^{18}F-fluoro-deoxyglucose positron emission tomography for post-treatment evaluation of malignant lymphoma: a systematic review

Josée M. Zijlstra
Gerda Lindauer-van der Werf
Otto S. Hoekstra
Lotty Hooft
Ingrid I. Riphagen
Peter C. Huijgens

Despite the increasing number of publications concerning ^{18}F-fluorodeoxyglucose positron emission tomography (FDG-PET) for post-treatment evaluation of lymphoma and the increasing availability of this novel diagnostic modality, its exact role in response assessment after therapy is still unknown. The aim of this study was to systematically review the literature regarding the diagnostic performance of dedicated FDG-PET in evaluation of first-line therapy of Hodgkin's disease and (aggressive) non-Hodgkin's lymphoma, and to calculate summary estimates of its sensitivity and specificity. The databases of PubMed and Embase were searched for relevant studies up to January 2004. Two reviewers independently assessed the methodological quality of each study. As a valid reference test, histology or follow-up of at least 12 months were accepted. A meta-analysis of the reported sensitivity and specificity of each study was performed. Fifteen studies, involving 705 patients, met the inclusion criteria. The studies had several design deficiencies. The majority of studies did not describe whether the reference test was interpreted without knowledge of the FDG-PET findings. In all studies, there was a description of the spectrum of patients included, i.e. all patients for post-treatment evaluation or only patients with substantial residual masses post-treatment. Pooled sensitivity and specificity for detection of residual disease in Hodgkin's lymphoma were 84% (95% CI 71-9192%) and 90% (95% CI 84-9394%), respectively. For non-Hodgkin's lymphoma, pooled sensitivity and specificity were 72% (95% CI 61-82%), and 100% (95% CI 97-100%), respectively. FDG-PET showed reasonable sensitivity and high specificity for evaluation of first-line therapy in Hodgkin's and in non-Hodgkin's lymphoma. Standardization of procedures is required before implementation in clinical practice.

Key words: Systematic review, Positron emission Tomography (PET), Lymphoma.

Haematologica 2006; 91:522-529

From the VU University Medical Center, Departments of Hematology (JMZ, GLvdW, PCH), Nuclear Medicine and PET research (JMZ, OSH), Clinical Epidemiology and Biostatistics (LH, OSH), Medical Library (IIR), P.O. Box 7057, 1007 MB Amsterdam, The Netherlands.

Correspondence:
Josée M. Zijlstra, VU University Medical Center, PO Box 7057 1007 MB Amsterdam, The Netherlands.
E-mail j.zijlstra@vumc.nl

Positron emission tomography (PET) is a rapidly developing new imaging technique for the diagnosis and staging of cancer. It combines excellent scanner performance (sensitivity, resolution) and a radioactive tracer with a favorable biodistribution and high affinity for cancer cells. Whole body ^{18}F-fluorodeoxyglucose (FDG)-PET has shown the ability to detect small tumor deposits with a diagnostic accuracy exceeding conventional imaging modalities.[1,2] New diagnostic technologies, such as PET, tend to diffuse rapidly into clinical practice before adequate evaluation of their clinical potential has taken place.[3,4] In malignant lymphoma, PET appears to be a useful diagnostic tool but there is no consensus regarding its place in treatment strategies. Potential indications include the detection of occult disease in small lymph nodes and of extranodal localizations at the time of presentation, assessment of response during treatment and evaluation following treatment. A number of publications claim a prominent role for PET, especially in patients with residual masses detected by computed tomography (CT) scanning following treatment.

The objective of this systematic review is to determine the diagnostic accuracy of dedicated (full ring) PET using FDG for response assessment after first-line therapy in lymphoma patients. To this end, all relevant scientific reports were identified using a comprehensive search strategy.[5,6] Accepted methodological standards for the evaluation of diagnostic tests were applied.[7-9] We summarize the existing data on the relevance of PET scanning in lymphoma following first-line treatment, as a first step towards the development of guidelines for the effective use of PET.

Design and Methods

Literature search for the identification of studies

A search of the bibliographic databases PubMed/MEDLINE (from 1966) and Embase (from 1988) was conducted up to January 2004, without any language restrictions. The search strategy for the identification of primary studies regarding diagnostic tests[9,10] was run in conjunction with a specific search for PET, FDG[6,11] and lymphoma, adapted for each database. All searches were performed using controlled indexing terms (MeSH in MEDLINE and EMTree in Embase) and free text words. To identify studies regarding lymphoma, the MeSH terms *hodgkin disease, lymphoma, non-hodgkin* and *soft tissue neoplasms* were used in MEDLINE whereas the EMTree terms *lymphoma, soft tissue neoplasms, hodgkin disease* and *non-hodgkin lymphoma* were applied in Embase. We augmented this search by manually reviewing the reference lists of the identified studies and relevant review articles. Unpublished data and conference proceedings were not included in this review.

Study selection

Criteria for inclusion of studies were: 1) histologically proven Hodgkin's disease (HD) or aggressive non-Hodgkin's lymphoma (NHL), 2) evaluation of post-treatment patients following first-line therapy, 3) the use of dedicated (ring) PET using FDG, and 4) a study population of at least ten patients. Exclusion criteria were: 1) central nervous system and AIDS related lymphoma, 2) the use of radiopharmaceuticals other than FDG, 3) animal studies and 4) abstracts, reviews, editorials, letters and comments. Using the above-mentioned inclusion and exclusion criteria, two reviewers (OSH and JMZ) independently selected the studies for possible inclusion in the review by checking titles and abstracts. All studies considered eligible, as well as studies for which it was unclear whether they were eligible, were retrieved and the final decision was based on the full article. Disagreement was resolved by consensus.

Methodological quality assessment

For the results of a PET accuracy study to be internally valid, an independent, blind comparison with a valid reference test to avoid review bias is essential. This reference test has to be measured in all patients independently of the results of the PET scan to avoid verification and work-up bias and the reference test has to be applied in a standardized manner.[7,8,12] Diagnostic accuracy is best determined by comparing test results with an appropriate so-called gold standard. In oncology, histological proof of presence or absence of viable tumor is an accurate reference test. However, in an often diffuse disease such as lymphoma, this approach may not be appropriate: surgical exploration of all initially involved sites is impossible, and since conventional staging techniques may understage patients, such surgical specimens would still be only partially representative. Due to sampling errors, especially after therapy, the same is true for biopsies unless they contain viable tumor. Therefore, clinical follow-up should be added: in this review, we considered a clinico-radiological follow-up of 12 months after completed therapy to be the minimum acceptable period in patients without histologically documented persistent disease immediately after first-line therapy. Consequently, we used the following criteria to decide whether patients had active disease or not: (i) tumor-positive biopsy or (ii) clinico-radiological follow-up of at least 12 months in all cases without positive biopsies. Three reviewers (JMZ, GLW, OSH) independently assessed the methodological quality of the selected studies using the criteria list recommended by the Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Some items on the list were modified for this specific review (Table 1). Internal validity criteria (IV) were scored as *positive* (adequate methods), *negative* (inadequate methods, potential bias), or *unclear* if insufficient information was present on a specific item. If authors did not explicitly state that the choice of patients who were assessed by the reference test(s) was independent of the PET result, we scored this item (IV3) as negative. The criteria for external validity (EV) were scored positive if sufficient information was provided to judge generalizablility of findings. The standard performance of FDG-PET was scored positive when the type of PET camera, the dose of FDG, the time between injection and scanning, and the methods of image reconstruction were described. Many studies have been performed on mixed lymphoma populations. If possible, data were extracted for Hodgkin's and non-Hodgkin's lymphoma separately.

Table 1. List of criteria used to assess the methodological quality of the studies.

Criteria of validity	*Positive Score*
Internal Validity (IV)	
1. Valid reference test FU > 12 months	Positive histology or clinico-radiological
2. Blind measurement of PET without knowledge of reference test	Mentioned in publication
3. Blind measurement of reference test without knowledge of PET	Mentioned in publication
4. Avoidance of verification bias test independent of PET results	Assessment by reference
5. Prospective study	Mentioned in publication
External Validity (EV)	
1. Description of spectrum of disease	Mentioned in publication
2. Demographic information	Age and gender given
3. Inclusion criteria	Mentioned in publication
4. Exclusion criteria	Mentioned in publication
5. Avoidance of selection bias	Consecutive patients
6. Standard performance of FDG-PET	Type of scanner, dose of FDG, time interval, reconstruction
7. PET interpreted independently of clinical information	Mentioned in publication

FDG: ^{18}F-fluorodeoxyglucose; PET: positron emission tomography; FU: follow-up.

Quantitative analysis (meta-analysis)

Data were reported according to the guidelines for meta-analyses evaluating diagnostic tests.[9] For each study, the sensitivity (proportion of tumor-positive patients correctly identified by PET), specificity (proportion of tumor-negative patients correctly identified by PET), positive predictive value (PPV), negative predictive value (NPV) and their 95% confidence intervals (CI) of FDG-PET in the detection of residual lymphoma were calculated from the original data. We also reported the positive and negative likelihood ratios. The likelihood ratio (LR) is the likelihood that a given test result would be expected in a patient with the target disorder compared to the likelihood that the same result would be expected in a patient without that disorder. These ratios provide a measure of the discriminating power of FDG- PET. The reviewers independently constructed 2x2 contingency tables with numbers of patients with and without relapse vs. the PET results. Studies that did not present data in sufficient detail to calculate the estimates were excluded from statistical pooling. We added 0.5 to a cell frequency of zero to calculate the estimates.

The Q test was used to assess the homogeneity (the consistency of findings) among studies. If homogeneity of both sensitivity and specificity was not rejected (p>0.05), summary estimates were calculated.[13] Because the Q test has limited power and may fail to detect heterogeneity, statistical pooling of the outcome measures was performed with a random-effects model[14] if the p value of the

Q test was between 0.05 and 0.10. A fixed-effects model was used when p>0.10. We used weighted models in which the weight of each study is its sample size. Analyses were performed using MetaDisc software (version Beta 1.1.0), specially designed for the meta-analysis of diagnostic and screening tests.

Results

Literature search

The search strategy yielded 447 publications in EMBASE and 450 in MEDLINE; 281 studies were identified in both databases. From the resulting 616 studies, 589 were excluded after reviewing the information provided in the title and abstract. Reviewing the full articles of the 27 remaining studies resulted in exclusion of another 12 articles due to an overlap in study population (four studies), inappropriate setting (staging after initial presentation, one study), use of a dual-head gamma camera (two studies) and missing essential data regarding PET findings (five studies; eg. multiple PET scans during follow-up without the specification of which PET scan was used vs. the reference test, follow-up of patients with negative PET scans only) (Figure 1).[15-18] Finally, 15 studies involving 705

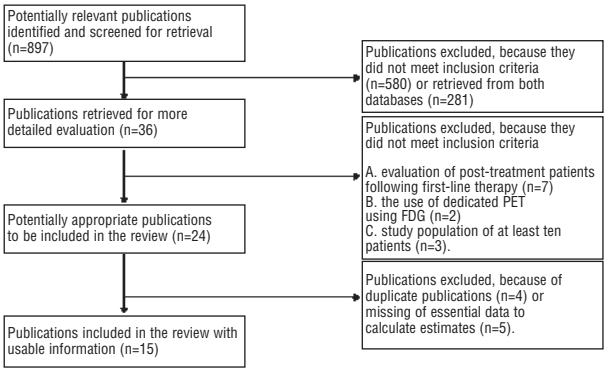

Figure 1. Flow diagram of publications included in the review according to QUOROM principles of publication selection.

Table 2. Characteristics of studies included: FDG-PET for post-treatment evaluation of malignant lymphoma.

Study	Year	No. of Patients	Design Spectrum*	HD	NHL	Treatment	Reference test	Endpoints
Bangerter[19]	1999	58 subgroup	Retrospective Broad	N=45	N=43	ChT/ ChT+RT	FU > 12 mo. Mean 36 mo.	Accuracy
Bangerter[20]	1999	36	Prospective Small; > 1 cm	N=14	N=22	ChT	FU > 11 mo. Median 28 mo. PA n=5	Accuracy
Cremerius[21]	2001	41 subgroup	Retrospective Broad	N=22	N=34	ChT/ ChT+RT	FU > 6 mo. Median 21 mo. PA n=2	PFS
De Wit[22]	2001	33 subgroup	Prospective Broad	N=37	–	ChT+RT	FU > 1.8 mo. Median 26 mo. PA n=5	Accuracy OS, DFS
Dittmann[23]	2001	26 subgroup	Prospective Small; > 1 cm	N=26	–	ChT+RT	FU > 6 mo. Mean unknown PA n=2	Accuracy
Hueltenschmidt[24]	2001	47 subgroup	Retrospective Broad	N=51	–	ChT/ ChT+RT	FU > 3 mo. PA n=8 Mean 20 mo.	Accuracy
Jerusalem[25]	1999	54	Prospective Broad	N=19	N=35	ChT/ ChT+RT	FU median 21 mo, range unclear	Accuracy PFS, OS
Mikhaeel[26]	2000	32	Retrospective Small; > 2 cm	N=15	N=17	ChT/ ChT+RT	FU > 18 mo. Median 38 mo. PA n=10	Accuracy
Mikhaeel[27]	2000	45 subgroup	Retrospective Broad	–	N=45	ChT/ ChT+RT	FU > 7 mo. Median 30 mo. PA n=15	Accuracy RFS
Naumann[28]	2001	58	Prospective + Retrospective Small; > 1 cm Also > 0.5 cm	N=43	N=15	ChT/ ChT+RT	FU > 15 mo. Median 34 mo. PA n=unknown	Accuracy PFS
Spaepen[29]	2001	93	Retrospective Broad	–	N=93	ChT	FU > 11 mo. Median 22 mo. PA n=8	PFS
Spaepen[30]	2001	60	Retrospective Broad	N=60	–	ChT/ ChT+RT	FU > 12 mo. Median 32 mo. PA n=10	PFS
Stumpe[31]	1998	50 subgroup	Retrospective Broad	N=35	N=15	unknown	FU > 6 mo.	Accuracy
Weihrauch[32]	2001	28	Prospective Small; > 2 cm	N=28	–	ChT/ ChT+RT	FU > 16 mo. Median 28 mo PA n=9	Accuracy DFS
Zinzani[33]	1999	44	Retrospective Broad	N=13	N=31	ChT/ ChT+RT	FU > 8 mo. Median 18 mo	RFS

*ChT: chemotherapy; RT: radiotherapy; FU: follow up; OS: overall survial; PFS: progression-free survival; DFS: disease-free survival; RFS: relapse-free survival Subgroup: publication describes FDG-PET in staging and/or restaging, data on PET in post-treatment evaluation were extracted. *small spectrum : only patients with 'residual masses' included (NB. Variable definitions of 'residual mass'), broad spectrum: inclusion of patients irrespective of CT findings.*

Table 3. Quality assessment of the diagnostic studies included.

Study	year	Internal Validity					External Validity							Total pos. IV score	Total pos. EV score
		IV1	IV2	IV3	IV4	IV5	EV1	EV2	EV3	EV4	EV5	EV6	EV7		
Bangerter[19]	1999	+	0	–	–	+	+	+	+	+	+	+	+	2	7
Bangerter[20]	1999	+	+	–	–	–	+	+	–	–	0	+	+	2	4
Cremerius[21]	2001	+	+	–	–	–	+	+	+	+	0	+	+	2	6
De Wit[22]	2001	–	+	+	–	+	+	+	+	–	0	+	–	3	4
Dittmann[23]	2001	–	+	–	–	+	+	+	+	–	0	+	–	2	4
Hueltenschmidt[24]	2001	+	+	–	–	–	+	+	–	–	0	+	–	2	3
Jerusalem[25]	1999	+	+	–	–	+	+	+	–	+	–	+	+	3	5
Mikhaeel[26]	2000	+	+	–	–	–	+	+	+	–	0	+	–	2	4
Mikhaeel[27]	2000	+	0	–	–	–	+	+	+	–	0	+	+	1	5
Naumann[28]	2001	+	+	–	–	+	+	+	+	–	+	+	–	3	5
Spaepen[29]	2001	+	+	–	–	–	+	+	+	–	–	+	+	2	5
Spaepen[30]	2001	+	+	–	–	–	+	+	+	–	–	+	+	2	5
Stumpe[31]	1998	0	0	–	0	–	–	+	–	–	0	+	0	0	2
Weihrauch[32]	2001	+	+	–	–	+	+	+	+	–	0	+	–	3	4
Zinzani[33]	1999	–	+	–	–	–	+	+	+	–	0	+	–	1	4

IV1-IV5: five criteria for internal validity; EV1-EV7: seven criteria for external validity (see Table 1); + = yes; - = no; 0 = unclear.

Table 4. Definitions of a positive PET scan.

Bangerter[19]	Any foci of increased FDG uptake over background uptake not located in area of physiologically increased uptake considered suspicious for lymphoma
Bangerter[20]	Any clearly delineated uptake in the hilar or mediastinal regions was considered as suspected lymphoma
Cremerius[21]	All foci of elevated FDG uptake which could not be explained by physiological or non-specific uptake were suspected of residual viable lymphoma
De Wit[22]	FDG-uptake was classified as PD if new sites with FDG accumulation were involved compared to prior morphological staging, or FDG uptake was more intense
Dittmann[23]	Focally increased uptake, exceeding that of the surrounding tissue was interpreted as viable tumor tissue
Hueltenschmidt[24]	Any focus of FDG uptake exceeding the normal FDG uptake in the respective area was considered to represent lymphoma involvement
Jerusalem[25]	Any focus of increased FDG uptake over background not located in areas of normal FDG uptake and/or excretion was considered positive for tumor
Mikhaeel[26]	Not defined
Mikhaeel[27]	Residual increased FDG uptake in previously diagnosed disease sites or the appearance of new uptake indicative of progressive disease
Naumann[28]	Pathologically raised FDG uptake outside as well as in the region of the residual mass with SUV > 3
Spaepen[29]	Any focal or diffuse area of increased activity in a location incompatible with normal anatomy and suspect for residual disease
Spaepen[30]	Idem
Stumpe[31]	All circumscribed lesions in abnormal locations with low or high FDG uptake were considered to be pathological
Weihrauch[32]	Focally increased uptake in the mediastinum was considered positive if visually detectable
Zinzani[33]	Not defined

patients were included in this systematic review.[19-33] There was no disagreement between the reviewers regarding the inclusion of the articles. The characteristics of the included studies are presented in Table 2. The total number of patients per study ranged from 28 to 93, and the patients' age ranged from 2 to 88 years. Eight studies included patients with Hodgkin's as well as non-Hodgkin's lymphoma (total number of patients: 418; 206 of whom with Hodgkin's lymphoma), five only included Hodgkin's disease (total number of patients: 202), and two only non-Hodgkin's lymphoma (total number of patients: 138). The prevalence of relapse varied from 14-46% (Table 6). Eight studies comprised both Hodgkin's and aggressive non-Hodgkin's lymphoma and in four of these, it was possible to extract data for both of patients groups separately.[26,28,31,33] These studies have been included in the subgroup analysis for Hodgkin's and non-Hodgkin's lymphoma. The data of each study are presented in Tables 5 and 6.

The definitions of a positive PET scan differed in the included studies (Table 4), and only a minority of authors used a semi-quantitative measurement of FDG-uptake besides visual assessment.[21,28] The reference test consisted of histopathology on biopsies in only a minority of patients, and of radiological and clinical follow-up in the majority.

Methodological quality assessment

Methodological quality was assessed by 12 items for each of the 15 selected studies. There was disagreement in 11 of the 180 scores (6%), solved by consensus. The scores for internal and external validity are presented in Table 3. All studies except three had a valid reference test (histology or follow-up of at least 12 months), but in another four studies not every patient had been followed up for 12 months. Even though the reference test was valid in most studies, statements about blinding of clinicians to PET results were often lacking. Six studies were prospective (40%), and in only two studies (13%) were consecutive patients included. All studies provided a description of the spectrum of patients included, i.e. all patients for post-treatment evaluation (broad spectrum) or only patients with residual masses post-treatment (narrow spectrum). In the majority of studies (73%), the inclusion criteria were described, but only a minority (20%) also described the exclusion criteria. The total score for the combined internal and external validity, expressed as a fraction of the maximum score, ranged from 17% to 75%, with a mean

Table 5. Parameters of diagnostic accuracy of PET for post-treatment evaluation of lymphoma.

Study	HD	NHL	TRP (95%CI) (sensitivity) Mixed	FPR (95%CI) (1-specificity) Mixed	TRP (95%CI) (sensitivity) NHL	FPR (95%CI) (1-specificity) NHL	TRP (95%CI) (sensitivity) HD	FPR (95%CI) (1-specificity) HD
Bangerter[19]	N=45	N=43	0.86 (0.49-0.97)	0.04 (0.01-0.13)				
Bangerter[20]	N=14	N=22	0.71 (0.36-0.92)	0.14 (0.06-0.31)				
Cremerius[21]	N=22	N=34	0.84 (0.62-0.94)	0.14 (0.05-0.33)				
De Wit[22]	N=37	–					1. 0 (0.72-1.0)	0. 22 (0.10-0.42)
Dittmann[23]	N=26	–					0.87 (0.53-0.98)	0.06 (0.01-0.26)
Hueltenschmidt[24]	N=51	–					0.95 (0.75-0.99)	0.11 (0.04-0.27)
Jerusalem[25]	N=19	N=35	0.43 (0.21-0.67)	0 (0-0.09)				
Mikhaeel[26]	N=15	N=17	0.80 (0.49-0.94)	0.05 (0.01-0.22)	0.71 (0.29-0.96)	0.00 (0-0.31)	1.0 (0.44-1.0)	0.08 (0.02-0.35)
Mikhaeel[27]	–	N=45			0.60 (0.32-0.84)	0.00 (0-0.12)		
Naumann[28]	N=43	N=15	0.86 (0.49-0.97)	0.12 (0.06-0.23)	0.83 (0.36-1.0)	0.00 (0-0.34)	1.0 (0.21-1.0)	0.14 (0.07-0.28)
Spaepen[29]	–	N=93			0.70 (0.53-0.84)	0.00 (0-0.06)		
Spaepen[30]	N=60	–					0.50 (0.24-0.76)	0 (0 -0.07)
Stumpe[31]	N=35	N=15	0.87 (0.68-0.95)	0.03 (0.01-0.16)				
Weihrauch[32]	N=28	–					0.67 (0.35-0.88)	0.20 (0.08-0.42)
Zinzani[33]	N=13	N=31	0.93 (0.68-0.99)	0 (0-0.11)	0. 91 (0.59-1.0)	0.02 (0-0.21)	1.0 (0.44-1.0)	0 (0-0.28)

Mixed: HD and NHL together.

Table 6. Parameters of diagnostic accuracy of PET for post-treatment evaluation of lymphoma.

Study	Patients	HD	NHL	Prev of relapse	PPV	95%CI	NPV	95%CI
Bangerter[19]	N=58	N=45	N=43	14%	85.7	48.7-97.4	96.1	86.8-98.9
Bangerter[20]	N=36	N=14	N=22	19%	71.4	35.9-91.8	86.2	69.4-94.5
Cremerius[21]	N=41	N=22	N=34	46%	84.2	62.4-94.5	86.4	66.7-95.3
De Wit[22]	N=33	N=37	–	30%	100	72.2-100	78.3	58.1-90.3
Dittmann[23]	N=26	N=26	–	31%	87.5	52.9-97.8	94.4	74.2-99.0
Hueltenschmidt[24]	N=47	N=51	–	40%	94.7	75.4-99.1	89.3	72.8-96.3
Jerusalem[25]	N=54	N=19	N=35	26%	42.9	21.4-67.4	100	91.2-100
Mikhaeel[26]	N=32	N=15	N=17	32%	80	49.0-94.3	95.5	78.2-99.2
Mikhaeel[27]	N=45	–	N=45	33%	60	35.7-80.2	100	88.6-100
Naumann[28]	N=58	N=43	N=15	12%	85.7	48.7-97.4	88.2	76.6-94.5
Spaepen[29]	N=93	–	N=93	40%	70.3	54.2-82.5	100	93.6-100
Spaepen[30]	N=60	N=60	–	17%	50	23.7-76.3	100	92.9-100
Stumpe[31]	N=50	N=35	N=15	42%	87	76.9-95.5	96.8	83.8-99.4
Weihrauch[32]	N=28	N=28	–	31%	66.7	35.4-87.9	80	58.4-91.9
Zinzani[33]	N=44	N=13	N=31	32%	92.9	68.5-98.7	100	88.6-100

Prev: prevalence; PPV: positive predictive value; NPV: negative predictive value.

of 54%. Since the PET acquisition parameters might affect the diagnostic performance, we have summarized these data in Table 7. There was, however no apparent association between each of the measures and the results.

Quantitative analysis (meta-analysis)

A fixed-effects model was used for the five relevant studies of NHL (QSE= 3,89; degrees of freedom(df) = 4; p=0.42; QSP= 1,15; df= 4; p=0.89). The overall sensitivity and specificity were 72% (95% CI 61-82) and 100% (95% CI 94-100), respectively (Figures 2A and 2B). The overall negative likelihood ratio (LR-) was 0.28 (95% CI 0.20-0.41) and the overall positive likelihood ratio (LR+) was 37 (95% CI 11-127). A random-effects model was used for the seven relevant studies of HD (QSE= 11.01; df= 6; p=0.09; QSP= 12,84; df= 6; p=0.05). The overall sensitivity and specificity were 84% (95% CI 71-92) and 90% (95% CI 84-94), respectively (Figures 3A and 3B). The overall LR- was 0.26 (95% CI 0.12-0.58) and the overall LR+ was 5.6 (95% CI 3.46-9.13). Several studies reported survival measures as a function of PET results (Table 2). Meta-analysis of such data was impossible due to the variability of the applied endpoints (overall survival, progression-free survival, relapse-free survival, disease-free survival) and the lack of specified data for Hodgkin's and non-Hodgkin's lymphoma as separate entities. There was no apparent inverse relation between sensitivity and specificity for either HD or NHL (Spearman -0.4 and -0.3, respectively).

Discussion

This systematic review analyzed the diagnostic accuracy of FDG-PET for post-treatment evaluation of lymphoma patients after first-line chemotherapy. Even though we identified several methodological deficiencies, the study results consistently show that FDG-PET has a very high specificity in this setting for a pooled sensitivity (vs. the gold standard of tumor positive biopsy/clinical follow-up of at least one year) of 72% and 84% for aggressive NHL and HD, respectively. The width of the 95% confidence intervals of these estimates obviously reflects the limited number of patients in most studies (mean of 50). Using the obtained point estimates of sensitivity and specificity, we calculated the predicted post-PET probabilities of viable lymphoma (Figure 4) as a function of pre-test probability (prevalence). These data suggest that in the case of a 50% probability of persisting viable tumor after first-line therapy for aggressive NHL, the probability of having persistent viable tumor after a positive PET scan is 97%, vs. 22% in the case of a negative PET result. Applying a 15% pre-test probability of relapse in HD, the projected probability of absence of relapse in the case of a negative PET result is 3%, vs. 60% in the case of a positive PET scan.

The studies included in our analysis were of moderate methodological quality, with a 40% score for internal validity, and a 64% score for external validity. For most studies, it was unclear whether the PET-scan results were blinded or could have influenced the clinical follow-up,

Table 7. Characteristics of PET studies.

Study	Spatial resolution (FWHM)*	FDG uptake period	FDG dose	Acquisition time (emission time/bed)	Attenuation correction	Exclusion hyperglycemia	PET interpretation	Interval end of therapy - PET
Bangerter[19]	7 mm p	50-60 min	270 MBq (250-350)	15 min	yes	unknown	visual	> 8 weeks
Bangerter[20]	7 mm p	50-60 min	200-300 MBq mean 270	15 min	yes	unknown	visual T-NT ratio	mean 2 months (1-5)
Cremerius[21]	7 mm p	45-60 min	220 ± 70 MBq	Unknown	yes	no	visual SUV	0.5-3 months
De Wit[22]	12 mm p	60 min	250-400 MBq	10 min	no	unknown	visual	mean 10 weeks (2-40)
Dittmann[23]	~5 mm s	45-60 min	400 MBq	5 min	yes	unknown	visual SUV	mean 2 months (0.7-5.5)
Hueltenschmidt[24]	5 mm s	60 min	370 MBq	8-10 min	yes	yes	visual	4-6 weeks
Jerusalem[25]	8 mm p	45-90 min	222-296 MBq	4 min	no	unknown	visual	1-3 months
Mikhaeel[26]	8-12 mm p	30 min	350 MBq	10 min	yes **	unknown	visual	<2 months
Mikhaeel[27]	8-12 mm p	30 min	350 MBq	10 min	yes **	unknown	visual	4-6 weeks
Naumann[28]	4 mm s	45-60 min	300-370 MBq	8 min	yes	unknown	visual SUV	median 12 weeks (1-24)
Spaepen[29]	~6 mm s	60 min	370-555 MBq	4 min	no	yes	visual	1-3 months
Spaepen[30]	~6 mm s	60 min	150-555 MBq	4 min	no	yes	visual	1-3 months
Stumpe[31]	~5 mm s	40 min	300-350 MBq	5 min	no	unknown	visual	unknown
Weihrauch[32]	6 mm s	unknown	370 MBq	unknown	yes only thorax	unknown	visual SUV	< 4 months
Zinzani[33]	6 mm s	45 min	444 MBq	unknown	unknown	unknown	visual	> 1-2 months

*FWHM: full width at half maximum; p: post-reconstruction resolution, s: system resolution; ** on indication.*

which might have led to diagnostic bias. In general, this tends to overestimate the diagnostic accuracy.[5,34] However, we expect that this is likely to affect especially the timing of relapse diagnosis rather than the diagnosis of recurrence itself, since these conclusions were drawn upon standard clinical methodology. Another type of bias is related to the patients' selection. If it was stated that consecutive series of patients were included in a study, is was often unclear whether these patients were seen consecutively at the nuclear medicine department (with potential referral bias) or at the hematology department. In our analysis, studies with a narrow spectrum (only patients with residual masses) showed diagnostic accuracy comparable to that of studies with a broad spectrum (all patients with and without residual masses included), but the number of included patients may have been too low to detect such a potential difference.

A particular problem was the difficulty in excluding overlap between study populations of several publications from one author or one institution. Cremerius *et al.,* published three papers[21,35,36] comprising partly the same patients. After personal communication, the most recent article was included in this review. Likewise, among the publications from Bumann *et al.*[37] and De Wit *et al.*,[22,38] the most recent article was included. Zinzani *et al.* have published an update of their series of patients with abdominal lymphoma, comprising another 15 patients.[39] In this new manuscript, data on NHL and HD could not be extracted separately, and were therefore not included in this systematic review. The two publications from Bangerter *et al.*[19,20] had a different design (retrospective versus prospective) and described different patient populations, so they were both included in this analysis. The same holds true for the articles by Spaepen *et al.*[29,30] and for the articles by Mikhaeel *et al.*[26,27]

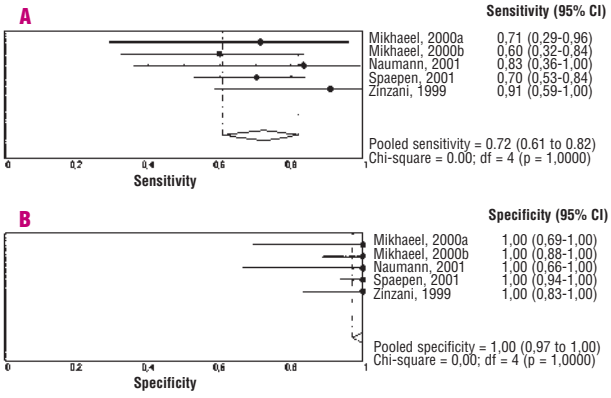

Figure 2. A. Sensitivities and 95% confidence intervals for studies assessing the diagnostic accuracy of FDG PET in patients with NHL. B. Specificity and 95% confidence intervals for studies assessing the diagnostic accuracy of FDG PET in patients with NHL. *The diamond represents the 95% CI of the pooled estimate.

The duration of follow-up differed between patients within each study. There is no consensus regarding the

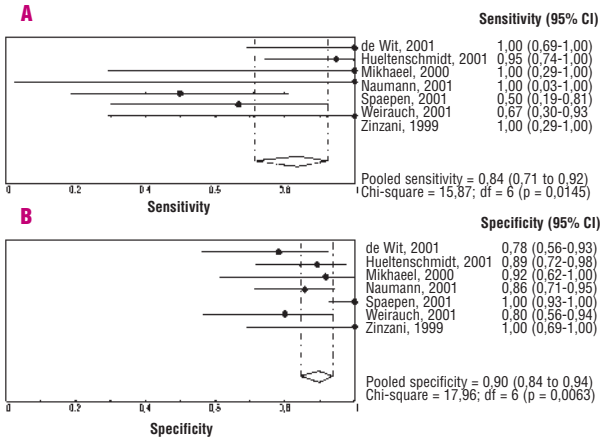

Figure 3. A. Sensitivities and 95% confidence intervals for studies assessing the diagnostic accuracy of FDG PET in patients with HD. B. Specificity and 95% confidence intervals for studies assessing the diagnostic accuracy of FDG PET in patients with HD. *The diamond represents the 95% CI of the pooled estimate.

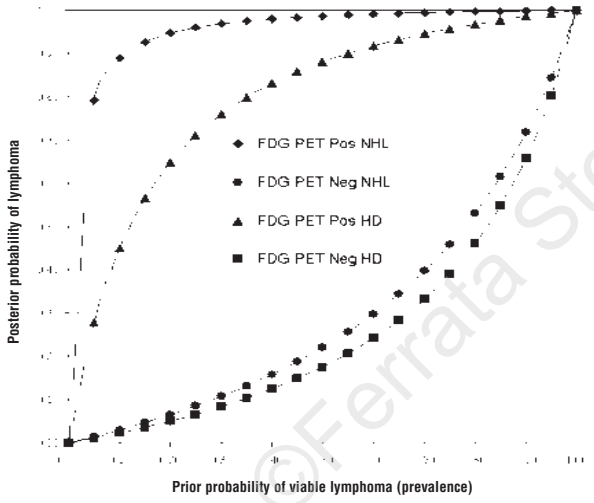

Figure 4. Predicted post-test probabilities of viable lymphoma after first-line therapy as a function of FDG PET results in non-Hodgkin's and Hodgkin's lymphoma. Post-test probabilities are shown as a function of pre-test probability in patients with positive FDG PET results in non-Hodgkin's lymphoma (NHL; diamonds) and Hodgkin's lymphoma (HD, triangles), and in patients with negative FDG PET results in NHL (circles) and HD (squares).

extent and duration of follow-up necessary to be valid as an accurate reference test. We considered 12 months to be the minimum period of follow-up (i.e. PET was truly negative if the patient is clinically free of recurrence after 12 months). Obviously, the sensitivity of PET is expected to be inversely related to the duration of follow-up, but the presented data did not reveal such trend, perhaps because on an aggregate level the follow-up intervals were relatively consistent.

The timing of a post-treatment PET-scan is another essential aspect. In several studies, the interval between the last administration of chemotherapy and the post-treatment PET-scan was not well documented or variable.[22,28,31] Evaluation with PET 40-100 weeks after treatment might suggest a bias because patients may have been referred for PET because of suspected relapse. The criteria for PET positivity and negativity were not completely consistent among the studies, even though it appears that this often likely reflected semantic differences (Table 4) rather than actual ones. However, for the implementation of FDG-PET in response monitoring of lymphoma patients, a uniform definition of a positive PET scan is essential. Only one publication[28] mentioned questionable PET findings for Hodgkin's and non-Hodgkin's lymphoma evaluation.

Even though PET interpretation is typically visual and therefore subject to observer variation, this aspect of reproducibility was assessed in only one study.[20] Finally, it was not always clear whether baseline PET scans were available or required for interpretation of the response to therapy, and this should be investigated as well.

The reviewed studies investigated diagnostic accuracy and not the impact of PET results on management or on patients' outcome resulting from management changes. The lack of such studies (e.g. randomizing patients to diagnostic /management strategies with vs. without PET) is a limitation of the currently available evidence.

Conclusion

The presently available evidence on the diagnostic performance of FDG-PET in evaluating the response to first-line therapy for HD and NHL is useful. Standardization of procedures is required before implementation in clinical practice. FDG-PET appears to be the most helpful non-invasive modality for differentiating tumor recurrence from fibrosis when CT scanning shows a residual mass. If abnormal FDG-uptake is seen, further investigation is mandatory. In the case of a negative PET-scan, no further investigations at that particular time point are necessary, but minimal residual disease and the risk of a late relapse cannot be completely excluded.

JMZ and OSH: conception and design of the study, writing of the manuscript; JMZ and GLW: collection, analysis and interpretation of data; LH and OSH: performed the statistical analysis and critical review of the manuscript; IIR and GLW: responsible for literature search and critical review of the manuscript; PCH: contributed fruitfully in revision of the manuscript and generation of the final version.

Manuscript received November 14, 2005. Accepted February 27, 2006.

References

1. Hoh CK, Schiepers C, Seltzer MA, Gambhir SS, Silverman DH, Czernin J, et al. PET in oncology: will it replace the other modalities? Semin Nucl Med 1997;27:94-106.
2. Kostakoglu L, Goldsmith SJ. Fluorine-18 fluorodeoxyglucose positron emission tomography in the staging and follow-up of lymphoma: is it time to shift gears? Eur J Nucl Med 2000;27: 1564-78.
3. Mowatt G, Bower DJ, Brebner JA, Cairns JA, Grant AM, McKee L. When and how to assess fast-changing technologies: a comparative study of medical applications of four generic technologies. Health Technol Assess 1997; 1:i-149.
4. Robert G, Milne R. Positron emission tomography: establishing priorities for health technology assessment. Health Technol Assess 1999;3:1-54.
5. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol 1999;52:943-51.
6. Mijnhout GS, Hooft L, van Tulder MW et al. How to perform a comprehensive search for FDG-PET literature. Eur J Nucl Med 2000;27:91-7.
7. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1994;271:389-91.
8. Kelly S, Berry E, Roderick P, Harris KM, Cullingworth J, Gathercole L, et al. The identification of bias in studies of the diagnostic performance of imaging modalities. Br J Radiol 1997;70: 1028-35.
9. Deville WL, Buntinx F, Bouter LM et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002;2:9.
10. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000;53:65-9.
11. Mijnhout GS, Riphagen II, Hoekstra OS. Update of the FDG PET search strategy. Nucl Med Commun 2004; 25: 1187-9.
12. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645-51.
13. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. Med Decis Making 1993;13:253-7.
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177-88.
15. Wiedmann E, Baican B, Hertel A et al. Positron emission tomography (PET) for staging and evaluation of response to treatment in patients with Hodgkin's disease. Leuk Lymphoma 1999;34:545-51.
16. Lavely WC, Delbeke D, Greer JP et al. FDG PET in the follow-up management of patients with newly diagnosed Hodgkin and non-Hodgkin lymphoma after first-line chemotherapy. Int J Radiat Oncol Biol Phys 2003;57: 307-15.
17. Jerusalem G, Beguin Y, Fassotte MF et al. Early detection of relapse by whole-body positron emission tomography in the follow-up of patients with Hodgkin's disease. Ann Oncol 2003; 14:123-30.
18. Filmont JE, Vranjesevic D, Quon A, Margolis DJ, Ko F, Safaei A, et al. Conventional imaging and 2-deoxy-2-[18F]fluoro-D-glucose positron emission tomography for predicting the clinical outcome of previously treated non-Hodgkin's lymphoma patients. Mol Imaging Biol 2003;5:232-9.
19. Bangerter M, Kotzerke J, Griesshammer M, Elsner K, Reske SN, Bergmann L. Positron emission tomography with 18-fluorodeoxyglucose in the staging and follow-up of lymphoma in the chest. Acta Oncol 1999; 38:799-804.
20. Bangerter M, Moog F, Griesshammer M, Elsner K, Kotzerke J, Heimpel H, et al. Role for whole body FDG-PET imaging in predicting relapse of malignant lymphoma in patients with residual masses after treatment. Radiography 1999;5:155-63.
21. Cremerius U, Fabry U, Neuerburg J, Zimny M, Bares R, Osieka R,et al. Prognostic significance of positron emission tomography using fluorine-18-fluorodeoxyglucose in patients treated for malignant lymphoma. Nuklearmedizin 2001;40:23-30.
22. de Wit M, Bohuslavizki KH, Buchert R, Bumann D, Clausen M, Hossfeld DK. 18FDG-PET following treatment as valid predictor for disease-free survival in Hodgkin's lymphoma. Ann Oncol 2001;12:29-37.
23. Dittmann H, Sokler M, Kollmannsberger C, Dohmen BM, Baumann C, Kopp A, et al. Comparison of 18FDG-PET with CT scans in the evaluation of patients with residual and recurrent Hodgkin's lymphoma. Oncol Rep 2001; 8:1393-9.
24. Hueltenschmidt B, Sautter-Bihl ML, Lang O, Maul FD, Fischer J, Mergenthaler HG, et al. Whole body positron emission tomography in the treatment of Hodgkin disease. Cancer 2001; 91: 302-310.
25. Jerusalem G, Beguin Y, Fassotte MF, Najjar F, Paulus P, Rigo P, et al. Whole-body positron emission tomography using 18F-fluorodeoxyglucose for posttreatment evaluation in Hodgkin's disease and non-Hodgkin's lymphoma has higher diagnostic and prognostic value than classical computed tomography scan imaging. Blood 1999;94: 429-33.
26. Mikhaeel NG, Timothy AR, Hain SF, O'Doherty MJ. 18-FDG-PET for the assessment of residual masses on CT following treatment of lymphomas. Ann Oncol 2000;11 Suppl 1:147-50.
27. Mikhaeel NG, Timothy AR, O'Doherty MJ, Hain S, Maisey MN. 18-FDG-PET as a prognostic indicator in the treatment of aggressive non-hodgkin's lymphoma-comparison with CT. Leuk Lymphoma 2000;39: 543-53.
28. Naumann R, Vaic A, Beuthien-Baumann B, Bredow J, Kropp J, Kittner T, et al. Prognostic value of positron emission tomography in the evaluation of post-treatment residual mass in patients with Hodgkin's disease and non-Hodgkin's lymphoma. Br J Haematol 2001;115:793-800.
29. Spaepen K, Stroobants S, Dupont P, Van Steenweghen S, Thomas J, Vandenberghe P, et al. Prognostic value of positron emission tomography (PET) with fluorine-18 fluorodeoxyglucose ([18F]FDG) after first-line chemotherapy in non- Hodgkin's lymphoma: is [18F]FDG-PET a valid alternative to conventional diagnostic methods? J Clin Oncol 2001;19:414-9.
30. Spaepen K, Stroobants S, Dupont P, Thomas J, Vandenberghe P, Balzarini J, et al. Can positron emission tomography with [(18)F]-fluorodeoxyglucose after first-line treatment distinguish Hodgkin's disease patients who need additional therapy from others in whom additional therapy would mean avoidable toxicity? Br J Haematol 2001;115:272-8.
31. Stumpe KD, Urbinelli M, Steinert HC, Glanzmann C, Buck A, von Schulthess GK. Whole-body positron emission tomography using fluorodeoxyglucose for staging of lymphoma: effectiveness and comparison with computed tomography. Eur J Nucl Med 1998; 25:721-8.
32. Weihrauch MR, Re D, Scheidhauer K, Ansen S, Dietlein M, Bischoff S, et al. Thoracic positron emission tomography using 18F-fluorodeoxyglucose for the evaluation of residual mediastinal Hodgkin disease. Blood 2001;98:2930-4.
33. Zinzani PL, Magagnoli M, Chierichetti F, Zompatori M, Garraffa G, Bendandi M, et al. The role of positron emission tomography (PET) in the management of lymphoma patients. Ann Oncol 1999;10:1181-4.
34. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667-76.
35. Cremerius U, Fabry U, Neuerburg J, Zimny M, Osieka R, Buell U. Positron emission tomography with 18F-FDG to detect residual disease after therapy for malignant lymphoma. Nucl Med Commun 1998;19:1055-63.
36. Cremerius U, Fabry U, Kroll U, Zimny M, Neuerburg J, Osieka R, et al. [Clinical value of FDG PET for therapy monitoring of malignant lymphoma results of a retrospective study in 72 patients]. Nuklearmedizin 1999;38:24-30.
37. Bumann D, de Wit M, Beyer W, Beese M, Lubeck M, Bucheler E, et al. [Computerized tomography and F-18-FDG positron emission tomography in staging of malignant lymphomas: a comparison]. Rofo Fortschr Geb Rontgenstr Neuen Bildgeb Verfahr 1998;168: 457-65.
38. de Wit M, Bumann D, Beyer W, Herbst K, Clausen M, Hossfeld DK. Whole-body positron emission tomography (PET) for diagnosis of residual mass in patients with lymphoma. Ann Oncol 1997;8 Suppl 1:57-60.
39. Zinzani PL, Chierichetti F, Zompatori M, Tani M, Stefoni V, Garraffa G, et al. Advantages of positron emission tomography (PET) with respect to computed tomography in the follow-up of lymphoma patients with abdominal presentation. Leuk Lymphoma 2002;43:1239-43.