MOLECULAR IMAGING AND BIOLOGY

Published Online: 3 July 2007

Mol Imaging Biol (2007) 9:318–322
DOI: 10.1007/s11307-007-0108-1

RESEARCH ARTICLE

Observer Variation of 2-Deoxy-2-[F-18] fluoro-D-Glucose-Positron Emission Tomography in Mediastinal Staging of Non-Small Cell Lung Cancer as a Function of Experience, and its Potential Clinical Impact

Sietske A. Smulders,[1]Chad M. Gundy,[2]Arthur van Lingen,[3]Emile F. Comans,[3] Frank W J M Smeenk,[1]Otto S. Hoekstra,[2,3] The Study Group of Clinical PET

[1]*Department of Pulmonary Diseases, Catharina Hospital Eindhoven, Amsterdam, The Netherlands*
[2]*Departments of Clinical Epidemiology and Biostatistics, VU University, Amsterdam, The Netherlands*
[3]*Nuclear Medicine and PET research, VU University Medical Centre Amsterdam, VU University, PO Box 7057, 1007 MB, Amsterdam, The Netherlands*

Abstract

Purpose: To test the extent of variation among nuclear medicine physicians with respect to staging non-small cell lung cancer with positron emission tomography (PET).

Procedures: Two groups of nuclear medicine physicians with different levels of PET experience reviewed 30 PET scans. They were requested to identify and localize suspicious mediastinal lymph nodes (MLN) using standardized algorithms. Results were compared between the two groups, between individuals, and with expert reading.

Results: Overall we found good interobserver agreement (kappa 0.65). Experience with PET translated into a better ability to localize MLN stations (68% vs. 51%, respectively), and experienced readers appeared to be more familiar with translating PET readings into clinically useful statements.

Conclusions: Although our results suggest that clinical experience with PET increases observers' ability to read and interpret results from PET adequately, there is room for improvement. Experience with PET does not necessarily improve the accuracy of image interpretation.

Key words: FDG-PET scanning, Interobserver variation, Lung cancer, Experience, Mediastinal lymph node metastases

Introduction

In non-small cell lung cancer (NSCLC), proven ipsi- (N2) or contralateral (N3) mediastinal lymph node involvement often precludes cure by surgery. 2-deoxy-2-[F-18]fluoro-D-glucose-positron emission tomography (FDG-PET) is used to stage NSCLC patients. The yield of whole-body PET pertains to typing the primary pulmonary lesion and on the preoperative identification of distant and lymph node metastases. Moreover, PET simplifies and improves lymph node evaluation by setting the indication for biopsy and improving its yield. Mediastinoscopy is the standard technique of invasive lymph node staging but the results in daily practice are quite variable [1]. It has been suggested that the proportion of tumor-positive procedures increases if

Correspondence to: Otto S. Hoekstra; *e-mail:* os.hoekstra@vumc.nl

guided by PET [2, 3]. So far, mediastinoscopy is the most often used invasive method, but more recently, endoscopic techniques [like transesophageal ultrasound-guided fine needle aspiration (EUS-FNA)] have been developed. Because the mediastinal areas covered by mediastinoscopy and EUS-FNA are largely complementary, proper localization of possible malignant nodes is important to assign patients to the appropriate procedure. FDG-PET criteria of test positivity for mediastinal lymph node staging are based on recognition of focally enhanced uptake ("hot spots") vs. background, rather than on quantitative assessment (like the 1-cm short axis criterion with CT scanning). Results from PET studies pertaining to its accuracy in mediastinal staging are robust [4], but as the technique is disseminating, observer variation and learning curves still need to be documented.

The aim of the present study was to measure the observer agreement and accuracy vs. expert readings of mediastinal lymph nodes in NSCLC staging with FDG-PET at various levels of complexity and as a function of experience.

Materials and Methods

Study Design

We used a set of 30 PET scans from the study by Joshi et al. [5] of consecutive patients referred for staging to the Department of Nuclear Medicine and PET Research of the Vrije University Medical Centre. To obtain an adequate case mix, we included scans of patients with a range of mediastinal lymph node sizes at CT scanning: (1) ≤10 mm short axis diameter (n=10), (2) 10.1–15 mm (n=10), and (3)>15 mm (n=10). PET scans had been performed according to the standard protocol in our institution using a full ring BGO PET scanner (ECAT EXACT HR+, CTI/Siemens, starting 60 min after 370 MBq ^{18}FDG) [5].

The scans were analyzed by 14 nuclear medicine physicians who had extensive experience with SPECT but variable expertise with PET and mediastinal lymph node staging in NSCLC: seven had no personal experience with PET (the "inexperienced group"), whereas the others had at least 1 year of experience with PET in NSCLC patients in their own clinical practice, which comprised access to mobile PET once every 1 or 2 weeks (the more "experienced group"). On average, the inexperienced group had reviewed 0–15 PET scans, each compared to a 100–150 (with at least 50% NSCLC) each in the experienced group. Prior to this study, the observers had been instructed in workshops by two expert PET readers, a pulmonologist, and a surgeon about the concepts, principles, and practice of mediastinal staging in NSCLC by PET and other methods. The results of all observers were compared to the combined judgment of two expert nuclear medicine physicians (EFC and OSH), and the latter readings were used as the gold standard. The expert readers had been working together in the same university hospital for numerous years and had a broad experience with PET [6–8].

We developed a software tool running Matlab 5.3, which allowed simultaneous visualization of PET images in the axial, coronal, and sagittal planes (at 5 or 10 mm slice thickness), with possible cross linking. Each observer was requested to identify and interpret any abnormal hot spot representing primary tumor or lymph node, blinded for the results of the other readers. This software tool was installed on the personal computer of each observer, and the results were electronically stored for analysis. To be able to accurately relate the results of different observers, the coordinates of each hot spot identified by an observer were stored and linked to the assigned interpretation. Because none of the observers had worked with this software before, we provided a test set (derived from the original data set) of three scans to each observer prior to the study. These three scans comprised 29 separate abnormal mediastinal lymph node localizations and therefore provided an adequate way to practice working with Naruke's map of lymph node localizations (adapted from Mountain and Dresler) [9]. Observers had knowledge of the clinical information provided with the original PET scan referral, except for the mediastinal stage at CT.

Data Acquisition

The observers were asked to interpret abnormal hot spots pertaining to the primary tumor and lymph nodes in terms of their *localization* and *likelihood of malignancy* using the classification systems shown in Table 1. The criterion for test positivity was the presence of focally enhanced uptake vs. background. Furthermore, observers were asked to formulate a *recommendation* with respect to the next *management step* to the referring clinician (Table 1). In this context, we instructed them to use the following protocol: (1) recommend biopsy of mediastinal lymph nodes in case of suspected (hilar or mediastinal) lymph node involvement and in case of tumors adjacent to the mediastinum or hilus, (2) recommend thoracotomy in case of a peripheral primary tumor without suspicious mediastinal lymph nodes at PET, and (3) recommend an expectative ("wait and see") policy in case PET shows no abnormal uptake in either the primary site nor in lymph nodes. For the purpose of the present investigation, they were instructed to ignore possible suspicious extrathoracic localizations in these management considerations.

Table 1. Classification system of tumor and lymph nodes

Characteristic	Classification
Primary tumor	
Presence	No tumor present
	Primary tumor
	Second primary
Localization	Peripheral
	Adjacent to mediastinum
	Adjacent to hilus
Lymph node localization[a]	No lymph nodes present
	N1 L/R
	N2 L/R
	N3
	N4 L/R
	N5/* N6/* N7
	N8 L/R
	N9 L/R
	N10 L/R
	Clavicular L/R
Likelihood of malignancy	Definitely benign
	Probably benign
	Equivocal
	Probably malignant
	Definitely malignant
Management recommendation	Invasive lymph node evaluation
	Thoracotomy
	Expectative policy

[a]According to the map of lymph node definitions by Mountain and Dresler

Data Analysis

Using the individual scores of the observers, we assigned an "N-stage according to PET" for each observer and each patient using the following classification:

1. N0, peripheral primary tumor, no mediastinal hot spot
2. N1, peripheral primary tumor and separate hot spot considered to be a hilar lymph node
3. N0–N1, primary tumor within hilar area, no separate mediastinal hot spot
4. N0–N2, primary tumor adjacent to mediastinum, no separate mediastinal hot spot
5. N2, hot spot compatible with ipsilateral mediastinal lymph node
6. N3, hot spot compatible with contralateral mediastinal or clavicular lymph node

We performed a more detailed analysis of the nature of the errors in the "management recommendation" classification vs. the expert reading, identifying whether these errors followed the observers' own interpretation of suspicious lymph node stations or resulted from true errors (protocol violation). For example, the former situation occurred if, in case of a peripheral primary tumor, an observer considered the ipsilateral right lower tracheobronchial station to be positive at PET, whereas the expert only identified the primary lesion. The resulting discrepant management recommendations (mediastinoscopy vs. thoracotomy, respectively) directly flow from these classifications. We coined such an incorrect answer as a *mistake* (M). However, if this observer would have advised to proceed directly to thoracotomy, this was considered a *protocol violation* (P).

We also measured how accurately readers could define and localize suspected mediastinal lymph node stations at PET. Compatible with known limitations of PET with respect to spatial resolution and accounting for different levels of clinical relevance, we accepted the following differences of nodal classifications: Naruke stations 1 and 2 [left (L)/right (R), respectively], 4R and 10R, 4L and 10L and 5, and 8 and 9 (L/R, respectively). Using this simplified system, we analyzed whether observers defined and localized suspected lymph node metastases vs. the expert readings "correctly," "incorrectly," or "not at all."

Statistical Analysis

Statistical analysis was done by SPSS version 13.0 software. To determine interobserver agreement regarding "management recommendation" and "N-stage," and to compare this to expert readings, we calculated the Kappa coefficients, using AGREE version 7.2. We used weighted kappa's for the N-stage analysis. Furthermore, to detect potential differences between the two groups of observers with different PET experience with respect to the nature of the management recommendation errors, and the classification of separate mediastinal hot spots, we used the Wilcoxon–Mann–Whitney test. Statistical significance was set at $p<0.05$.

Results

The 30 PET scans comprised a total of 89 locations of suspected malignancy, according to the gold standard (expert reading). Thirty-four represented tumor locations, 55 were lymph nodes (10 hilar, 39 mediastinal, and six supraclavicular). According to expert readers, there was a mean of three sites (primary lesion and lymph nodes) per patient (range 1–13). The experts classified (according to Table 1) 82 lesions as "definitely malignant," five as "probably malignant," and two as "equivocal." In the final analysis, these "probably" and "definitely" malignant locations were classified as malignant. The expert N-stage classifications included nine "N0," three "N1," one "N0–N1," three "N0–N2," nine "N2," and five "N3," according to the definitions mentioned earlier.

Management recommendations were correct in 80% of cases (86 errors out of 420 recommendations, 42 in the experienced group and 44 in the inexperienced group). The accuracy vs. expert reading was moderate (kappa 0.59) at either level of experience (Table 2). The level of agreement among inexperienced observers tended to be lower but did not reach significance. Four scans accounted for a total of 38 errors (44%), while not a single mistake by any observer was made in eight.

In the group of inexperienced readers, 29 (of 44; 66%) of the incorrect management recommendations were protocol violations (type "P"), vs. 17 (of 42; 40%) in the experienced readers group ($p=0.12$). On the contrary, errors that directly flow from reading errors (type "M") were significantly more prevalent in the group of experienced readers (25 out of 42=59%), vs. 15 out of 44 (34%) in the inexperienced readers group ($p=0.03$).

Table 2. Interobserver agreement and accuracy as a function of experience with respect to the classification of "N-stage" and "management recommendation"

	Inexperienced observers (n=7)	Experienced observers (n=7)	Overall
Management recommendation[a]			
Agreement vs. expert	0.60 (0.42–0.77)	0.58 (0.37–0.79)	0.59 (0.42–0.76)
Pair wise agreement	0.48 (0.35–0.62)	0.56 (0.41–0.71)	0.50 (0.37–0.63)
N-stage[b]			
Agreement vs. expert	0.58 (0.36–0.80)	0.72 (0.55–0.88)	0.65 (0.47–0.83)
Pair wise agreement	0.56 (0.44–0.68)	0.61 (0.49–0.74)	0.58 (0.46–0.69)

[a]Kappa (95% confidence interval)
[b]Weighted kappa (95% confidence interval)

Common errors (type "P", protocol violations) were, e.g., to recommend "expectative policy" or "directly to thoracotomy" in a patient without enhanced PET uptake in primary tumor and mediastinal lymph nodes. However, the provided clinical information stated that bronchoalveolar cell carcinoma had been proven histologically. Therefore, "mediastinal lymph node evaluation" should have been recommended because the mediastinum in a patient with adenocarcinoma without FDG uptake of the primary tumor cannot be reliably evaluated so that histological confirmation of the mediastinum is required.

N-stage classifications were correct in 68% of cases (286 out of 420 assigned N-stages, 138 in the inexperienced group and 148 in the experienced group). Experienced observers tended to have a better agreement with the expert reading than inexperienced ones (weighted kappa's 0.72 and 0.58, respectively). N-stages were overestimated in 17.4% (16.7% by the experienced and 18.1% by the inexperienced observers) and underestimated in 14.5% of cases (12.9 and 16.2%, respectively). The individual scores of the observers (Table 3) reveal that errors in either direction were made by most of them.

Because we used three scans to practice on localizing mediastinal lymph nodes, 27 scans remained with 26 separate lymph node localizations. The detection rate of individual mediastinal lymph node stations was similar for inexperienced and experienced observers (71 and 74%, respectively, Table 4), and the variation within the groups was also comparable. However, experienced readers were better at localizing the stations than inexperienced readers were (correct in 68 vs. 51%, respectively). The most common mislocalizations (Table 5) were to classify right tracheobronchial stations (4R) as upper-right paratracheal (2R), subcarinal (7) as right tracheobronchial (4R), and left para-esophageal (8/9L) as left tracheobronchial (4L).

Table 3. Details on N-stage (using the classification system described in the methods section) in 30 scans for each observer

	N-stage classified correctly [% (*n*)][a]	N-stage overestimated [% (*n*)][b]
Inexperienced observers		
INEXP 1	70.0 (21)	20.0 (6)
INEXP 2	56.7 (17)	20.0 (6)
INEXP 3	70.0 (21)	13.3 (4)
INEXP 4	63.3 (19)	23.3 (7)
INEXP 5	66.7 (20)	20.0 (6)
INEXP 6	66.7 (20)	20.0 (6)
INEXP 7	66.7 (20)	10.0 (3)
Total	65.7 (138)	18.1 (38)
Experienced observers		
EXP 1	63.3 (19)	30.0 (9)
EXP 2	76.7 (23)	6.7 (2)
EXP 3	73.3 (22)	10.0 (3)
EXP 4	73.3 (22)	20.0 (6)
EXP 5	73.3 (22)	13.3 (4)
EXP 6	73.3 (22)	16.7 (5)
EXP 7	60.0 (18)	20.0 (6)
Total	70.5 (148)	16.7 (35)

[a]Percentage of N-stages classified correctly vs. expert reading
[b]Percentage of overestimated N-stages vs. expert reading

Table 4. Accuracy of inexperienced and experienced observers to detect and localize the 26 mediastinal lymph node stations present according to the expert reading

	Identified [% (*n*)][a]	Correctly localized [% (*n*)][b]
Inexperienced observers		
INEXP 1	76.9 (20)	30.0 (6)
INEXP 2	84.6 (22)	63.6 (14)
INEXP 3	61.5 (16)	62.5 (10)
INEXP 4	80.8 (21)	23.8 (5)
INEXP 5	69.2 (18)	55.6 (10)
INEXP 6	65.4 (17)	64.7 (11)
INEXP 7	57.7 (15)	66.7 (10)
Total	70.9% (129)	51.2% (66)
Experienced observers		
EXP 1	76.9 (20)	65.0 (13)
EXP 2	61.5 (16)	81.3 (13)
EXP 3	69.2 (18)	83.3 (15)
EXP 4	73.1 (19)	89.5 (17)
EXP 5	84.6 (22)	77.3 (17)
EXP 6	80.8 (21)	42.9 (9)
EXP 7	69.2 (18)	38.9 (7)
Total	73.6% (134)	67.9% (91)

[a] Percentage of identified nodal stations vs. expert reading
[b] Percentage of correctly localized nodal stations vs. expert reading (e.g., INEXP 1 identified 20 out of the 26 stations, and 6 out of 20 were localized correctly)

Discussion

Observer variation is the Achilles' heel of diagnostic imaging [10] and especially of tests that apply visual interpretation. It is therefore surprising that the clinical PET literature contains few studies on observer variation beyond the level of occasional reports on variation between two observers participating in an accuracy study. The present study reports on the results of 14 observers stratified by their experience with PET, and it accounts for several aspects of the clinical context of NSCLC staging (management recommendation, N-stage, nodal stations). We found that the accuracy (vs. expert reading) was moderate to

Table 5. Mediastinal lymph node stations by experienced and inexperienced observers, according to Mountain and Dresler

Expert (CA)	Experienced and inexperienced observers										
	2R	3	4L	4R	6	7	8R	8L	SC	T[a]	Missed
2 R (1 R)	4									1	9
3	7			3						1	3
4 L (5, 10 L)			24		1					14	31
4 R (10 R)	18	1		85		1	1		1	14	19
7		1	2	13		15				1	25
8 R (9 R)				4			8			1	1
8 L (9 L)			9		1			8		1	9
SC [c]	2			2					14	6	3

Mediastinal lymph node stations using the simplified system mentioned in the "Materials and Methods" section regarding the acceptance of different lymph node classifications, consistent with clinical practice, for expert and both groups of observers
CA = correct alternative according to simplified system, SC = supra- or infraclavicular lymph nodes, T = tumor
[a]Observer identified pertaining mediastinal lymph node as primary or second primary tumor

substantial at moderate levels of interobserver agreement. Our results suggest that clinical experience with PET improves the ability of readers to localize mediastinal hot spots correctly, and this is relevant with respect to the next clinical step: i.e., to decide which invasive verification method should follow and to enhance the yield of such procedures. Moreover, within the more experienced group, the agreement of assigning N-stages and management recommendations tended to be better. Finally, familiarity with clinical practice and staging protocols for NSCLC patients may have contributed to fewer inconsistencies in management recommendations. Our management advice constructs were designed to account for generally recognized limitations of PET in mediastinal staging.

With slightly different endpoints, the interobserver agreement of CT reading appears to be similar to what we have reported for PET: in CT evaluation of mediastinal lymph node size, Guyatt et al. reported a kappa of 0.61 regarding the presence of any nodes greater than 1 cm on CT scan [11]. However, agreement in different nodal groups varied widely, and it appeared to be far more difficult for the left superior mediastinal nodes. In our study, we found that some mistakes were made relatively more often regarding localizing separate lymph nodes (Table 5). With the increasing clinical methods to verify imaging findings (transesophageal, transbronchial EUS-FNA, mediastinoscopy, video-assisted thoracoscopy), the relevance of interpreting images at the nodal level is growing. PET–CT helps to improve the yield of PET and CT reading in patients newly presenting with lung cancer, but also in restaging after neoadjuvant therapy. Using PET–CT in this study, instead of PET alone, would probably have been more clinically relevant. However, we believe that the errors related to localizing suspicious foci will improve with PET–CT, but this is not the case for detection and interpretation errors. Other limitations of our study were the relative unfamiliarity of the observers with the display and registration software and, perhaps, the lack of standardized computer screens.

In the Netherlands, the availability of FDG-PET is rapidly expanding, even in smaller hospitals, and this has major implications for local nuclear medicine physicians, as well as for residents. To our knowledge, the duration of time that is needed before results on PET are adequately reviewed and interpreted ("the learning curve") by nuclear medicine physicians is unknown. We had anticipated striking differences between experienced and inexperienced readers, but this was not the case. However, there was obvious room for improvement in the experienced group and we suggest that optimal performance is not acquired by experience alone but requires higher levels of direct feedback [12]. We propose that such feedback could be achieved efficiently in experimental settings like those applied in our study. We believe that data sets like that of the present study should play a key role in the training of residents because they can learn and demonstrate improving skills at any time during their training. However, for example, in the Dutch setting, this requires that residents should spend more time in such skill labs and less in daily clinical production.

Conclusion

Emerging alternatives to invasively stage the mediastinum in NSCLC puts high levels of skill to interpret PET and CT scans in NSCLC patients. Observer variation of PET in mediastinal staging appears to be similar to CT reading, as reported in literature, with obvious room for improvement. Training of imaging specialists may require higher levels of feedback, which can more efficiently be obtained in skill labs using existing databases than are currently achievable in local daily clinical practice.

Acknowledgements. Financial support was provided by GlaxoSmith Kline Beecham, the Netherlands. The authors would like to thank all participating nuclear medicine physicians for their effort and enthusiasm.

The Study Group of Clinical PET consists of Geesje J. Abels-Fransen, Jim Baas, Peter Barneveld, Robbert Boer, Vivian Bongers, Filiz Celik, Roel Claessens, Michela A. Edelbroek, Dyde A. Huysmans, Ing Han Liem, Peter van Noorden, Imad al Younis, Dolores Zanin, and Ton Zwijnenburg

References

1. Smulders SA, Smeenk FW, Janssen-Heijnen ML, Wielders PL, de Munck DR, Postmus PE (2005) Surgical mediastinal staging in daily practice. Lung Cancer 47:243–251
2. Stroobants S, Verschakelen J, Vansteenkiste J (2003) Value of FDG-PET in the management of non-small cell lung cancer. Eur J Radiol 45:49–59
3. Kernstine KH (2003) Positron emission tomography with 2-[18f]fluoro-2-deoxy-D-glucose: can it be used to accurately stage the mediastinum in non-small cell lung cancer as an alternative to mediastinoscopy? J Thorac Cardiovasc Surg 126:1700–1703
4. Gould MK, Maclean CC, Kuschner WG, Rydzak CE, Owens DK (2001) Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. JAMA 285:914–924
5. Joshi U, Hoekstra OS, Boellaard R, Comans EF, Raijmakers PG, Pijpers RJ, et al. (2004) Initial experience with a prototype dual-crystal (LSO/NaI) dual-head coincidence camera in oncology. Eur J Nucl Med Mol Imaging 31:596–598
6. van Tinteren H, Hoekstra OS, Smit EF, van den Bergh JH, Schreurs AJ, Stallaert RA, et al. (2002) Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: the PLUS multicentre randomised trial. Lancet 359:1388–1393
7. Herder GJ, van Tinteren H, Comans EF, Hoekstra OS, Teule GJ, Postmus PE, et al. (2003) Prospective use of serial questionnaires to evaluate the therapeutic efficacy of 18F-fluorodeoxyglucose (FDG) positron emission tomography (PET) in suspected lung cancer. Thorax 58:47–51
8. Hoekstra CJ, Stroobants SG, Hoekstra OS, Vansteenkiste J, Biesma B, Schramel FJ, et al. (2003) The value of [18F]fluoro-2-deoxy-D-glucose positron emission tomography in the selection of patients with stage IIIA-N2 non-small cell lung cancer for combined modality treatment. Lung Cancer 39:151–157
9. Mountain C, Dresler C (1997) Regional lymph node classification for lung cancer staging. Chest 111:1718–1723
10. Robinson PJ (1997) Radiology's Achilles' heel: error and variation in the interpretation of the Rontgen image. Br J Radiol 70:1085–1098
11. Guyatt GH, Lefcoe M, Walter S, Cook D, Troyan S, Griffith L, et al. (1995) Interobserver variation in the computed tomographic evaluation of mediastinal lymph node size in patients with potentially resectable lung cancer. Canadian Lung Oncology Group. Chest 107:116–119
12. Brehmer B (1980) In one word: not from experience. Acta Psychol (Amst) 45:223–241