Journal of Nuclear Medicine, published on September 16, 2009 as doi:10.2967/jnumed.109.064121

Observer Variation in Interpreting ^{18}F-FDG PET/CT Findings for Lymphoma Staging

Michael S. Hofman[1,2], Nigel C. Smeeton[3], Sheila C. Rankin[1,2], Tom Nunan[1,2], and Michael J. O'Doherty[1,2]

[1]PET Imaging Centre, Kings College London, London, United Kingdom; [2]Division of Imaging, Guy's and St. Thomas' NHS Foundation Trust, London, United Kingdom; and [3]Division of Health and Social Care Research, King's College London, London, United Kingdom

Many studies demonstrate a high accuracy for PET in staging lymphoma, but few assess observer variation. This study quantified agreement for staging lymphoma with PET/CT. **Methods:** The PET/CT images of 100 patients with lymphoma who had been referred for staging were reviewed by 3 experienced observers, with 2 observers reviewing each series a second time. Ann Arbor stage and individual nodal and extranodal regions were assessed. Weighted κ (κ_w) and intraclass correlation coefficient were used to compare ratings. **Results:** Intra- and interobserver agreement was high for Ann Arbor stage (κ_w = 0.79–0.91), number of nodal regions involved (intraclass correlation coefficient, 0.83–0.93), and presence of extranodal disease (κ = 0.74–0.86). High agreement was also observed for all nodal regions (κ_w > 0.60) except hilar (κ_w = 0.56–0.82) and infraclavicular (κ_w = 0.14–0.55). Lower agreement was observed for bowel involvement (κ_w = 0.37–0.71). **Conclusion:** Experienced observers had a high level of agreement using PET/CT for lymphoma staging, supporting its use as a robust noninvasive staging tool. Further research is needed to evaluate observer variability for restaging during and after chemotherapy.

Key Words: fluorodeoxyglucose; FDG; positron emission tomography; PET/CT; lymphoma; reporter agreement

J Nucl Med 2009; 50:1594–1597
DOI: 10.2967/jnumed.109.064121

Many studies have assessed the sensitivity and specificity of PET and PET/CT for staging lymphoma (*1*), but few have analyzed variation between observers. Although a high level of reproducibility does not necessarily equate to high accuracy, low levels of reproducibility cannot be associated with high accuracy. Observer variation can be substantial, and differences between observers can outweigh purported differences between imaging techniques (*2*). Assessment is particularly pertinent with the newer imaging modalities that generate hundreds of images for review from a single patient encounter. Quantifying interpretative variability is important, especially in view of increasing PET/CT use in multicenter trials designed to establish how functional imaging results can be used to alter patient management.

This study quantified both intra- and interobserver variation for staging lymphoma with PET/CT. Primary outcome variables were agreement on Ann Arbor stage, number of nodal regions, and presence of extranodal involvement. Secondary outcome variables were agreement on specific nodal and extranodal sites.

Received Mar. 12, 2009; revision accepted Jul. 10, 2009.
For correspondence or reprints contact: Michael S. Hofman, Centre for Molecular Imaging, Peter MacCallum Cancer Centre, Locked Bag 1, A'Beckett St., Melbourne 8006, Australia.
E-mail: nucmedpet@drhofman.com

MATERIALS AND METHODS

Patients

One hundred consecutive PET/CT studies of patients with biopsy-proven lymphoma who underwent PET/CT staging before therapy were reviewed. The identity of all patients was masked, and no history or correlative imaging was provided. The observers used a standardized form and rated individual nodal groups as negative, equivocal, or positive for disease. Negative included inflammatory, reactive, or other benign etiologies. Nodal regions, as defined at the Rye Symposium in 1965, included cervical (including supraclavicular, occipital, and preauricular), axillary, infraclavicular, mediastinal, hilar, periaortic, mesentery, pelvic, and inguinal or femoral. Extranodal sites included spleen, bone or bone marrow, lung, liver, bowel (including gastric), and other (including sites such as muscle, subcutaneous tissue, and breast). Ann Arbor stage was also assigned as per the sixth edition classification of the American Joint Committee on Cancer (*3*).

Three observers reviewed the same 100-patient series. Two of these observers reviewed the same patient series a second time in a different order to assess intraobserver variability, with reviews separated by several weeks to reduce the effect of memory. Thus, 500 reviews were conducted in total. All observers were experienced in reporting PET/CT. Two were nuclear medicine physicians, and one was a radiologist. The observers had a minimum of 8 y and a maximum of 17 y of experience with PET reporting, and all had 3 y of experience of PET/CT reporting in a unit performing approximately 4,000 studies per year, of which 800 were lymphoma.

The characteristics of the study population are summarized in Tables 1 and 2. Patient age ranged from 11 to 80 y (median, 53 y). The study population included patients with high-grade non-Hodgkin lymphoma (57%), Hodgkin lymphoma (32%), follicular lymphoma (9%), and other subtypes (2%). The average number of nodal regions was 4.2, with extranodal involvement in 45% of [Table 1] [Table 2]

TABLE 1. Lymphoma Subtypes

Subtype	*n*
Diffuse large B cell	40
Hodgkin	32
Follicular	9
T cell	9
High-grade non-Hodgkin, unspecified	3
Burkitt	1
Anaplastic large cell	2
Mantle cell	2
Uncertain	1
Posttransplant lymphoproliferative disorder	1

patients. The proportions of patients with Ann Arbor stages 1, 2, 3 and 4 were 15.6%, 22.8%, 28.2%, and 33.4%, respectively.

PET/CT Acquisition

The studies were performed from skull base to upper thighs 90 min after injection of ^{18}F-FDG on a dual-modality PET/CT scanner (Discovery ST; GE Healthcare). The images were acquired in 2-dimensional mode and were reconstructed with an iterative technique.

PET/CT Interpretation

All cases were reviewed on a Hermes workstation (Nuclear Diagnostics) volume display. Viewing conditions such as the physical environment, monitor brightness, or background lighting were not standardized. All images were scaled to a standardized uptake value upper threshold of 10 using a gray scale. Areas of increased ^{18}F-FDG uptake not considered physiologic were generally reported as areas of nodal or extranodal involvement. Correlative low-dose CT findings were incorporated for anatomic localization and further differentiation between physiologic, inflammatory, and lymphomatous etiologies.

Our unit has used a reporting routine of 2 observers who read the scans independently. If there is disagreement, a third observer issues the consensus report. This routine is likely to result in similar thresholds for reporting, as reinforced by feedback from multidisciplinary meetings of hematologists, oncologists, and pathologists.

TABLE 2. Percentage of Patients with Lymphomatous Involvement at Specific Nodal and Extranodal Sites

Site	Percentage
Nodal	
Cervical	55.8 (L); 47.4 (R)
Axillary	28.8 (L); 28.0 (R)
Infraclavicular	6.4 (L); 4.2 (R)
Hilar	23.0
Mediastinal	38.8
Periaortic	37.8
Pelvic	31.4 (L); 31.6 (R)
Inguinal	33.6 (L); 28.6 (R)
Extranodal	
Spleen	17.8
Bone marrow or bone	20.0
Lung	5.4
Liver	10.0
Bowel or gastric	5.4
Other nodal sites*	8.0

*Muscle, subcutaneous tissue, breast, and uterus.

Statistical Analysis

Levels of agreement were quantified using weighted κ (κ_w) (*4,5*), and intraclass correlation coefficient for continuous measures (*6*). Weighting was defined as zero credit being given for the most extreme discrepancies possible and the highest partial credit being given for the least discrepant pairs of ratings. For analysis, a conservative interpretative threshold was used, with nodal and extranodal regions assigned a value of 0 for benign and equivocal, and 1 for malignant. For nodal stations, left- and right-sided scores were added together; that is, a value of 2 was assigned if both sides were involved. κ-values are reported using the benchmarks of Landis and Koch (*7*) (with 0.81–1 being almost perfect agreement; 0.61–0.8, substantial agreement; 0.41–0.6, moderate agreement; 0.21–0.4, fair agreement; 0.01–0.2, slight agreement; and ≤0, poor agreement). Bootstrapping was used to calculate 95% confidence intervals (*8*). Analyses were performed using the statistical package Stata (version 9.2; Stata Corp).

Postanalysis Review

For those variables with the lowest κ-agreement, the 3 observers were asked to review and reach a consensus without any further information and then to review again with the aid of relevant clinical history and correlative imaging. The observers were also asked to postulate the main reasons for the observer disagreement. The instances of disagreement included 18 patients with infraclavicular nodal involvement and 8 patients with bowel involvement.

RESULTS

Results are summarized in Table 3. Intra- and interobserver agreement was high for overall Ann Arbor stage (κ_w = 0.79–0.91). For intraobserver agreement, 95% confidence intervals were within the "almost perfect" range of κ-values, whereas for interobserver agreement, there was crossover into the substantial-agreement range. Similarly, agreement was high for the total number of nodal groups involved (intraclass correlation coefficient, 0.83–0.93) and for presence of extranodal involvement (intra- and interobserver κ_w = 0.82–0.84 and 0.74–0.86, respectively). [**Table 3**]

For specific nodal groups, there was substantial or greater agreement for cervical (κ_w = 0.77–0.86), axillary (0.69–0.80), pelvic (0.65–0.82), inguinal (0.69–0.82), mediastinal (0.73–0.77), periaortic (0.75–0.81), and mesenteric (0.61–0.67) nodal groups. Agreement was lower for hilar (0.56–0.82) and infraclavicular (0.14–0.55) nodal groups. For extranodal involvement, there was substantial or greater agreement for spleen (0.69–0.84) and bone marrow (0.76–0.94). Agreement was lower for lung (0.58–0.90), liver (0.59–0.95), and bowel (0.37–0.71). The higher κ-values in these ranges reflect greater intraobserver agreement.

In the postanalysis review of patients with variability of infraclavicular nodal classification, all patients had distant disease and the variability did not change the overall stage. On consensus review, the variability was clearly due to variable definition of the boundary between infraclavicular

jnm064121-sn ■ 9/8/09

TABLE 3. Intra- and Interobserver Variability

Parameter	Intraobserver		Interobserver		
	1	2	1 vs. 2	2 vs. 3	1 vs. 3
Ann Arbor	0.91 (0.82–0.97)	0.88 (0.80–0.95)	0.81 (0.69–0.90)	0.79 (0.68–0.86)	0.87 (0.77–0.94)
Extranodal	0.82 (0.70–0.93)	0.86 (0.76–0.96)	0.74 (0.61–0.87)	0.82 (0.71–0.93)	0.76 (0.63–0.89)
No. of nodal groups	0.93 (0.90–0.96)	0.91 (0.87–0.94)	0.83 (0.76–0.89)	0.92 (0.89–0.95)	0.88 (0.83–0.92)
Nodal sites					
Cervical*	0.84 (0.72–0.92)	0.86 (0.78–0.93)	0.81 (0.71–0.90)	0.77 (0.66–0.86)	0.79 (0.68–0.88)
Axillary	0.80 (0.68–0.89)	0.74 (0.61–0.85)	0.69 (0.56–0.81)	0.69 (0.53–0.83)	0.73 (0.60–0.84)
Infraclavicular	0.55 (−0.01–0.89)	0.39 (0.10–0.73)	0.23 (−0.02–0.55)	0.37 (−0.01–0.68)	0.14 (−0.04–0.50)
Hilar	0.82 (0.70–0.95)	0.65 (0.48–0.83)	0.56 (0.37–0.75)	0.58 (0.36–0.79)	0.63 (0.45–0.81)
Pelvic	0.82 (0.70–0.91)	0.68 (0.53–0.81)	0.65 (0.53–0.76)	0.71 (0.60–0.82)	0.68 (0.55–0.79)
Inguinal or femoral	0.82 (0.72–0.91)	0.69 (0.55–0.82)	0.71 (0.59–0.82)	0.76 (0.64–0.87)	0.69 (0.55–0.82)
Mediastinal	0.75 (0.61–0.88)	0.75 (0.61–0.88)	0.75 (0.61–0.88)	0.77 (0.64–0.90)	0.73 (0.59–0.87)
Periaortic	0.77 (0.65–0.90)	0.76 (0.63–0.89)	0.75 (0.62–0.88)	0.81 (0.69–0.93)	0.78 (0.65–0.90)
Mesentery	0.65 (0.47–0.83)	0.61 (0.41–0.81)	0.63 (0.44–0.81)	0.61 (0.41–0.81)	0.67 (0.49–0.85)
Extranodal sites					
Spleen	0.84 (0.69–0.99)	0.81 (0.67–0.96)	0.75 (0.57–0.92)	0.81 (0.67–0.96)	0.69 (0.51–0.88)
Bone marrow	0.94 (0.85–0.99)	0.81 (0.67–0.96)	0.76 (0.60–0.92)	0.76 (0.60–0.92)	0.93 (0.84–0.99)
Lung	0.82 (0.58–0.99)	0.90 (0.72–0.99)	0.58 (0.21–0.95)	0.58 (0.21–0.95)	0.65 (0.32–0.97)
Liver	0.95 (0.84–0.99)	0.88 (0.71–0.99)	0.84 (0.66–0.99)	0.78 (0.57–0.99)	0.59 (0.32–0.87)
Bowel	0.71 (0.40–0.99)	0.56 (0.11–0.99)	0.64 (0.35–0.93)	0.59 (0.28–0.91)	0.37 (−0.03–0.76)

*Cervical includes supraclavicular and all head and neck nodal stations.

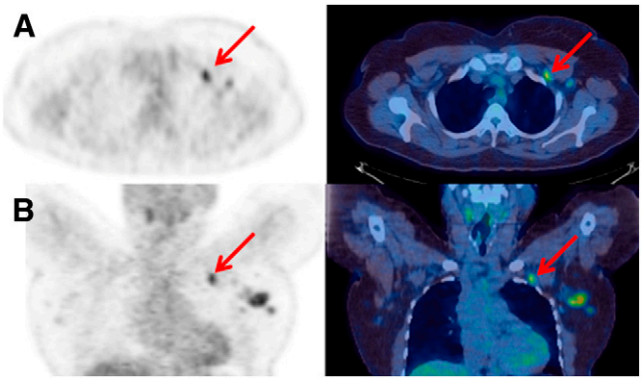
Data are intraclass correlation coefficient for number of nodal groups and κ_w for all other variables, with 95% confidence intervals in parentheses.

nodes and adjacent nodal regions, including medial axillary and supraclavicular (Fig. 1). Review of the 8 patients with disagreement on bowel involvement indicated that 2 were related to gastric involvement and 6 to large-bowel involvement. On consensus review, all these patients had either stage 3e or 4 disease. In the cases of large-bowel involvement, the disagreement was related to differentiating colonic involvement from adjacent mesenteric nodal disease (Fig. 2).

[Fig. 1] [Fig. 2]

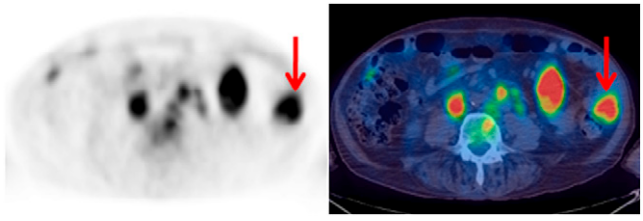
DISCUSSION

Few studies have assessed observer variability in lymphoma imaging. Fletcher et al. (*9*) evaluated interobserver variability for CT detection of cervical–thoracic Hodgkin disease. For individual nodal stations, agreement ranged from poor to moderate, with κ-scores ranging from 0.13 for left paratracheal nodes to 0.72 for right lower cervical nodes. Agreement between the majority of reviewers and the primary report was poor ($\kappa < 0.40$) for two thirds of the sites. Zijlstra et al. (*10*) measured observer variation for PET staging and restaging. For experts using sensitive and conservative models, concordance was 61% and 56%,



RGB

FIGURE 1. Axial (A) and coronal (B) PET and PET/CT images demonstrating focal increased metabolic activity in subcentimeter node (arrow) that was variably reported as supraclavicular, infraclavicular, or axillary. This did not result in a change in overall stage due to the presence of widespread disease, and this variability would therefore not change management.



RGB

FIGURE 2. Axial PET and PET/CT images demonstrating large-bowel lymphomatous involvement (arrow) in addition to mesenteric nodal involvement. Some reviewers reported only mesenteric nodal involvement, but involvement of bowel wall was clearly evident on review.

respectively, for staging and 82% and 94%, respectively, for restaging.

In this study, intra- and interobserver agreement was high for Ann Arbor stage, number of nodal regions involved, and presence of extranodal disease. Most of the study patients (89%) had high-grade lymphoma, in which high glucose use results in high lesion-to-background contrast. This facilitates easy visual perception of active sites of disease and is further assisted by the use of contemporaneous CT, which allows precise anatomic correlation and the identification of physiologic variants or other pathologic processes.

The lower agreement for extranodal bowel involvement appears to be due largely to difficulty in differentiating bowel involvement from adjacent mesenteric nodal disease. This difficulty is not unexpected on imaging alone and is unlikely to alter the management strategy. Interpretation is limited by the low prevalence of patients with bowel involvement (<5.5%); κ-statistics also tend to weigh disagreements more heavily when the prevalence of a positive finding approaches zero (*11*).

For specific nodal regions, agreement was lowest for infraclavicular nodes, as is consistent with previous findings for CT staging (*9*). The cause was disagreement about the definitions of infraclavicular, supraclavicular, and axillary nodal regions, but this disagreement did not change the overall stage, and locoregional radiotherapy would not have been considered. There was also lower agreement about hilar nodes than about other nodal stations. Hilar ^{18}F-FDG uptake is not uncommon, because of inflammatory changes secondary to a reactive or granulomatous process (*12*). In staging lung carcinoma with PET/CT, we have previously described lower agreement in the hilar region than in mediastinal nodal stations (*13*). Caution is thus warranted in interpretation of hilar activity, especially if the intensity is discordant with uptake at other sites.

This study had several potential sources of error. The form-based method of data collection may have reduced errors by ensuring systematic review and standardization of terminology. PET/CT findings were not compared with histology or patient follow-up, and therefore no comment can be made on accuracy. Other studies, however, have demonstrated high sensitivity and specificity (*1*). In this study, reviewers were unaware of clinical history and were not provided with correlative imaging results—a discrepancy with routine clinical practice. Availability of this information, however, would likely result in higher agreement. For example, knowledge of lymphoma subtype will help the reviewer refine expected distribution, likely improving agreement.

This study was limited by the use of experienced PET/CT observers from a single center. As such, they could be expected to show a significant degree of concordance in their approach, especially as the center adopts a dual-reporting system. The findings are still of interest, as multiinstitutional trials increasingly use a central core laboratory for reporting. It would be useful to extend the study to assess variation across different centers and also investigate agreement with less experienced observers or trainees.

Our study did not assess observer agreement for restaging lymphoma after chemotherapy. Observer agreement may be lower in these patients because the intensity of ^{18}F-FDG uptake may be low. Although the Imaging Subcommittee of the International Harmonization Project in Lymphoma has published positivity criteria for restaging after completion of chemotherapy (*14*), defining positivity when using PET for restaging during a course of therapy is less well established. Defining appropriate criteria and assessing observer variability in these patients warrant further investigation.

CONCLUSION

Among experienced physicians in a single center, there was a high level of intra- and interobserver agreement using PET/CT for lymphoma staging. This result complements the results of other studies demonstrating a high accuracy and supports the use of PET/CT as a robust noninvasive staging tool. Further research is needed to evaluate observer variability for restaging during and after completion of chemotherapy.

REFERENCES

1. Kirby AM, Mikhaeel NG. The role of FDG PET in the management of lymphoma: what is the evidence base? *Nucl Med Commun.* 2007;28:335–354.
2. Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. *Br J Radiol.* 1997;70:1085–1098.
3. Greene FL, American Joint Committee on Cancer, American Cancer Society. *AJCC Cancer Staging Manual.* 6th ed. New York, NY: Springer; 2002.
4. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
5. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;70:213–220.
6. Dunn G, Everitt B. *Clinical Biostatistics: An Introduction to Evidence-Based Medicine.* New York, NY: Edward Arnold; 1995.
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
8. Efron B, Tibshirani R. *An Introduction to the Bootstrap.* New York, NY: Chapman and Hall; 1993.
9. Fletcher BD, Glicksman AS, Gieser P. Interobserver variability in the detection of cervical-thoracic Hodgkin's disease by computed tomography. *J Clin Oncol.* 1999;17:2153–2159.
10. Zijlstra JM, Comans EF, van Lingen A, et al. FDG PET in lymphoma: the need for standardization of interpretation: an observer variation study. *Nucl Med Commun.* 2007;28:798–803.
11. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol.* 1988;41:949–958.
12. Karam M, Roberts-Klein S, Shet N, Chang J, Feustel P. Bilateral hilar foci on ^{18}F-FDG PET scan in patients without lung cancer: variables associated with benign and malignant etiology. *J Nucl Med.* 2008;49:1429–1436.
13. Hofman MS, Smeeton NC, Rankin SC, Nunan T, O'Doherty MJ. Observer variation in FDG PET-CT for staging of non-small-cell lung carcinoma. *Eur J Nucl Med Mol Imaging.* 2009;36:194–199.
14. Juweid ME, Stroobants S, Hoekstra OS, et al. Use of positron emission tomography for response assessment of lymphoma: consensus of the Imaging Subcommittee of International Harmonization Project in Lymphoma. *J Clin Oncol.* 2007;25:571–578.