
International Validation Study for Interim PET in ABVD-Treated, Advanced-Stage Hodgkin Lymphoma: Interpretation Criteria and Concordance Rate Among Reviewers

Alberto Biggi¹, Andrea Gallamini², Stephane Chauvie³, Martin Hutchings⁴, Lale Kostakoglu⁵, Michele Gregianin⁶, Michel Meignan⁷, Bogdan Malkowski^{8,9}, Michael S. Hofman¹⁰, and Sally F. Barrington¹¹

¹Nuclear Medicine Department, Azienda Ospedaliera S. Croce e Carle, Cuneo, Italy; ²Hematology Department, Azienda Ospedaliera S. Croce e Carle, Cuneo, Italy; ³Medical Physics Department, Azienda Ospedaliera S. Croce e Carle, Cuneo, Italy; ⁴Departments of Hematology and Oncology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark; ⁵Department of Radiology, Division of Nuclear Medicine, Mount Sinai Medical Center, New York, New York; ⁶Radiotherapy and Nuclear Medicine Unit, Istituto Oncologico Veneto IOV, IRCCS, Padova, Italy; ⁷Department of Nuclear Medicine, Centre Universitaire Hospitalier Henri Mondor, Creteil, Paris, France; ⁸Department of Nuclear Medicine, Oncology Centre, Bydgoszcz, Poland; ⁹Department of PET and Molecular Imaging, Collegium Medicum Bydgoszcz, University of N. Copernicus, Torun, Poland; ¹⁰Centre for Cancer Imaging, Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, Australia; and ¹¹Division of Imaging, PET Centre, Guy's and St. Thomas' Hospital and King's College, London, United Kingdom

At present, there are no standard criteria that have been validated for interim PET reporting in lymphoma. In 2009, an international workshop attended by hematologists and nuclear medicine experts in Deauville, France, proposed to develop simple and reproducible rules for interim PET reporting in lymphoma. Accordingly, an international validation study was undertaken with the primary aim of validating the prognostic role of interim PET using the Deauville 5-point score to evaluate images and with the secondary aim of measuring concordance rates among reviewers using the same 5-point score. This paper focuses on the criteria for interpretation of interim PET and on concordance rates. **Methods:** A cohort of advanced-stage Hodgkin lymphoma patients treated with doxorubicin, bleomycin, vinblastine, and dacarbazine (ABVD) were enrolled retrospectively from centers worldwide. Baseline and interim scans were reviewed by an international panel of 6 nuclear medicine experts using the 5-point score. **Results:** Complete scan datasets of acceptable diagnostic quality were available for 260 of 440 (59%) enrolled patients. Independent agreement among reviewers was reached on 252 of 260 patients (97%), for whom at least 4 reviewers agreed the findings were negative (score of 1–3) or positive (score of 4–5). After discussion, consensus was reached in all cases. There were 45 of 260 patients (17%) with positive interim PET findings and 215 of 260 patients (83%) with negative interim PET findings. Thirty-three interim PET-positive scans were true-positive, and 12 were false-positive. Two hundred three interim PET-negative scans were true-negative, and 12 were false-negative. Sensitivity, specificity,

and accuracy were 0.73, 0.94, and 0.91, respectively. Negative predictive value and positive predictive value were 0.94 and 0.73, respectively. The 3-y failure-free survival was 83%, 28%, and 95% for the entire population and for interim PET-positive and -negative patients, respectively ($P < 0.0001$). The agreement between pairs of reviewers was good or very good, ranging from 0.69 to 0.84 as measured with the Cohen kappa. Overall agreement was good at 0.76 as measured with the Krippendorff α . **Conclusion:** The 5-point score proposed at Deauville for reviewing interim PET scans in advanced Hodgkin lymphoma is accurate and reproducible enough to be accepted as a standard reporting criterion in clinical practice and for clinical trials.

Key Words: interim PET; Hodgkin lymphoma; interpretation criteria; concordance rate; clinical trial

J Nucl Med 2013; 54:683–690
DOI: 10.2967/jnumed.112.110890

Preliminary reports have shown that ¹⁸F-FDG PET performed early during doxorubicin, bleomycin, vinblastine, and dacarbazine (ABVD) treatment of patients with Hodgkin lymphoma predicts the treatment outcome (1,2). Moreover, interim PET is a more effective predictor of treatment response than well-established clinical prognostic factors such as the International Prognostic Score (3). Further reports have confirmed these findings, with an overall sensitivity and specificity for interim PET in predicting treatment outcome ranging between 43% and 100% and between 67% and 100%, respectively (4). One of the most relevant factors affecting the variation in sensitivity and specificity observed was the heterogeneity of interim PET interpretation,

Received Jul. 7, 2012; revision accepted Nov. 6, 2012.
For correspondence or reprints contact: Alberto Biggi, Nuclear Medicine Department, Azienda Ospedaliera S. Croce e Carle, Via M. Coppino 26, 12100, Cuneo, Italy.
E-mail: biggi.a@ospedale.cuneo.it
Published online Mar. 20, 2013.
COPYRIGHT © 2013 by the Society of Nuclear Medicine and Molecular Imaging, Inc.

as no standard criteria were established for interim PET reporting.

Thus, in April 2009 an international workshop attended by hematologists and nuclear medicine experts was held in Deauville, France, with the aim of developing simple and reproducible rules for interim reporting in lymphoma. A consensus among experts was reached on the appropriateness of a qualitative determination of residual ^{18}F -FDG uptake by visual assessment, the so-called Deauville 5-point scale (5-PS). The 5-PS compares residual uptake, if present, in sites of initial disease with mediastinal blood-pool structures and the liver (5). The experts proposed that the criteria should be validated in a large cohort of patients affected by diffuse large B-cell lymphoma and Hodgkin lymphoma. In October 2009, the International Validation Study in Hodgkin Lymphoma was launched. A homogeneous cohort of advanced-stage, ABVD-treated Hodgkin lymphoma patients was retrospectively enrolled. Scans were reviewed by an international panel of nuclear medicine experts. Patients had to be staged at baseline and after 2 ABVD courses with a PET/CT scan, without any treatment change based on interim PET. Reviewers reported the scans according to the 5-PS.

This paper focuses mainly on the interpretation criteria for interim PET to formulate a clear set of instructions for PET/CT reporters and to measure concordance rates among reviewers using the 5-PS with the defined instructions. The prognostic role of interim PET in the International Validation Study in Hodgkin Lymphoma is the subject of a separate paper.

MATERIALS AND METHODS

Patient Data Retrieval

Four hundred forty consecutive patients from 17 clinical centers in Australia, Denmark, France, Israel, Italy, Poland, the United Kingdom, and the United States, in whom Hodgkin lymphoma was diagnosed between January 2002 and December 2009, were considered eligible and retrospectively enrolled in the study.

The inclusion criteria were as follows: advanced-stage (stages IIB–IVB) Hodgkin lymphoma or stage IIA Hodgkin lymphoma with adverse prognostic factors, treatment with 4–6 cycles of ABVD with or without involved-field radiotherapy or consolidation radiotherapy, PET/CT staging at baseline and after 2 ABVD courses, no treatment change based on interim PET results, a minimum follow-up of 1 y after treatment completion, and informed written consent. Patients treated with intensified chemotherapy for progressive or resistant lymphoma during ABVD chemotherapy were eligible only if the treatment change was decided on the basis of clinical or radiologic evidence of disease progression. Patients were excluded from the study if they had not been examined with PET/CT (e.g., C-PET [BC Technical, Inc.] or multiring PET without CT), had been scanned on different PET/CT cameras at baseline and after 2 ABVD courses, had fasting glucose values greater than 200 mg/dL at the time of ^{18}F -FDG administration, did not have a full DICOM dataset for PET and CT images, or had images of nondiagnostic quality.

The centers were asked for various clinical data at diagnosis and for data relating to treatment outcome. All the data requested, including disease status at latest follow-up, cause of death, and duration of last follow-up, were available for every patient.

The study was approved by the Ethical Committee of the coordinating center in Cuneo and conducted according to the Helsinki declaration. Patient written informed consent was obtained for PET scanning and for use of anonymized data and images for teaching and research purposes. Specific informed written consent to be included in this study was not required as all data were anonymized from participating academic centers. Data collection conformed to specific institutional and national requirements.

PET/CT Scanning

All patients underwent PET/CT scans before chemotherapy (baseline PET) and after 2 cycles of ABVD (interim PET). Both scans were obtained according to the usual scanning protocol of the individual PET center.

PET/CT Acquisition

The administered ^{18}F -FDG activity was 362 ± 88 MBq (mean \pm SD; range, 51–694 MBq) for baseline PET and 355 ± 82 MBq (range, 48–699 MBq) for interim PET. The interval between ^{18}F -FDG injection and PET acquisition (uptake time) was 85 ± 43 min for baseline PET and 79 ± 24 min for interim PET. The uptake time was uniformly distributed between 55 and 100 min, with only 30% of the patients having images acquired in the 60 ± 10 min range, as is regarded to be standard in oncologic imaging (6). Interim PET scans were obtained 12.3 ± 4.9 d (range, 7–22 d) after administration of day 15 chemotherapy during the second ABVD cycle. Images were attenuation-corrected using iterative reconstruction, with SUV normalized according to body weight and administered activity.

PET/CT Data Retrieval and Review Scheme

After anonymization, baseline and interim scans were uploaded from the participating PET centers to a dedicated Web site called WIDEN, which is a Web-based tool for imaging exchange (7). Once received in the Core Lab for the study in Cuneo, all scans were checked for image quality. The field of view had to encompass an area from the base of the skull to below the pelvis and include the femoral heads. PET/CT scans were then transferred to a dedicated workstation and distributed via a central server hosted by Keosys to reviewers. All scans were viewed remotely using the same software (Positoscope; Keosys). Reviewers were masked to patient history and clinical data and were asked to report the scans independently.

Criteria of Interpretation for Interim PET/CT

Interim PET scans were compared with baseline PET and analyzed using 5-PS, where a score of 1 indicated no residual uptake above the background level, 2 indicated residual uptake less than or equal to the mediastinum, 3 indicated residual uptake greater than the mediastinum but not greater than the liver, 4 indicated residual uptake moderately increased compared with the liver, and 5 indicated residual uptake markedly increased compared with the liver or new sites of disease. These criteria were used for grading nodal and extranodal disease, with scores of 1–3 regarded as negative and scores of 4 and 5 regarded as positive. Information on how the scan had been reported by the local center during the course of the patients' treatment was also obtained from participating centers.

A more detailed set of instructions was drawn up to deal with potential confounding variables such as the interpretation of marrow uptake, which required further clarity according to the experience of reviewers who had used the 5-PS previously in the

clinic or in trials. The panel of reviewers agreed to the following instructions before starting the review process:

Nodal and extranodal focal ^{18}F -FDG uptake in interim PET represents residual lymphoma if the intensity is greater than uptake in normal liver (score of 4 or 5) at sites involved on baseline PET.

A new lesion (not present on baseline PET) in a patient who is responding to treatment at other sites is unlikely to be lymphoma. The scan should be scored accordingly as 1, 2, or 3 depending on the residual level of uptake, if any, in initial disease sites.

A new lesion or lesions (not present on baseline PET) in a patient with residual lymphoma is likely to represent a new site of lymphoma. The scan should be scored as 5 (progressive disease) unless there is a clear alternative explanation, such as increased focal uptake in the lungs and CT correlative changes suggestive of infection.

Diffusely increased uptake in the bone marrow—even if more intense than the liver—is usually due to marrow stimulation after chemotherapy, especially if growth factors have been used. Such uptake should not be misinterpreted as marrow involvement even if focal uptake was present in marrow at baseline PET.

Diffusely increased uptake in the spleen—even if more intense than the liver—in association with diffuse bone marrow uptake is usually due to chemotherapy effects. Such uptake should not be misinterpreted as splenic involvement even if focal uptake was present in the spleen at baseline PET.

A focal reduction of uptake in sites of marrow involvement at baseline PET occurs because of marrow ablation with successful treatment. Focal increased uptake may occur at sites where there was no disease on baseline PET because of chemotherapy stimulation. The patterns of uptake at baseline PET and interim PET may therefore mirror each other, with sites of initial disease becoming cold and sites of normal marrow becoming hot on the treatment scan. Focal uptake in the marrow with this pattern should not be misinterpreted as disease.

Symmetric tonsillar uptake (on baseline and interim PET) is most likely to represent nonspecific inflammatory uptake in Hodgkin lymphoma and should not be misinterpreted as lymphoma. Asymmetric uptake on interim PET should be regarded as disease only in the presence of clear evidence of tonsillar involvement at baseline.

Agreement Among Reviewers

The review panel comprised 6 nuclear medicine experts from 5 countries. The panel agreed that the final interim PET score for each patient, positive (score of 4 or 5) or negative (score of 1, 2, or 3), was assigned by the majority view, that is, agreement between at least 4 reviewers.

Statistical Analysis

The concordance between pairs of reviewers with respect to binary reporting (positive vs. negative) was measured with the Cohen κ (8) for the 15 combinations of the 6 reviewers. κ -values between 0.21 and 0.40, 0.41 and 0.60, 0.61 and 0.80, and 0.81 and 1.0 indicate fair, moderate, good, and very good agreement, respectively (9). The overall concordance between reviewers with respect to binary reporting (positive vs. negative) was measured using the Krippendorff α -coefficient (10). Survival curves were measured using the Kaplan-Meier method and evaluated using Log-rank regression (11).

RESULTS

PET/CT Data Retrieval

Baseline and interim PET scans were available for 335 of 440 patients with complete clinical datasets. Image retrieval

was not possible in 105 of these because scans could not be retrieved from files archived locally or were not accessible in DICOM format. Images were uploaded from participating PET centers to the Core Lab via WIDEN with a median upload time of 2 min and 3 s for paired PET/CT scans (baseline PET and interim PET). Seventy-five patients had to be excluded because of incomplete image data ($n = 61$), poor-quality scans ($n = 6$), or miscellaneous reasons ($n = 8$), leaving 260 patients (59%) of the 440 initially enrolled available for the review process.

PET/CT Reporting

At the end of the review process, independent agreement was reached on 252 of 260 patients (97%); more than 4 reviewers agreed that the scan result was positive or negative. All 6 reviewers agreed in 212 of 260 cases (82%), at least 5 reviewers agreed in 240 cases (92%), and at least 4 reviewers in 252 cases (97%). There were 42 of 252 patients (17%) with a positive interim PET result and 210 of 252 with a negative interim PET result (83%). A consensus session was held in London to discuss the 8 true-discordant cases—those for which the opinion of the reviewers was equally split as to whether the scan was positive or negative (Table 1). Reviewers requested clinical information on 1 patient because of the presence of a suspected unilateral parotid adenoma. Consensus was reached on all patients with a majority agreement, resulting finally in 45 of 260 patients (17%) with a positive interim PET result and 215 of 260 patients (83%) with a negative interim PET result. Of the 45 patients with a positive interim PET result, 33 were true-positive and 12 were false-positive; of the 215 patients with a negative interim PET result, 203 were true-negative and 12 were false-negative. Sensitivity, specificity, and accuracy were 0.73, 0.94, and 0.91, respectively. The positive predictive value and the negative predictive value were 0.73 and 0.94, respectively. After a median follow-up of 37 mo, the 3-y failure-free survival (FFS) for the entire patient population and for the interim PET-positive and -negative patients was 83%, 28%, and 95%, respectively ($P < 0.0001$) (Fig. 1). The 3-y FFS for PET-positive and PET-negative patients according to local interpretation was 54% and 94%, respectively (Fig. 1).

Agreement Among Reviewers

The Cohen κ for agreement between pairs of reviewers ranged from 0.69 to 0.84 (“good and very good”). Overall agreement among reviewers measured with the Krippendorff α was 0.76 (“excellent”) (Table 2). Twelve patients had false-positive interim PET results: the score was 4 in 8 patients and 5 in 4 patients. All patients were alive after a median follow-up of 51 mo. Residual uptake that turned out to be a false-positive result was located in the mediastinum in 6 patients; in the cervical region in 2 patients; and in the axilla, at the lung hilum, in lung parenchyma, and in bone in 1 patient each. In 4 patients the residual uptake was at a site of initial bulk disease, and in 7 patients more than one false-positive site of residual uptake was identified.

TABLE 1
Initial and Final Review Results of Patients Discussed at Joint Session in London

Patient no.	Initial review (positive/negative)	Final review (positive/negative)	Reason for disagreement	Consensus after discussion
025	3/3	0/6	Focal left parotid uptake (adenoma or residual disease?)	Left parotid adenoma (confirmed by fine-needle aspiration)
168	3/3	0/6	Sternum (healing process or residual disease?)	Healing of pathologic fracture
215	3/3	0/6	Liver or gut uptake?	"Liver" uptake due to misregistered physiologic gut uptake
229	3/3	6/0	Left cervical node (positive or negative?)	¹⁸ F-FDG uptake in residual node higher than liver despite very low liver uptake (maximum standardized uptake value, 1.7)
231	3/3	6/0	Subcarinal node or cardiac uptake?	Subcarinal node
242	3/3	6/0	Left humerus (diseased or not?)	Baseline and interim PET scans performed with different arm positions (measuring distance of lesion from humeral head, reviewers concurred there was true focal uptake in bone marrow present on baseline and on interim scans higher than uptake in normal liver)
247	3/3	2/4	Sternum (healing process or residual disease?)	Healing of pathologic fracture
293	3/3	0/6	Cervical node or brown fat?	Brown fat

Figures 2, 3, and 4 depict representative false-positive and discordant cases. Twelve patients had false-negative interim PET scans; the score was 3 in 7 patients, 2 in 1 patient, and 1 in 4 patients. The negative predictive value was 87% for scans scored as 3 and 97% for scans scored 1 or 2.

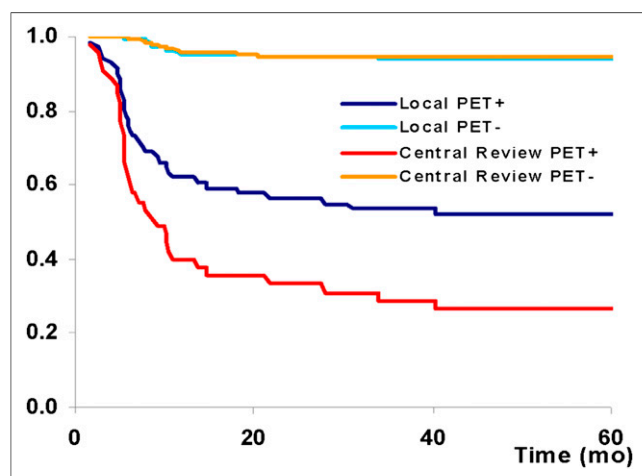


FIGURE 1. Three-year FFS of interim PET-positive and interim PET-negative patients according to review panel using 5-PS and according to local review.

DISCUSSION

The use of interim PET in patients with lymphoma has emerged as a powerful prognostic tool, particularly in Hodgkin lymphoma, when compared with well-established clinical parameters such as the International Prognostic Score. A relatively wide range of sensitivity and specificity has been reported in the literature for interim PET, possibly related to the use of different interpretation criteria used by various groups (4).

The criteria suggested by the International Harmonization Project for interpretation of PET (12) were originally proposed for end-of-treatment assessment and are affected by a high percentage of false-positive results when applied to interim PET interpretation (13). The high false-positive rate likely depends on the relatively low threshold used to define a positive scan. The reference background in the International Harmonization Project criteria is mediastinal blood-pool activity for a residual mass with a diameter of at least 2 cm and local background activity for a residual mass with a diameter of less than 2 cm. Inflammation induced by treatment may result in higher uptake on interim scans than on end-of-treatment scans, and therefore it is assumed that patients may have a degree of residual uptake higher than that in mediastinal blood-pool structures or local background on interim scans and still achieve a complete

TABLE 2
Agreement Between Pairs of Reviewers with Respect to Negative vs. Positive PET Scans Using Cohen κ

	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
Reviewer 1	1	0.73	0.77	0.77	0.75	0.73
Reviewer 2	0.73	1	0.71	0.75	0.69	0.70
Reviewer 3	0.77	0.71	1	0.83	0.77	0.77
Reviewer 4	0.77	0.75	0.83	1	0.83	0.84
Reviewer 5	0.75	0.69	0.77	0.83	1	0.78
Reviewer 6	0.73	0.70	0.77	0.84	0.78	1

metabolic response (5). On the basis of this assumption and to increase the positive predictive value of interim PET in a clinical trial where patients are escalated from ABVD to a regimen of bleomycin, etoposide, doxorubicin, cyclophosphamide, vincristine, procarbazine, and prednisone (Response-Adapted Therapy in Hodgkin Lymphoma [RATHL]; www.cancer.gov/clinicaltrials/ct2/show/NCT00678327), criteria were developed for interim PET reporting, previously referred to as the London criteria (14). The criteria consisted of a 5-point scale, with the intensity of uptake at sites involved on a baseline scan scored by comparison with the uptake in normal mediastinum and liver. The use of a graded visual response assessment reflects the fact that ^{18}F -FDG uptake is a continuum, with the likelihood of malignancy increasing as the level of ^{18}F -FDG uptake increases, rather than a black-or-white phenomenon indicating the presence or absence of malignancy (15). For the purposes of RATHL, interim scans on which any site involved at diagnosis shows residual uptake higher than the liver are regarded as positive scans. Nonetheless, using a graded assessment, the threshold chosen to define a positive or negative scan can be adapted to fit the clinical or research context; a lower threshold such as the mediastinum might be preferred in a situation in which treatment is deescalated. A high level of agreement was reported for 4 European centers using the London criteria in a population of 50 patients with stages II–V Hodgkin lymphoma (15).

During the First International Workshop on Interim PET in Lymphoma, which took place in Deauville, France, it was proposed that the London criteria be adopted and their use validated in interim PET interpretation of Hodgkin lymphoma and diffuse large B-cell lymphoma (5).

The International Validation Study in Hodgkin Lymphoma reported here is the first, to our knowledge, to investigate predefined criteria for interim PET reporting in a homogeneous population of Hodgkin lymphoma patients treated with ABVD from clinical institutions worldwide in which interim PET was not used to change treatment. Images were acquired in actual clinical environments using local protocols for PET/CT. The use of a Web-based tool for image exchange and the central Core Lab enabled images to be collated from centers using different imaging platforms and software programs. The panel of reviewers analyzed images using identical software (16) and reported results according to the 5-PS using a predefined set of instructions to limit the variability of interpretation.

The International Validation Study in Hodgkin Lymphoma confirmed that interim PET identifies a cohort of patients with negative scan results who have significantly better FFS than patients with positive results. Using the 5-PS, patients with advanced-stage Hodgkin lymphoma treated with 2 ABVD cycles had a 3-y FFS of 95% (scores of 1–3, compared with a 3-y FFS of 28% (scores of 4 or 5). The FFS for patients with negative results was similar for patients using 5-PS criteria and local interpretation, but the central review process with 5-PS criteria was better at discriminating patients with poor outcomes than was local review without predefined reporting criteria. FFS in patients with PET-positive scans was 28% according to the review panel and 54% according to the local review. These data confirm that criteria for interpretation of interim PET can significantly affect clinical results. The fact that NPV was better for scans scored 1 or 2 than those scored 3 likely reflects the uncertainty as to whether uptake in the gray area between mediastinal and liver uptake represents inflammation or low-volume disease. This supports the view that a graded visual assessment is meaningful and that outcomes in patients with different levels of ^{18}F -FDG uptake should ideally be measured in clinical trials.

The agreement between reviewers was good or very good, similar to that previously reported in 50 test cases using the 5-PS before its use in the RATHL trial (15). The percentage of discordant cases was lower, at 3% vs. 12%, in the smaller study. Perhaps the use of a clear set of operating instructions to clarify issues such as the interpretation of marrow and splenic uptake to assist in interpretation may have been beneficial. The discordant cases were typical of cases that are challenging in daily practice, as previously reported. There was difficulty distinguishing the healing process from residual disease in pathologic fractures, separating physiologic from pathologic uptake with prominent brown fat uptake, separating misregistered physiologic uptake in the gut from liver uptake, interpreting uptake when the arms were positioned differently on 2 scans, and differentiating a parotid adenoma from lymphoma, and some sites of disease were overlooked.

This was a retrospective study with a wide variation in ^{18}F -FDG uptake times. There was less variation in the timing of scans in relation to chemotherapy, with centers attempting to scan as late after chemotherapy as possible, suggesting that late scanning is generally accepted to be

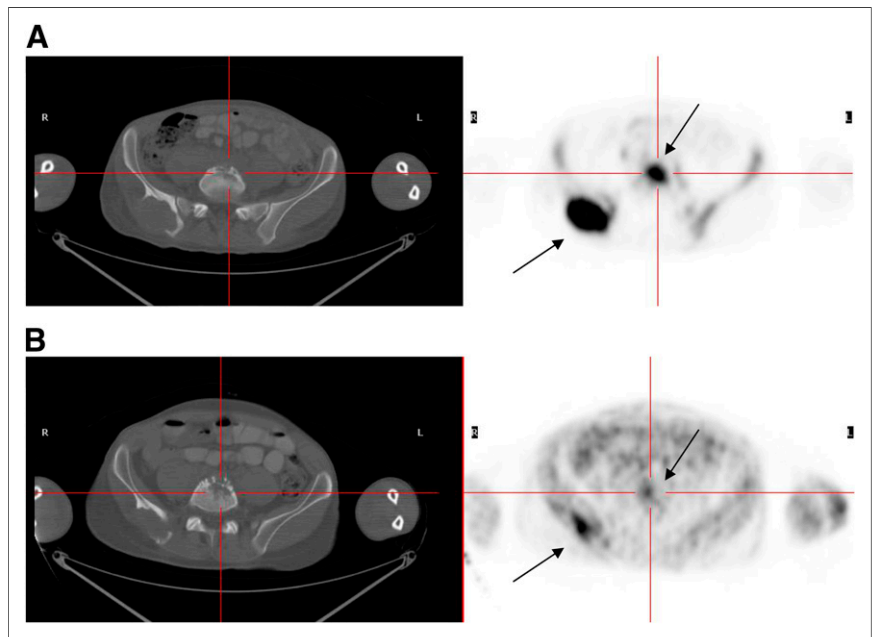


FIGURE 2. Patient 199, false-positive case: baseline PET/CT (A); interim PET/CT (B). Score 5 residual lesion is seen in right posterior ileum, and score 4 residual lesion in body of S1. Patient was alive in complete remission after 72 mo.

important for reliable interpretation. Despite differences in scan acquisition, the high concordance between reviewers suggests that the Deauville criteria are sufficiently robust to use in clinical practice and clinical trials. Standardization of PET/CT acquisition methods and quality control is gaining widespread acceptance and should further improve the reliability of PET/CT and allow comparable data to be collected for the purposes of multicenter trials (6).

Most response-adapted clinical trials using PET in advanced Hodgkin lymphoma are designed to intensify treatment in poor responders to improve disease control and to either leave treatment unchanged or reduce it in good responders. Thus, a high PPV associated with a highest NPV

is desirable. Our study suggests that, at least for patients treated with ABVD, a sensible choice would be to define a positive scan by setting an empiric threshold for the amount of residual activity higher than the liver.

The 3-y FFS of interim PET-positive patients in our study population is in the previously reported range of 13%–53% (3,17,18). There were 12 false-positive results (27% of interim PET-positive patients).

False-positive interim results could be related to any of several potential causes. The first is the disease itself, with treatment-related inflammation, delayed treatment response, and successful salvage accounting for good patient outcomes despite treatment failure. The second is a lack of

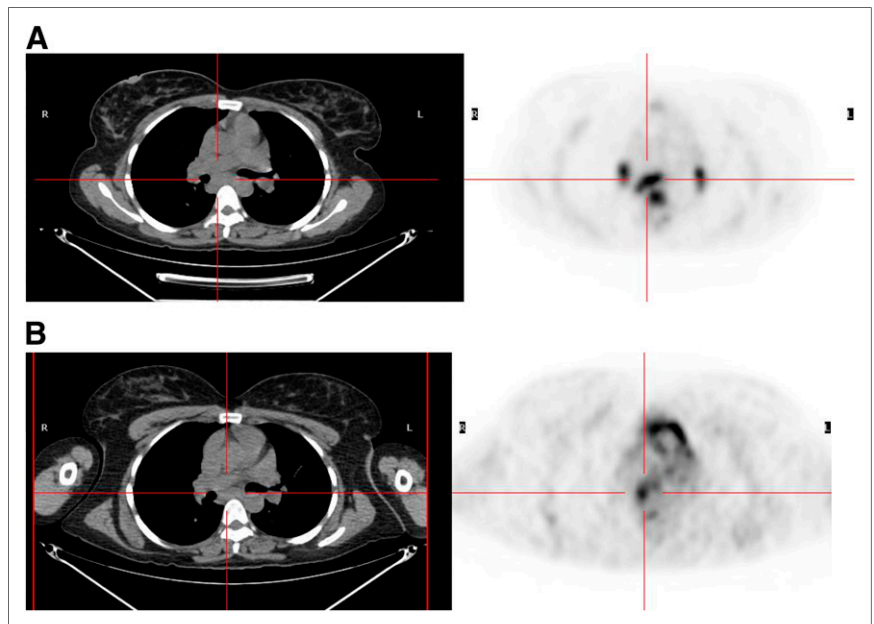


FIGURE 3. Patient 231, discordant case: baseline PET/CT (A); interim PET/CT (B). Subcarinal node shows abnormal uptake, which was either missed or thought to be physiologic heart uptake by some reviewers. Final consensus was score 4 residual subcarinal node. Patient was alive in complete remission after 42 mo.

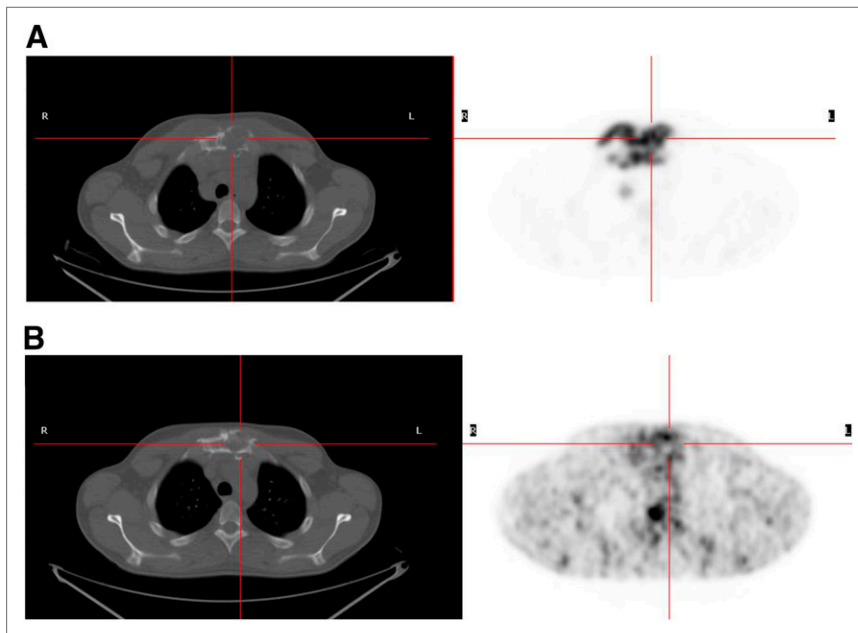


FIGURE 4. Patient 168, discordant case: baseline PET/CT (A); interim PET/CT (B). Increased uptake seen in sternum at site of pathologic fracture could be due to healing of fracture. Final consensus was score 3 residual lesion. Patient was alive in complete remission after 35 mo.

clinical information that could have assisted in scan interpretation, as would be the case in real life. In one patient (patient 25), clinical information requested by the reviewers modified the final interpretation with identification of a parotid adenoma; in another (patient 213), the scan was called positive because of an obvious lesion at the right lung hilum previously involved by disease, yet the patient had an intercurrent respiratory infection that might have influenced the false-positive interpretation. A third potential cause of false-positive interim results is difficulties in interpretation of uptake within pathologic fractures. Three patients were initially scored by some reviewers as having residual disease at fracture sites (patients 168, 199, and 247). One of these was scored as positive by most reviewers, yet all 3 patients turned out to be true-negative. This is an important learning point. Finally, a fourth cause is the variation in scan preparations and protocols between centers, as assessment relied on comparing residual activity with liver uptake, which may not be stable over time.

In general, a review process undertaken without access to clinical information will affect the false-positive rather than the false-negative rate because of the nonspecific behavior of ^{18}F -FDG. This might also explain the higher number of false-positive results and higher FFS in this study than the study previously reported by Gallamini et al., in which reviewers had access to relevant clinical information that might account for false-positive findings (3). Quite recently, preliminary reports have suggested that the specificity of interim PET could be improved by scanning at 2 time points and calculating a retention index to discriminate inflammatory from residual tumor uptake (19). Nonetheless, some patients will inevitably be overtreated in trials in which treatment is escalated, and this possibility should be considered in the trial design. It is reassuring to note

that in the present study, if treatment of PET-positive patients had been escalated on the basis of the PET interpretation, the result would have been overtreatment of less than 5% of the study population.

The 3-y FFS of PET-negative patients in our study population was 95% and is consistently high in studies reported previously after 2 cycles (1,17,18) and even after 1 cycle of treatment (20). There were only 12 false-negative results (6% of interim PET-negative patients); 7 of these were scored as 3, which may reflect the problems with differentiating low-volume disease from inflammation or the problems with using the liver as a reference organ when activity may vary from patient to patient, especially if the patients did not have an identical scan preparation as was the case in this study. In other cases, the residual disease might be too minor to identify using current PET technology, with its limited spatial resolution.

CONCLUSION

Our study confirmed that interim PET performed after 2 cycles of ABVD in advanced-stage Hodgkin lymphoma identifies patients with a significantly worse FFS when scans are positive than when scans are negative. The 5-PS combined with a detailed set of instructions is sufficiently robust to be accepted as a standard reporting tool for interpretation of interim PET scans; moreover, use of the liver to define a positive scan is the most accurate threshold to maintain a high negative predictive value while optimizing the positive predictive value.

DISCLOSURE

The costs of publication of this article were defrayed in part by the payment of page charges. Therefore, and solely to indicate this fact, this article is hereby marked

“advertisement” in accordance with 18 USC section 1734. No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We acknowledge the kind assistance of Anna Cavallo, from the Hematology Department in Cuneo, for data editing and text formatting; the technical support of Cerello Piergiorgio of the National Institute of Nuclear Physics (INFN) in Turin, Italy, for providing the imaging exchange tool WIDEN; the technical assistance of Keosys Company for providing the Positroscope network to distribute images to reviewers; and the assistance of Emanuele Roberto with statistical analysis.

We thank the following contributors who enrolled patients, obtained and contributed the scans, or sent the PET scan images for central review: Seymour John and Rod Hicks, Department of Hematology and Centre for Cancer Imaging, Peter MacCallum Cancer Centre, Melbourne, Australia; Lena Specht, Department of Radiotherapy, Rigshospitalet, Copenhagen University Hospital, Denmark; Alina Berriolo, Department of Nuclear Medicine, Centre Hospitalier de Dijon, France; Casasnovas Olivier, Department of Hematology, Hopital Le Bocage, Dijon, France; Elif Hindie and Brice Pauline, Department of Nuclear Medicine and Department of Hematology, Centre Hospitalier Universitaris St. Louis, Paris, France; Rachel Bar-Shalom and Eldad Dann, Department of Nuclear Medicine and Department of Hematology and Bone Marrow Transplantation, Rambam Medical Center, Haifa, Israel; Zaucha Jan Maciej, Hematology and BMT Department, Gdansk Medical University, Poland; Mikhaeel George and Lucy Pike, Department of Clinical Oncology and Division of Imaging Sciences, Guy’s and St. Thomas’ Hospital, London, United Kingdom; and Coleman Morton, Hemato-Oncology Division, Center for Lymphoma and Myeloma, Weill Cornell Medical Center, New York.

The following centers enrolled patients on behalf of Fondazione Italiana Linfomi (FIL): Salvi Flavia and Alessandro Levis, Hematology Department, SS Antonio e Biagio Hospital, Alessandria; Roberto Emanuele and Roberto Sorasio, Medical Physics and Hematology Department, S. Croce e Carle Hospital, Cuneo; Rigacci Luigi, Benedetta Puccini, and Luca Vaggelli, Department of Hematology and Department of Nuclear Medicine, University of Florence Careggi Hospital, Florence; Viviani Simonetta and Flavio Crippa, Medical Oncology 2 Department and PET Center, National Cancer Institut, Milan; Rusconi Chiara and Cristina Gabutti, Department of Hematology, and Emma Gay, Department of Nuclear Medicine, Niguarda Ca’ Granda Hospital, Milan; Massimo Federico, Stefano Luminari, and Bruno Bagni, Medical Oncology Department, Department of Onco-Hematology and Department of Nuclear Medicine, University of Modena; Luca Guerra, Department of Nuclear Medicine, and Enrico Pogliani and Bolis Silvia, Hematology Department, San Gerardo University Hospital, Monza; Giovanni

Semenzato, Renato Zambello, Anna Colpo, and Livio Trentin, Hematology Chair, University of Padua, Padua; Pulsoni Alessandro and Angela Rago, Hematology Chair, University “La Sapienza,” Rome; Agostino Chiaravalloti, Nuclear Medicine Department, University Tor Vergata, Rome; and Emanuele Nicolai, Institute of Diagnostic and Nuclear Medicine Development, Naples.

This study was presented orally at the annual meeting of the Society of Nuclear Medicine and Molecular Meeting in Miami, Florida, on June 9–12, 2012.

REFERENCES

1. Hutchings M, Loft A, Hansen M, et al. FDG-PET after two cycles of chemotherapy predicts treatment failure and progression-free survival in Hodgkin lymphoma. *Blood*. 2006;107:52–59.
2. Gallamini A, Rigacci L, Merli F, et al. The predictive value of positron emission tomography scanning performed after two courses of standard therapy on treatment outcome in advanced stage Hodgkin’s disease. *Haematologica*. 2006;91:475–481.
3. Gallamini A, Hutchings M, Rigacci L, et al. Early interim 2-[¹⁸F]fluoro-2-deoxy-D-glucose positron emission tomography is prognostically superior to international prognostic score in advanced-stage Hodgkin’s lymphoma: a report from a joint Italian-Danish study. *J Clin Oncol*. 2007;25:3746–3752.
4. 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.
5. Meignan M, Gallamini A, Haioun C. Report on the first international workshop on interim-PET scan in lymphoma. *Leuk Lymphoma*. 2009;50:1257–1260.
6. Boellaard R, O’Doherty MJ, Weber WA, et al. PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging*. 2010;37:181–200.
7. Chauvie S, Stancu A, Cerello P, Biggi A, Gallamini A. A clinical trial toolkit for diagnostic imaging exchange through the WEB [abstract]. *Eur J Nucl Med Mol Imaging*. 2009;36(suppl):S355.
8. Cohen J. A coefficient for agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37–46.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
10. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas*. 2007;1:77–89.
11. Kaplan EL, Meier P. Non parametric estimation from incomplete observations. *J Am Statistic Assoc*. 1958;53:4547–4581.
12. Cheson BD, Pfister B, Juweid ME, et al. Revised response criteria for malignant lymphoma. *J Clin Oncol*. 2007;25:579–586.
13. Biggi A, Chauvie S, Bianchi A, et al. Interim PET in Hodgkin lymphoma: comparison of the different criteria to evaluate chemotherapy response [abstract]. *Eur J Nucl Med Mol Imaging*. 2009;36(suppl):S252.
14. Horning SJ, Juweid ME, Schoder 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.
15. Barrington SF, Qian W, Somer EJ, et al. Concordance between four European centres of PET reporting criteria designed for use in multicentre trials in Hodgkin lymphoma. *Eur J Nucl Med Mol Imaging*. 2010;37:1824–1833.
16. Meignan M, Itti E, Bardet S, et al. Development and application of a real-time on-line blinded independent central review of interim PET scans to determine treatment allocation in lymphoma trials. *J Clin Oncol*. 2009;27:2739–2741.
17. Cerci JJ, Pracchia LF, Linardi CG, et al. ¹⁸F-FDG PET after 2 cycles of ABVD predicts event-free survival in early and advanced Hodgkin lymphoma. *J Nucl Med*. 2010;51:1337–1343.
18. Zinzani PL, Rigacci L, Stefoni V, et al. Early interim ¹⁸F-FDG PET in Hodgkin’s lymphoma: evaluation on 304 patients. *Eur J Nucl Med Mol Imaging*. 2012;39:4–12.
19. Gallamini A, Bianchi A, Borra A, et al. Dual-point FDG-PET: a novel scanning technique in Hodgkin lymphoma with bulky disease [abstract]. *J Clin Oncol*. 2012;30(suppl):8077P.
20. Kostakoglu L, Goldsmith SJ, Leonard JP, et al. FDG-PET after 1 cycle of therapy predicts outcome in diffuse large cell lymphoma and classic Hodgkin disease. *Cancer*. 2006;107:2678–2687.