.* 2010;28:e488–e489.
21. Jacene HA, Ishimori T, Engles JM, Leboulleux S, Stearns V, Wahl RL. Effects of pegfilgrastim on normal biodistribution of ^{18}F-FDG: preclinical and clinical studies. *J Nucl Med.* 2006;47:950–956.
22. Hofman MS, Smeeton NC, Rankin SC, Nunan T, O'Doherty MJ. Observer variation in interpreting ^{18}F-FDG PET/CT findings for lymphoma staging. *J Nucl Med.* 2009;50:1594–1597.
23. Cheson BD, Pfistner B, Juweid ME, et al.; International Harmonization Project on Lymphoma. Revised response criteria for malignant lymphoma. *J Clin Oncol.* 2007;25:579–586.
24. Juweid ME, Stroobants S, Hoekstra OS, et al.; Imaging Subcommittee of International Harmonization Project in Lymphoma. Use of positron emission tomography for response assessment of lymphoma: consensus of the Imaging Subcommittee of International Harmonization Project in Lymphoma. *J Clin Oncol.* 2007;25:571–578.
25. Ceriani L, Barrington S, Biggi A, et al. Training improves the interobserver agreement of the expert positron emission tomography review panel in primary mediastinal B-cell lymphoma: interim analysis in the ongoing International Extranodal Lymphoma Study Group-37 study. *Hematol Oncol.* 2017;35:548–553.
26. Boellaard R, Oyen WJ, Hoekstra CJ, et al. The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *Eur J Nucl Med Mol Imaging.* 2008;35:2320–2333.
27. Boellaard R, Delgado-Bolton R, Oyen WJ, et al.; European Association of Nuclear Medicine (EANM). FDG PET/CT: EANM procedure guidelines for tumour imaging—version 2.0. *Eur J Nucl Med Mol Imaging.* 2015;42:328–354.