

---

---

# $^{18}\text{F}$ -FDG PET for Posttherapy Assessment of Hodgkin's Disease and Aggressive Non-Hodgkin's Lymphoma: A Systematic Review

Teruhiko Terasawa<sup>1</sup>, Takashi Nihashi<sup>2</sup>, Tomomitsu Hotta<sup>1</sup>, and Hirokazu Nagai<sup>1</sup>

<sup>1</sup>Clinical Research Center for Blood Diseases, National Hospital Organization Nagoya Medical Center, Nagoya, Japan; and

<sup>2</sup>Department of Radiology, National Center for Geriatrics and Gerontology, Obu City, Japan

---

Although studies have shown that  $^{18}\text{F}$ -FDG PET, when used to assess the response of malignant lymphoma after treatment, has a strong ability to predict relapse, its diagnostic accuracy in clinical practice remains unclear. The aim of this study was to systematically review the diagnostic accuracy of  $^{18}\text{F}$ -FDG PET in detecting residual disease at the completion of first-line therapy of Hodgkin's disease (HD) and aggressive non-Hodgkin's lymphoma (NHL). **Methods:** We searched relevant articles from 1966 to July 2006 using MEDLINE, EMBASE, SCOPUS, Biological Abstracts, bibliographies, review articles, and textbooks without language restriction. One assessor (for non-English-language studies) or 2 assessors (for English-language studies) independently reviewed each article to abstract relevant study characteristics and results. Relevant individual patient data or subgroup data were provided by the investigators if they were unavailable from the publications. We estimated summary receiver operating characteristic curves and confidence regions for summary sensitivity and specificity. **Results:** Nineteen studies consisting of 474 HD and 254 aggressive NHL patients were included. These studies had heterogeneity and suboptimal methodologic quality and reporting. Reported ranges for the sensitivity and specificity of  $^{18}\text{F}$ -FDG PET in predicting disease relapse were 0.50–1.00 and 0.67–1.00, respectively, for HD and 0.33–0.77 and 0.82–1.00, respectively, for NHL. These estimates were similar when conventional imaging tests showed a residual mass. For HD studies, the summary receiver operating characteristic curves were similar irrespective of whether a residual mass was detected by conventional tests. Factors explaining the variability of diagnostic estimates were not identified. **Conclusion:** Although currently available evidence is still limited,  $^{18}\text{F}$ -FDG PET seems to have good diagnostic accuracy for assessing residual HD at the completion of first-line treatment. Clinical data on this use of  $^{18}\text{F}$ -FDG PET for aggressive NHL are more limited. Prospective studies with a more rigorous research design, conduct, and reporting would more reliably reveal the clinical diagnostic accuracy of this imaging modality.

**Key Words:**  $^{18}\text{F}$ -FDG PET; lymphoma; response assessment; residual disease

**J Nucl Med 2008; 49:13–21**

DOI: 10.2967/jnumed.107.039867

---

**P**ersistence of a mass after completion of first-line therapy is a common problem in assessing the response of both Hodgkin's disease (HD) and non-Hodgkin's lymphoma (NHL) (1,2). Based on the likelihood of residual disease, clinicians need to proceed to further therapeutic interventions such as applying radiotherapy to the residual mass or salvage chemotherapy with or without high-dose chemotherapy followed by stem cell transplantation. Therefore, more accurate diagnostic tests to distinguish a posttherapy disease mass from fibrosis or scar tissue are needed to reduce the likelihood of unnecessary and potentially toxic therapy (3).

$^{18}\text{F}$ -FDG PET is a promising functional imaging test for patients with malignant lymphoma and other malignancies and has gained wide use during the last decade (4). Its routine use has been recommended to assess the posttherapy response of HD, especially if CT reveals a residual mass (3). Others have also recommended its use for the same purpose on the basis of its high negative predictive value for HD and high positive predictive value for NHL (5). A recent survey of physicians caring for lymphoma patients found that intended management plans were often changed on the basis of  $^{18}\text{F}$ -FDG PET results (6).

Recently, a systematic review explored the diagnostic accuracy of  $^{18}\text{F}$ -FDG PET for this purpose and assessed the quality of the included studies (7). Similar to studies of diagnostic tests in other medical fields, this review revealed several methodologic problems affecting both the internal and the external validity of the published studies. The authors, however, estimated summary diagnostic accuracy without considering the effect of major methodologic variability on diagnostic tests underlying the original studies. For example, some of the studies included a mix of NHL histologies consisting of indolent, aggressive, and highly

---

Received Jan. 15, 2007; revision accepted Aug. 18, 2007.

For correspondence or reprints contact: Teruhiko Terasawa, MD, Institute for Clinical Research and Health Policy Studies, Tufts-New England Medical Center, 750 Washington St., Tufts-NEMC #63, Boston, MA 02111.

E-mail: tterasawa@tufts-nemc.org

COPYRIGHT © 2008 by the Society of Nuclear Medicine, Inc.

aggressive lymphomas, and others included a mix of patients consisting of those who received first-line therapy and those who received salvage therapy. The diagnostic accuracy estimates reported from these studies have limited external validity and cannot be directly applied to specific clinical scenarios; each distinct histologic subtype has its unique clinical profiles, such as treatment strategies, responses, and prognoses. Also, the review did not consider the different ways that the primary studies presented results (e.g., individual patient vs. each lymphoma lesion, all involved sites vs. only bulky disease, or single vs. multiple inclusions of a patient).

This study was an updated systematic review of the diagnostic accuracy of  $^{18}\text{F}$ -FDG PET in assessing the response of HD and aggressive NHL after first-line therapy, with special emphasis on the methodologic issues discussed above.

## MATERIALS AND METHODS

### Study Identification

We searched MEDLINE and EMBASE from January 1966 through July 2006 without language restriction. The detailed search strategy can be found in Supplemental Table 1 (supplemental materials are available online only at <http://jnm.snmjournals.org>). This search was augmented by SCOPUS and Biological Abstracts. We also manually reviewed the reference lists of eligible studies, review articles, and textbooks.

### Study Selection

Two of us reviewed the pertinent studies to determine eligibility. We included prospective or retrospective studies evaluating posttherapy response assessment by  $^{18}\text{F}$ -FDG PET exclusively for patients with HD or aggressive NHL followed by clinical follow-up, with or without pathologic confirmation as a reference standard. We included studies that evaluated at least 10 patients. We considered aggressive NHL to be mantle cell lymphoma; follicular center lymphoma, grade III; diffuse large B-cell lymphoma (DLBCL); primary mediastinal large B-cell lymphoma; peripheral T-cell lymphoma, unspecified; angioimmunoblastic T-cell lymphoma; angiocentric lymphoma; and anaplastic large cell lymphoma, T- and null-cell type by the Revised European–American Classification of Lymphoid Neoplasms (REAL)/World Health Organization classification, or corresponding subcategories for the International Working Formulation classification, the Kiel classification, or the Rappaport classification (8). We included studies of only those patients who completed first-line chemotherapy, radiotherapy, or combined-modality therapy and underwent conventional imaging tests such as CT, ultrasonography, or MRI for posttreatment restaging just before or after undergoing PET. We focused our analysis on studies that reported individual patients as the unit of analysis irrespective of the number of relapses or of the sites of relapse or residual disease, because this is the most appropriate perspective for clinical decision making. We excluded abstracts, editorials, comments, letters, review articles, and case reports. We excluded studies enrolling patients with HIV-associated or post-transplant lymphoproliferative disorders.

Many studies did not meet all the rigorous inclusion criteria but did partially include a relevant patient population. For these studies, we contacted the authors by mail or email to ask for individual patient data or subgroup data relevant to our inclusion criteria. When there was no response within 3 wk, another correspondence was

sent. When there was no response to the third communication attempt, we considered the request rejected.

### Data Abstraction

Two independent, board-certified hematologists abstracted relevant data for English-language articles. For non-English-language articles, data were extracted by a single reviewer working with a physician native speaker of the relevant language. We used an abstraction form consisting of items recommended in the Standards for Reporting of Diagnostic Accuracy (9). One nuclear medicine specialist evaluated the technical specifications and quality of the PET procedure by using recommended guidelines (10). The reviewers knew in which journals the studies had been published. Based on the enrollment of participants before PET, we categorized studies into 2 groups: “posttherapy evaluation,” or studies that included patients irrespective of restaging results on conventional imaging tests, and “residual mass evaluation,” or studies that evaluated only patients for whom visible residual mass lesions were shown on conventional imaging tests. In the posttherapy evaluation, we also abstracted the data on the subgroup of patients who had a residual mass shown on imaging. If the relevant data were unavailable from the published literature, we contacted the authors of the paper to request the subgroup data. Inconsistencies between reviewers were either clarified by the paper authors or resolved by consensus.

### Assessment of Study Quality and Applicability

To evaluate the quality and applicability of the studies included in this review, we used an established quality rating system for diagnostic studies (11) and a recently proposed quality evaluation tool (12). In the established system, we examined 6 aspects of study quality: quality and application of the reference standard, independence of test interpretation, description of patient characteristics, cohort assembly, and sample size. We then rated each study as “a” (the highest quality), “b,” “c,” or “d” (the lowest quality) according to the predefined score. The recently proposed quality evaluation tool, which was designed exclusively for studies of diagnostic accuracy, comprehensively explores both methodologic quality and reporting. The tool consists of 14 items addressing patient spectrum, reference standard, disease progression bias, partial and differential verification bias, test review bias, clinical review bias, incorporation bias, test execution of both index test and reference standard, study withdrawals, and indeterminate test results.

### Data Synthesis and Statistical Analysis

For each study, we constructed a  $2 \times 2$  contingency table consisting of true-positive, false-positive, false-negative, and true-negative, where all patients were categorized as being PET-positive or -negative and as being positive or negative for disease according to the results determined through  $^{18}\text{F}$ -FDG PET and the reference standard, respectively. We defined as disease-positive a patient who had biopsy-confirmed residual disease or whose disease had relapsed during clinical follow-up. We did not independently combine the sensitivity and specificity of the included studies because this approach does not take into account the interdependence of these 2 test parameters. We instead estimated summary receiver operating characteristic (ROC) curves and elliptic 95% confidence regions of summary sensitivity and specificity by the hierarchical summary ROC method (13). This model is a more sophisticated approach than the conventional linear regression model to compute summary diagnostic measures taking account of variations both within a study and between studies. We fitted the model by using maximum-

likelihood estimation implemented in the NLMIXED procedure of SAS/STAT (version 8; SAS Institute) (14). Then, we depicted the summary ROC curves and confidence regions for summary sensitivity and specificity by using Stata (version 8.2.; Stata Corp.) (15). Also, we estimated the area under the curve and the  $Q^*$  statistic, the point on the curve where sensitivity equals specificity, as global measures for the summary ROC curves. We explored heterogeneity between studies by visual assessment of ROC plots for the following predetermined items: study design (prospective vs. retrospective), rates of patients with residual mass as found on conventional imaging tests, relapse rates, follow-up period, timing of PET scan after completion of therapy, and publication year. We also performed post hoc subgroup analyses for the items used for the technical specifications of PET and quality assessment.

## RESULTS

### Study Characteristics

We included 19 studies: 9 studies (16–24) that met all eligibility criteria and 10 studies (25–34) with relevant published data or with unpublished data obtained by contacting the authors of the paper (Table 1). The detailed process by which articles were selected can be found in the Supplemental Appendix and in Supplemental Figure 1. Because one group of researchers reported a “residual therapy evaluation” for HD and aggressive NHL and a “posttherapy evaluation” exclusively for aggressive NHL—and included in both groups the data on some participants (23,29)—we separately extracted the relevant data. Another group of researchers published an updated “posttherapy evaluation” for HD, with some participants overlapping from the previous report (27,31). We contacted the principal investigator and obtained updated combined results. For 5 studies that reported both HD and aggressive NHL (25–28,30), the contacted investigators provided unpublished subgroup data for each category. As a result, we finally had for further evaluation 15 studies consisting of 474 patients with HD and 8 studies consisting of 254 patients with aggressive NHL (Table 2).

Most studies were retrospective (Table 1). Four studies exclusively evaluated patients for whom a posttherapy visual residual mass was found on conventional imaging tests (25,29,32,34). Generally, patients underwent PET 1–3 mo after completing therapy, but some underwent PET at less than 1 mo. Most studies included both adult and adolescent patients, but 3 studies also included pediatric patients (16,21,24).

For HD studies, nodular sclerosis was the leading histologic subcategory and either ABVD (doxorubicin, bleomycin, vinblastine, dacarbazine) or MOPP (nitrogen mustard, vincristine, procarbazine, prednisone) plus ABVD or ABV (doxorubicin, bleomycin, vincristine) with or without radiotherapy was the most widely reported first-line treatment (Table 2). In posttherapy evaluations, 35%–72% of patients were found to have a residual mass on conventional imaging (Supplemental Table 2). Relapse rates were similar for both posttherapy evaluations and residual mass evaluations, ranging from 4% to 55% and 0% to 50%, respectively.

For aggressive NHL studies, DLBCL was the leading histologic subcategory, and the most widely adopted first-line therapy was CHOP (cyclophosphamide, doxorubicin, vincristine, prednisone) or comparable doxorubicin-containing regimens with or without radiotherapy (Table 2). In posttherapy evaluations, 26%–56% of patients were found to have a residual mass on conventional imaging (Supplemental Table 2). Relapse rates were 33%–60% for posttherapy evaluations and 33%–67% for residual mass evaluations.

Concerning imaging techniques and technologies, although their reporting was limited, the included studies generally followed the guidelines of the Society of Nuclear Medicine for performing  $^{18}\text{F}$ -FDG PET (Supplemental Table 3). Only a single study used PET/CT (30). Most studies adopted qualitative diagnostic criteria: foci of elevated  $^{18}\text{F}$ -FDG uptake unexplained by physiologic uptake. Two studies (32,34) also adopted quantitative diagnostic criteria: standardized uptake values (the ratio of  $^{18}\text{F}$ -FDG uptake in tumor sites to that in normal sites). Only 2 studies clearly reported that all the included participants underwent pretherapy PET (17,23). Generally, experienced nuclear medicine physicians interpreted the results.

### Sensitivity, Specificity, and Summary ROC Curves

HD studies reported widely ranging sensitivities and specificities for  $^{18}\text{F}$ -FDG PET. For posttherapy evaluations, sensitivity ranged from 0.50 to 1.00 and specificity ranged from 0.67 to 1.00 (Supplemental Table 2; Fig. 1). For residual mass evaluations, reported estimates had a similarly wide range: 0.43–1.00 for sensitivity and 0.67–1.00 for specificity. The summary ROC curves and confidence regions for summary sensitivity and specificity for posttherapy evaluations and residual mass evaluations were similar (Fig. 2): The area under the curve for the summary ROC curve was 0.94 for posttherapy evaluations and 0.93 for residual mass evaluations, and the  $Q^*$  statistic was 0.88 for posttherapy evaluations and 0.86 for residual mass evaluations.

Studies of aggressive NHL also reported a widely ranging sensitivity but a narrower high specificity. For posttherapy evaluations, sensitivity ranged from 0.33 to 0.77 and specificity ranged from 0.82 to 1.00 (Supplemental Table 2; Fig. 1). For residual mass evaluations, sensitivity ranged from 0.33 to 0.87 and specificity ranged from 0.75 to 1.00. We did not perform the metaanalysis because there were too few data points to reliably estimate the summary ROC curves and confidence regions for summary sensitivity and specificity.

### Investigating Heterogeneity

We did not identify any clinical or  $^{18}\text{F}$ -FDG PET test characteristics, or any items that assessed the quality and applicability of each study, to explain the heterogeneity of sensitivity and specificity (data not shown).

### Quality Assessment of Published Studies

Overall, the quality and reporting of the included studies were limited (Supplemental Table 4), suggesting that they are subject to bias and variation limiting the internal and

**TABLE 1**  
Studies Included in the Systematic Review

Study	Clinical indication	Study design	Time from therapy completion to PET (mo)			Follow-up (mo)			No. of patients	No. of female patients	Age (y)		Exclusion of patients with PIF or PD during treatment
			Median	Range	Range	Median	Range	Median			Range		
<b>HD and NHL</b>													
de Wit et al. (25)	Residual mass	Retrospective	2	<10	13	1-26	34	NR	NR	43	21-73	NR	
Foo et al. (26)	Posttherapy	Retrospective	NR	NR	NR or 13*	3-65	24 <sup>†</sup>	33% <sup>‡</sup>	33% <sup>‡</sup>	33 <sup>‡</sup>	20-67	NR	
Jerusalem et al. (27)	Posttherapy	Prospective	NR	1-3	23 or 21*	NR	54	27 (50%)	27 (50%)	47	15-80	Yes	
Mikosch et al. (28)	Posttherapy <sup>§  </sup>	Retrospective	NR	>1	NR	>12	93	NR	NR	NR	NR	NR	
Mikhaeel et al. (29)	Residual mass	Retrospective	NR	<1	38	16-68	32	31 (31%)	31 (31%)	NR	NR	NR	
Schaefer et al. (30)	Posttherapy <sup>¶</sup>	Retrospective	3	NR	9.3 <sup>#</sup>	NR	41 <sup>†</sup>	38% <sup>‡</sup>	38% <sup>‡</sup>	39.6 <sup>#</sup>	NR	NR	
<b>HD</b>													
Filmont et al. (16)	Posttherapy	Retrospective	5.2	1-87	14	6-39	32	22 (53%)	22 (53%)	30	6-65	NR	
Friedberg et al. (17)	Posttherapy	Prospective	NR	<1	24	10-32	32 <sup>†</sup>	36% <sup>‡</sup>	36% <sup>‡</sup>	30 <sup>‡</sup>	18-60	NR	
Guay et al. (18)	Posttherapy	Retrospective	2	0-55	15	1-60	48	25 (50%)	25 (50%)	38	17-78	NR	
Jerusalem et al. (31)	Posttherapy <sup>§</sup>	Prospective	1	NR	NR <sup>§</sup>	NR	36	23 (64%) <sup>‡</sup>	23 (64%) <sup>‡</sup>	28	13-71	Yes	
Keresztes et al. (32)	Residual mass <sup>  </sup>	Retrospective	3 <sup>#</sup>	2-5	50 or 62**	5-98	31 <sup>†</sup>	42% <sup>‡</sup>	42% <sup>‡</sup>	31 or 37**	13-70	NR	
Mocikova et al. (33)	Posttherapy <sup>  </sup>	Retrospective	NR	NR	25	NR	94	44 (47%) <sup>‡</sup>	44 (47%) <sup>‡</sup>	30	28-85	NR	
Rigacci et al. (19)	Posttherapy	Prospective	4	2-8	45	NR	28	14 (50%)	14 (50%)	30.6	16-73	NR	
Spaepen et al. (20)	Posttherapy	Retrospective	NR	1-3	30	12-48	60	33 (45%)	33 (45%)	35	11-75	NR	
Weihrauch et al. (34)	Residual mass <sup>§</sup>	Prospective	NR	<4	28	16-68	28	8 (26%)	8 (26%)	36	18-76	NR	
Wickmann et al. (21)	Posttherapy	Retrospective	NR	0-7	32	1-56	48	36	36	NR	<18	NR	
<b>NHL</b>													
Juweid et al. (22)	Posttherapy	Retrospective	1	0-5	35	25-60	54	NR	NR	58	21-79	No	
Mikhaeel et al. (23)	Posttherapy	Retrospective	NR	<2	30	7-56	45 <sup>†</sup>	37% <sup>‡</sup>	37% <sup>‡</sup>	NR	NR	Yes	
Spaepen et al. (24)	Posttherapy	Retrospective	NR	1-3	21	11-50	93	34 (37%)	34 (37%)	47	2-77	NR	

\*Patients without or with residual mass, respectively.

<sup>†</sup>Only subset of participants who were evaluated with PET after completion of therapy.

<sup>‡</sup>Data abstracted from total participants, not limited to patients who were evaluated with PET at end of therapy.

<sup>§</sup>Some PET scans were taken at the end of salvage therapy.

<sup>||</sup>Some participants underwent PET multiple times.

<sup>¶</sup>Each lymphoma lesion was evaluated as the unit of analysis.

<sup>#</sup>Mean.

\*\*Patients with or without bulky mass before therapy, respectively.

NR = not reported; PD = progressive disease; PIF = primary induction failure.

**TABLE 2**  
Patient Characteristics for Studies Included in the Systematic Review

Study	No. of patients*	Histologic class	Pretherapy stage†	Type of therapy
HD				
de Wit et al. (25)	5	NA	NA	Chemotherapy ± radiotherapy
Filmont et al. (16)	32	22 NS, 7 MC, 2 LP, 1 LD	NA	Chemotherapy ± radiotherapy ± surgery or autologous transplantation
Foo et al. (26)	9	7 NS, 1 MC, 1 unclassifiable	8 II, 1 III	Chemotherapy ± radiotherapy
Friedberg et al. (17)	32	20 NS, 6 MC, 3 LP, 7 unclassifiable	3 I, 23 II, 6 III, 4 IV	ABVD or MOPP + ABVD ± radiotherapy
Guay et al. (18)	48	NA	2 I, 23 II, 15 III, 8 IV	MOPP or ABVD ± radiotherapy
Jerusalem et al. (27,31)	31	27 NS, 3 MC, 1 LD	NA	Chemotherapy ± radiotherapy
Keresztes et al. (32)	26	NA	NA	C-MOPP + ABVD or C-MOPP + ABV + CEP ± radiotherapy
Mikosch et al. (28)	31	NA	NA	Chemotherapy ± radiotherapy
Mikhaeel et al. (29)	15	NA	8 I + II, 7 III + IV	ABVD ± radiotherapy
Mocikova et al. (33)	71	NA	NA	Chemotherapy ± radiotherapy
Rigacci et al. (19)	28	22 NS, 5 MC, 1 LD	21 II, 5 III, 2 IV	ABVD ± radiotherapy
Schaefer et al. (30)	18	17 NS, 1 MC	NA	NA
Spaepen et al. (24)	60	45 NS, 9 MC, 3 LP, 3 unclassifiable	25 II, 19 III, 16 IV	Stanford V or MOPP + ABV ± radiotherapy
Weirauch et al. (34)	20	NA	1 I, 9 II, 7 III, 2 IV	ABVD, C-MOPP + ABVD, or BEACOPP ± radiotherapy‡
Wickmann et al. (21)	48	NA	NA	GPOH-HD 95§
Aggressive NHL				
de Wit et al. (25)	6	Aggressive by REAL/WHO criteria	NA	Chemotherapy ± radiotherapy
Foo et al. (26)	12	11 diffuse large, 1 PTCL	3 I, 5 II, 2 III, 2 IV	Chemotherapy ± radiotherapy
Jerusalem et al. (31)	35	27 DLBCL, 8 FL grade 3	NA	Chemotherapy ± radiotherapy
Juweid et al. (22)	54	47 diffuse large cell, 4 ALCL, 2 diffuse mixed, 1 angioimmunoblastic T-cell	19 I + II, 35 III + IV	22 CHOP, 29 R-CHOP, 2 ProMACE-CytaBOM, 1 CNOP
Mikosch et al. (28)	24	Aggressive by REAL/WHO criteria	NA	Chemotherapy ± radiotherapy
Mikhaeel et al. (23)	45	“Aggressive” (diffuse mixed, diffuse large cell, large cell immunoblastic)	30 I + II, 19 III + IV <sup>  </sup>	Radiotherapy only or CHOP ± radiotherapy
Schaefer et al. (30)	5	5 DLBCL	NA	NA
Spaepen et al. (24)	73	50 DLBCL, 8 ALCL, 7 ALCL HD-like, 5 MCL, 3 PTCL	NA	Doxorubicin-containing regimen or UKCCSG protocol (pediatric cases)

\*Who met inclusion criteria of this systematic review.

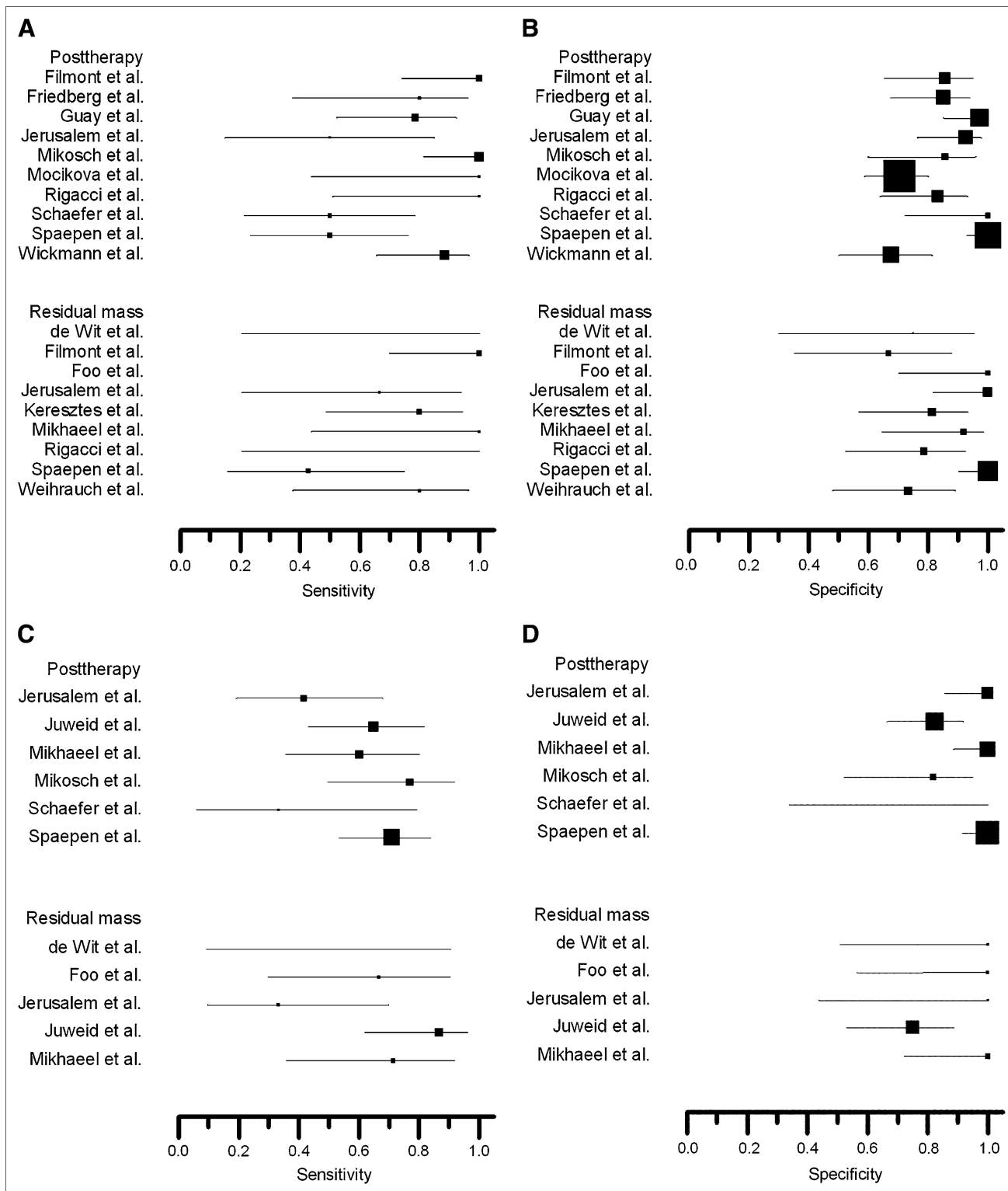
†According to Ann Arbor staging system.

‡Eight relapsed cases (excluded) received salvage therapy with or without high-dose therapy followed by stem-cell transplantation.

§Treatment strategy for HD in children and adolescents (35).

<sup>||</sup>Data abstracted from total participants, not limited to patients who were evaluated with PET at end of therapy.

ALCL = anaplastic large cell lymphoma; BCL = B-cell lymphoma; BEACOPP = bleomycin, etoposide, doxorubicin, cyclophosphamide, vincristine, procarbazine, prednisone; CEP = CCNU, etoposide, prednimustine; C-MOPP = cyclophosphamide, vincristine, procarbazine, prednisone; CNOP = cyclophosphamide, mitoxantrone, vincristine, procarbazine, prednisone; FL = follicular lymphoma; LD = lymphocyte depleted; LP = lymphocyte predominance; MC = mixed cellularity; MCL = mantle cell lymphoma; NA = not available; NS = nodular scleriosis; ProMACE-CytaBOM = prednisone, cyclophosphamide, doxorubicin, etoposide, cytarabine, bleomycin, vincristine, methotrexate; PTCL = peripheral T-cell lymphoma; R-CHOP = rituximab plus CHOP; Stanford V = doxorubicin, vinblastine, mechlorethamine, vincristine, bleomycin, etoposide, prednisone; UKCCSG = United Kingdom children cancer study group; WHO = World Health Organization.

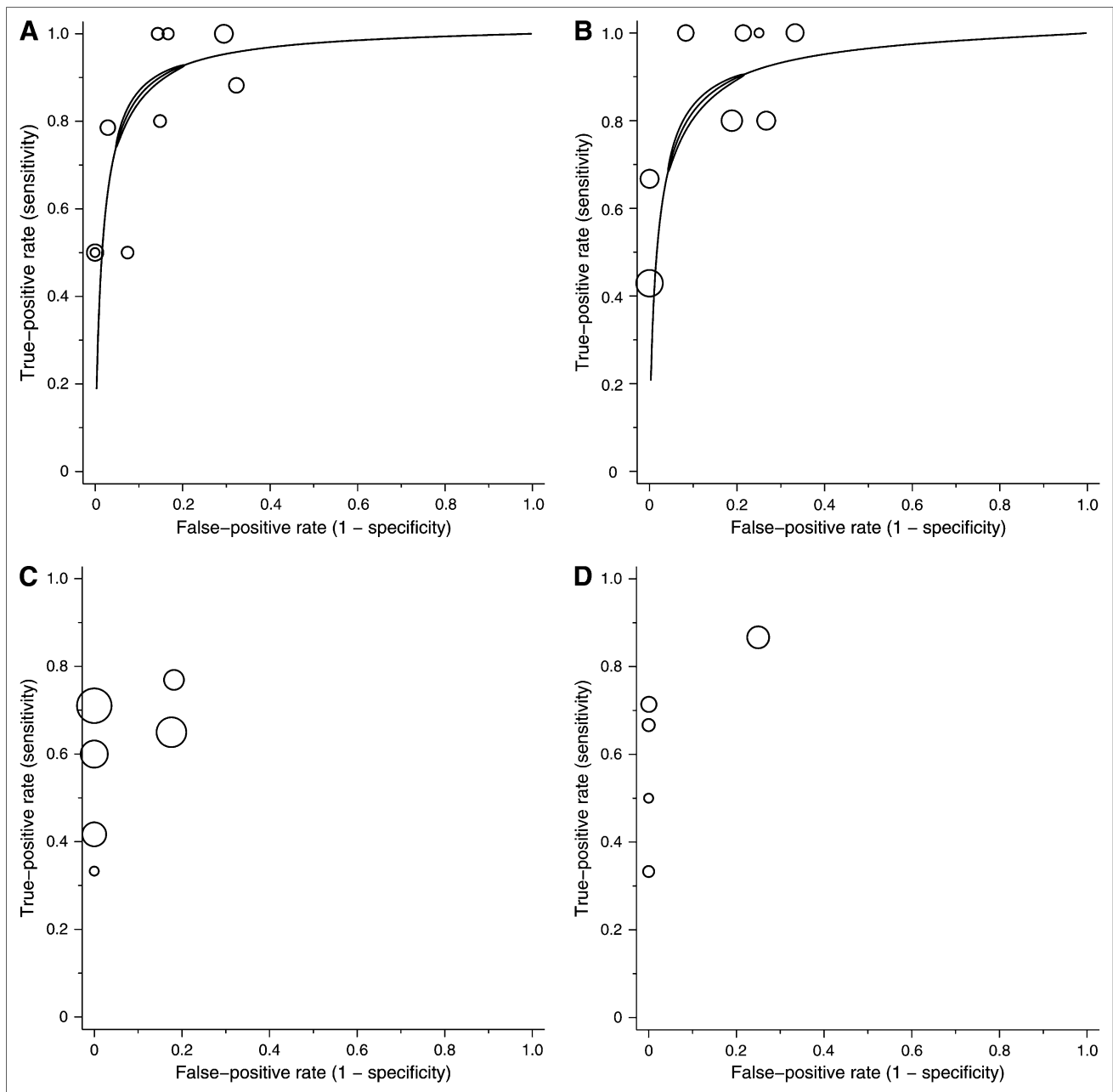


**FIGURE 1.** Sensitivity and specificity for HD (A and B) and aggressive NHL (C and D). Size of square plotting symbol is proportional to sample size (for sensitivity, number of patients who relapsed; for specificity, number of patients in remission) for each study. Horizontal lines are 95% confidence intervals. One study (Foo et al.) is omitted for sensitivity for HD because there were no cases of relapse.

external validities, respectively, of the test results. The detailed results of the quality assessment can be found in the Supplemental Appendix.

## DISCUSSION

This systematic review showed that the reported sensitivities and specificities of <sup>18</sup>F-FDG PET to assess residual



**FIGURE 2.** ROC plotting and summary ROC curve of  $^{18}\text{F}$ -FDG PET for HD and aggressive NHL. Individual study estimates of sensitivity and  $1 - \text{specificity}$  are shown for posttherapy evaluation and residual mass evaluation of HD (A and B) and aggressive NHL (C and D). Size of each circle is proportional to sample size for each study (all study participants). Solid crescent boundary represents 95% confidence region for summary sensitivity and specificity.

disease for patients with HD who completed first-line chemo- or chemoradiotherapy had large between-study heterogeneity. In addition, the summary ROC curves and confidence regions for sensitivity and specificity showed similar and good overall diagnostic accuracy irrespective of the presence of a visible residual mass on conventional imaging modalities such as CT or MRI. For aggressive NHL, although our review showed a relatively stable high specificity and variable sensitivity, there were too few studies to calculate a reliable overall diagnostic accuracy.

Many potential factors may explain the heterogeneity. For test characteristics, differences in the type of PET scanner, in the timing of PET after the completion of therapy, in positive test criteria, and in the clinical experience of the interpreters are relevant. For patient characteristics, several differences should be considered: in the type of histology, especially for aggressive NHL (e.g., DLBCL vs. other aggressive NHLs); in therapeutic strategy (e.g., chemotherapy vs. combined-modality therapy); in the presence or absence of a visible residual mass on conventional imaging tests; and in relapse

risk groups, such as the international prognostic score for advanced-stage HD (36) or the international prognostic index for aggressive NHL (8). For study characteristics, differences such as method of patient selection (e.g., prospective vs. retrospective), type and application of a reference standard, and patient follow-up have been reported to affect the variability of sensitivity and specificity (9,37). Because the available data were limited, we could not identify specific factors that explain the heterogeneity, which should be further addressed in future studies. Although the reported positive PET criteria appeared almost identical in all HD studies, the large crescent shape of the confidence regions implies a negative correlation between sensitivity and specificity across the included studies, suggesting that a variation in threshold may partially explain the between-study heterogeneity.

Our study had several important limitations. Because we selected only those studies for which pertinent data were available, several important investigations (some of which were included in the previous metaanalysis) may have been excluded. Also, because most data were derived from retrospective studies with poor-quality design and reporting, our conclusions are subject to the bias and variations in the original studies (37). In addition, our review included only a single study in which patients with DLBCL received rituximab in addition to CHOP (22). Because the combination of rituximab with CHOP is a current standard therapy (38), our results may be less applicable to clinical practice. Further, we included only a single study (30) that used a relatively new and promising additional technique—PET/CT—that may overcome the current technical limitations of PET (39).

In the recently revised consensus recommendations,  $^{18}\text{F}$ -FDG PET has become an important component of posttherapy response assessments in clinical trials for HD and DLBCL (40). Although the currently available data have limitations, our systematic review would probably support the clinical relevance of the response criteria for high-risk HD; incomplete response defined by positive PET findings would have an excellent ability to predict relapse irrespective of whether a residual tumor mass is found on conventional imaging. For favorable-risk HD, patients labeled as incomplete responders after first-line therapy should still have a moderate possibility of long-term remission. Thus, clinical investigators adopting the recommendations into therapeutic efficacy trials would need to decide how to manage patients in this category before implementing the trials. For DLBCL and other aggressive NHLs, our results based on the limited clinical evidence would not suffice to support the criteria, and further research is necessary to validate them.

Our review shows that currently available data do not suffice to answer the phase 3 question of diagnostic accuracy studies: What is the diagnostic accuracy of  $^{18}\text{F}$ -FDG PET for posttherapy response assessment of malignant lymphoma? (41). Reliable clinical evidence is especially limited for aggressive NHL. Further investigation should include prospective diagnostic accuracy studies (phase 3) of PET or

PET/CT that adopt a more rigorous research methodology. We propose that, ideally, diagnostic accuracy studies should accompany prospective clinical trials to answer efficacy questions. Data on additional therapy, such as involved-field radiotherapy or high-dose chemotherapy with autotransplantation, for posttherapy PET-positive patients are limited. Before the routine clinical implementation of a treatment strategy based on posttherapy PET findings, randomized studies should assess the impact on patients' clinical outcomes, if appropriate (phase 4) (41). Also, determination of the cost-effectiveness of treatment strategies adopting PET is necessary to allow a better understanding of the role of posttherapy PET.

## CONCLUSION

The currently available data show that  $^{18}\text{F}$ -FDG PET has good overall accuracy in detecting residual disease in patients with HD who have completed first-line therapy. The current literature has methodologic weaknesses that may overestimate accuracy. Because data from original studies are limited, our review could not find robust evidence to answer the question of whether clinicians should routinely use PET to assess the posttherapeutic response, suggesting that they should be cautious about making clinical decisions based solely on a PET result.

## ACKNOWLEDGMENTS

We thank Maike de Wit, MD, Jerusalem Guy, MD, Malik Juweid, MD, Katalin Keresztes, MD, Peter Mikosch, MD, Heidy Mocikov, MD, Robert Pichler, MD, Niklaus Schaefer, MD, and Martin Weihrauch, MD, for providing the data on their original work; Yinghua Li, MD, and Leo Kawaguchi, MD, for assistance in translating Chinese- and German-language publications; Jonathan Deeks, PhD, Kengo Ito, MD, and Kazuo Kubota, MD, for helpful comments; Petra Macaskill, PhD, for providing statistical programs; and Bruce Cheson, MD, and Joseph Lau, MD, for reviewing the manuscript. Part of this work was supported by Grant-in-Aid 15-2 for Cancer Research from the Ministry of Health, Labor, and Welfare of Japan, by the Japanese Society of Nuclear Medicine, and by Nihon Medi-Physics Corporation Limited.

## REFERENCES

1. Radford JA, Cowan RA, Flanagan M, et al. The significance of residual mediastinal abnormality on the chest radiograph following treatment for Hodgkin's disease. *J Clin Oncol*. 1988;6:940-946.
2. Cheson BD, Horning SJ, Coiffier B, et al. Report of an international workshop to standardize response criteria for non-Hodgkin's lymphomas. NCI Sponsored International Working Group. *J Clin Oncol*. 1999;17:1244-1253.
3. Meyer RM, Ambinder RF, Stroobants S. Hodgkin's lymphoma: evolving concepts with implications for practice. *Hematology Am Soc Hematol Educ Program*. 2004;184-202.
4. Gambhir SS, Czernin J, Schwimmer J, Silverman DH, Coleman RE, Phelps ME. A tabulated summary of the FDG PET literature. *J Nucl Med*. 2001;42(suppl):1S-93S.
5. Burton C, Ell P, Linch D. The role of PET imaging in lymphoma. *Br J Haematol*. 2004;126:772-784.



6. Hillner BE, Tunuguntla R, Fratkin M. Clinical decisions associated with positron emission tomography in a prospective cohort of patients with suspected or known cancer at one United States center. *J Clin Oncol.* 2004;22:4147–4156.
7. Zijlstra JM, Lindauer-van der Werf G, Hoekstra OS, Hooft L, Riphagen II, Huijgens PC. <sup>18</sup>F-Fluoro-deoxyglucose positron emission tomography for post-treatment evaluation of malignant lymphoma: a systematic review. *Haematologica.* 2006;91:522–529.
8. A predictive model for aggressive non-Hodgkin's lymphoma. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. *N Engl J Med.* 1993;329:987–994.
9. Bossuyt PM, Reitsma JB, Bruns DE, et al. Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med.* 2003;138:40–44.
10. Schelbert RS, Hoh CK, Royal HD, et al. Society of Nuclear Medicine procedure guideline for tumor imaging using F-18 FDG, version 2.0 [SNM Web site]. 1999. Available at: <http://interactive.snm.org/index.cfm?PageID=772&RPID=969>. Accessed October 18, 2007.
11. Kent DL, Haynor DR, Larson EB, Deyo RA. Diagnosis of lumbar spinal stenosis in adults: a metaanalysis of the accuracy of CT, MR, and myelography. *AJR.* 1992;158:1135–1144.
12. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol.* 2003;3:25.
13. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med.* 2001;20:2865–2884.
14. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol.* 2004;57:925–932.
15. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics.* 2007;8:239–251.
16. Filmont JE, Yap CS, Ko F, et al. Conventional imaging and 2-deoxy-2-[<sup>18</sup>F]fluoro-D-glucose positron emission tomography for predicting the clinical outcome of patients with previously treated Hodgkin's disease. *Mol Imaging Biol.* 2004;6:47–54.
17. Friedberg JW, Fischman A, Neuberger D, et al. FDG-PET is superior to gallium scintigraphy in staging and more sensitive in the follow-up of patients with de novo Hodgkin lymphoma: a blinded comparison. *Leuk Lymphoma.* 2004;45:85–92.
18. Guay C, Lepine M, Verreault J, Benard F. Prognostic value of PET using <sup>18</sup>F-FDG in Hodgkin's disease for posttreatment evaluation. *J Nucl Med.* 2003;44:1225–1231.
19. Rigacci L, Castagnoli A, Dini C, et al. <sup>18</sup>F-FDG-positron emission tomography in post treatment evaluation of residual mass in Hodgkin's lymphoma: long-term results. *Oncol Rep.* 2005;14:1209–1214.
20. Spaepen K, Stroobants S, Dupont P, 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;115:272–278.
21. Wickmann L, Luders H, Dorffel W. <sup>18</sup>F-FDG-PET-findings in children and adolescents with Hodgkin's disease: retrospective evaluation of the correlation to other imaging procedures in initial staging and to the predictive value of follow up examinations [German]. *Klin Padiatr.* 2003;215:146–150.
22. Juweid ME, Wiseman GA, Vose JM, et al. Response assessment of aggressive non-Hodgkin's lymphoma by integrated International Workshop Criteria and fluorine-18-fluorodeoxyglucose positron emission tomography. *J Clin Oncol.* 2005;23:4652–4661.
23. Mikhaeel NG, Timothy AR, O'Doherty MJ, Hain S, Maisey MN. <sup>18</sup>F-FDG-PET as a prognostic indicator in the treatment of aggressive non-Hodgkin's lymphoma: comparison with CT. *Leuk Lymphoma.* 2000;39:543–553.
24. Spaepen K, Stroobants S, Dupont P, et al. Prognostic value of positron emission tomography (PET) with fluorine-18 fluorodeoxyglucose (<sup>18</sup>F]FDG) after first-line chemotherapy in non-Hodgkin's lymphoma: is [<sup>18</sup>F]FDG-PET a valid alternative to conventional diagnostic methods? *J Clin Oncol.* 2001;19:414–419.
25. 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):57–60.
26. Foo SS, Mitchell PL, Berlangieri SU, Smith CL, Scott AM. Positron emission tomography scanning in the assessment of patients with lymphoma. *Intern Med J.* 2004;34:388–397.
27. Jerusalem G, Beguin Y, Fassotte MF, et al. Whole-body positron emission tomography using <sup>18</sup>F-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;94:429–433.
28. Mikosch P, Gallowitsch HJ, Zinke-Cerwenka W, et al. Accuracy of whole-body <sup>18</sup>F-FDP-PET for restaging malignant lymphoma. *Acta Med Austriaca.* 2003;30:41–47.
29. Mikhaeel NG, Timothy AR, Hain SF, O'Doherty MJ. <sup>18</sup>F-FDG-PET for the assessment of residual masses on CT following treatment of lymphomas. *Ann Oncol.* 2000;11(suppl):147–150.
30. Schaefer NG, Hany TF, Taverna C, et al. Non-Hodgkin lymphoma and Hodgkin disease: coregistered FDG PET and CT at staging and restaging—do we need contrast-enhanced CT? *Radiology.* 2004;232:823–829.
31. 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;14:123–130.
32. Keresztes K, Lengyel Z, Devenyi K, Vadasz G, Miltenyi Z, Illes A. Mediastinal bulky tumour in Hodgkin's disease and prognostic value of positron emission tomography in the evaluation of post-treatment residual masses. *Acta Haematol.* 2004;112:194–199.
33. Mocikova H, Markova J, Belohlavek O, et al. *Cas Lek Cesk.* 2004;143:476–479.
34. Weihrauch MR, Re D, Scheidhauer K, et al. Thoracic positron emission tomography using <sup>18</sup>F-fluorodeoxyglucose for the evaluation of residual mediastinal Hodgkin disease. *Blood.* 2001;98:2930–2934.
35. Dorffel W, Luders H, Ruhl U, et al. Preliminary results of the multicenter trial GPOH-HD 95 for the treatment of Hodgkin's disease in children and adolescents: analysis and outlook. *Klin Padiatr.* 2003;215:139–145.
36. Hasenclever D, Diehl V. A prognostic score for advanced Hodgkin's disease. International Prognostic Factors Project on Advanced Hodgkin's Disease. *N Engl J Med.* 1998;339:1506–1514.
37. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med.* 2004;140:189–202.
38. Cheson BD. What is new in lymphoma? *CA Cancer J Clin.* 2004;54:260–272.
39. Barrington SF, O'Doherty MJ. Limitations of PET for imaging lymphoma. *Eur J Nucl Med Mol Imaging.* 2003;30(suppl):S117–S127.
40. Cheson BD, Pfistner B, Juweid ME, et al. The International Harmonization Project on Lymphoma: revised response criteria for malignant lymphoma. *J Clin Oncol.* 2007;25:579–586.
41. Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ.* 2002;324:539–541.