



EU–USA pathology panel for uniform diagnosis in randomised controlled trials for HRCT screening in lung cancer

F.B. Thunnissen*, K.M. Kerr[#], E. Brambilla[†], C.E. Comin⁺, W.A. Franklin[§],
B. Guldhammerskov[‡], W.H. Westra^{**} and D.B. Flieder^{##}

ABSTRACT: Randomised controlled trials for lung cancer screening using high-resolution computed tomography are now underway. In order to allow effective future comparison of the different trials, as well as strengthening conclusions based upon the analysis of larger data sets, uniformity and consistency of pathology diagnosis are essential. The aim of the present study was to determine the effectiveness of the learning process in this difficult area of diagnostic pathology.

Eight pathologists received two CD-ROMs, each with digital images of 30 cases. After diagnosing the first series, selected background reading was provided. Kappa (κ) scores were calculated for each pathologist and category, and were compared to the consensus score.

The readings of the first series showed a moderate agreement κ score: mean \pm SD for category numbers 8 (all eight categories) and 2 were 0.53 ± 0.05 and 0.65 ± 0.04 , respectively. The κ 2 score distinguished between categories denoting benign and malignant lesions. The second series resulted in a good agreement κ score: 0.65 ± 0.06 for category number 8 and 0.81 ± 0.02 for category number 2.

In conclusion, this study demonstrates that screen-detected cases pose particular problems for pathologists and that a trained pathology panel serving randomised controlled trials is likely to lead to more consistent and accurate tissue diagnosis.

KEYWORDS: Diagnosis, lung cancer, panel, pathology, screening

Randomised controlled trials (RCTs) using high-resolution computed tomography (HRCT) to screen for lung cancer are underway in the USA, the Netherlands, Denmark and Italy and a trial may soon start in Germany. While many of the administrative, epidemiological and radiological difficulties have been addressed and solved, little has been written about pathological diagnoses of resected lesions. Most of the programmes have reached their enrolment goals, and these expensive and time-consuming exercises must be supported by accurate pathological data. Thus, it is necessary for the pathological diagnoses made in the different trials to follow the accepted World Health Organization (WHO) classification scheme. If pathology diagnoses are uniform across the different trials, then comparing and combining data will be possible. Screening for lung cancer using low-density HRCT detects small parenchymal lesions, with a bias towards peripheral adenocarcinomas.

Not only is there uncertainty about the biology of these small adenocarcinomas, but also a lack of diagnostic uniformity among surgical pathologists. In preparation for the RCT screening trials, a National Cancer Institute-sponsored protocol for pathological handling of screen-detected lesions was formulated [1]. To guide the process of uniform pathology diagnosis, a panel consisting of pathologists involved in the RCT trials and other experts was established. The pathology panel studied 60 difficult cases in order to standardise their diagnostic criteria within the 2004 WHO classification of lung tumours. The results of this undertaking are discussed herein.

MATERIAL AND METHODS

The eight panel pathologists first received digital images of 30 lesions likely to be encountered in the screening process. The cases represented a range of small glandular and other lesions encountered in the periphery of lung resection specimens. To achieve an effective learning

AFFILIATIONS

*Dept of Pathology, Canisius Wilhelmina Hospital, Nijmegen, The Netherlands.

[#]Dept of Pathology, Aberdeen University School of Medicine, Aberdeen Royal Infirmary, Aberdeen, UK.

[†]Dept of Pathology, INSERM U578, Grenoble, France.

⁺Dept of Human Pathology and Oncology, University of Florence, Florence, Italy.

[‡]Dept of Pathology, KAS Herlev, div. Gentofte, Hellerup, Denmark.

[§]Dept of Pathology, University of Colorado Health Science Centre, Aurora, CO

^{**}Dept of Pathology, The Johns Hopkins Medical Institutions, Baltimore, MD, and

^{##}Fox Chase Cancer Centre, Philadelphia, PA, USA.

CORRESPONDENCE

F.B. Thunnissen
Dept of Pathology
VU Medical Centre
De Boelelaan 1117
1081 HV Amsterdam
The Netherlands
Fax: 31 204444586
E-mail: e.thunnissen@vumc.nl

Received:

March 29 2006

Accepted after revision:

July 31 2006

SUPPORT STATEMENT

The authors would like to acknowledge the support of the National Cancer Institute, USA for the panel workshop in Amsterdam.

European Respiratory Journal
Print ISSN 0903-1936
Online ISSN 1399-3003

phase, cases were selected with a high bias towards difficult histopathological diagnoses. Digital images were selected by D.B. Flieder. Each case consisted of three images: an overview ($2\times$ or $4\times$ original magnification) and two high-magnification detailed images ($20\times$ and $40\times$ original magnification). Panel members were provided with a CD-ROM, with instructions to classify each lesion according to the current WHO classification [2].

Two scores were given: one for the diagnosis and one for the certainty of pathologist's opinion. Since there may have been doubt about the diagnosis, it was qualified by a score of "certain" or probable. If doubt existed between diagnostic categories, this case was coded "9". For example, doubt between adenocarcinoma and nonmucinous bronchioloalveolar carcinoma was given as category number 9 and then qualified by the categories in the differential diagnosis (table 1). Any doubt implies that, in the clinical setting, histopathological consultation is preferred before starting treatment. All panel pathologists were advised beforehand that they should approach each case such that a certain diagnosis would lead to definitive treatment, while a case with a probable diagnosis would be referred for another opinion. It is noteworthy that the pathologist who provided the cases did not record either category 9 or uncertainty levels for any of the cases.

After reading and submitting diagnoses, panel members received a review article addressing many issues concerning diagnosis of small pulmonary adenocarcinomas [3]. Subsequently, 30 additional cases were similarly distributed, studied and diagnosed by panel members. A group meeting in Amsterdam (the Netherlands) ensued in which findings were reviewed and discussed.

Data handling

Some of the scores were received electronically and some on paper; all were entered into the database and double-checked for errors.

For any case given a score of 9, the alternative diagnoses were compared with the consensus opinion and if the latter was one of the alternatives, this was considered as agreement,

otherwise the score of 9 was retained. The kappa (κ) 2 score established the agreement between benign and malignant diagnoses after combining, into two categories, diagnostic scores 1, 2, 6 and 7 for a generic "malignant" diagnosis, and similarly scores 3, 4, 5 and 8 for a generic "benign" diagnosis. The κ 8 score was calculated using all eight diagnostic categories. κ scores were determined using a standard method [4]. All the κ scores were calculated for all panel members through comparison with the final consensus diagnosis.

For each panel member, a calculation was made of the number of diagnoses that agreed with the final consensus diagnosis expressed as a percentage.

RESULTS

In order to test the panel in a realistic way, a set of particularly difficult and potentially controversial cases were selected; cases which would be unusual in standard pulmonary surgical pathology practice. During a consensus meeting the 60 cases were reviewed. Consensus diagnosis is shown in table 2.

The readings of the first series resulted in a moderate κ -score: the mean \pm SD for categories 8 and 2 were 0.53 ± 0.05 and 0.65 ± 0.04 , respectively. The κ 2 score distinguished between lumped categories, denoting benign and malignant lesions (fig. 1). For the second series of readings a higher κ score was obtained: 0.65 ± 0.06 for category 8 and 0.81 ± 0.02 for category 2.

In many cases, agreement by all pathologists was not reached. In 30 cases at least five panel members were in agreement; in 26 cases at least four of the eight members agreed the diagnosis. For the remaining four cases, two diagnoses were each favoured by three panel members. In order to facilitate percentage agreement calculations, these latter two cases were excluded.

Some cases existed in which one or more members' diagnosis deviated from the consensus. Each pathologist offered at least one deviant diagnosis. In four cases, only three of the eight panel members initially gave the final consensus diagnosis, emphasising the difficult nature of the material under review. In two instances the differential diagnosis was between malignant bronchioloalveolar carcinoma (BAC) *versus* benign

TABLE 1 Options for histological classification

Diagnostic categories	Category number	Confidence level
Adenocarcinoma (with obvious or focal invasion)	1	1 or 3
Nonmucinous bronchioloalveolar carcinoma	2	1 or 3
Atypical adenomatous hyperplasia	3	1 or 3
Scar lesion or reactive atypia	4	1 or 3
Bronchialisation of alveolar septa (Lambertosis)	5	1 or 3
Small cell carcinoma	6	1 or 3
Squamous cell carcinoma or large cell carcinoma	7	1 or 3
Other, please specify e.g. papillary adenoma	8	1 or 3
Doubt between one or more categories (specify)	9	Stipulate categories considered e.g. category 1 and 2

Confidence level 1: certain; 3: probable.

TABLE 2 Consensus diagnosis (cases with at least four equal scores)

Consensus diagnosis	Score	Agreement	
		Cases 1–30	Cases 31–60
Adenocarcinoma mixed subtype	1	83 (9)	84 (8)
BAC	2	62 (5)	81 (6)
AAH	3	73 (7)	76 (4)
Reactive scar	4	67 (3)	67 (6)
Bronchialisation-lambertosis	5	100 (1)	
Meningothelial nodule/tumour let	8	(1)	
<4 equal scores		(4)	(2)

Data are presented as % (n). BAC: bronchioloalveolar carcinoma; AAH: atypical adenomatous hyperplasia.

atypical adenomatous hyperplasia (AAH); between adenocarcinoma and BAC and reactive in 2 cases.

In subjective terms the pathologists felt that the second series of 30 cases was more difficult than the first. The total number of instances with doubt between categories (score 9, table 1) was 14 out of 210 (6.7%) and 23 out of 210 (11%), in series one and two, respectively. The number of diagnoses in which the confidence level was low (confidence “probable”, table 1 and includes all category 9 diagnoses) was 36.7% and 38%, in series one and two, respectively.

DISCUSSION

Since RCTs of low-dose HRCT screening for early detection of lung cancer are so expensive and several European countries are undertaking studies with relatively small numbers of cases in the screen and control arms, the individual studies are unlikely to be large enough to result in significant power for reliable evaluation of the outcome. Therefore, combining the

data from those studies with a similar study design will be essential. Consistent high-quality pathology data are thus remarkably important for the success of the controlled trials.

The assembled EU–USA pathology panel consists of both recognised “experts” and of responsible pathologists from the individual RCT trials. The panel members read the cases beforehand, and met with the sole purpose of improving their understanding and ability to apply diagnostic criteria to lesions according to the WHO classification [2]. The panel had intense exposure to the type of difficult lesions encountered in screening.

The cases selected comprised mainly reactive bronchioloalveolar proliferative lesions, atypical adenomatous hyperplasia, nonmucinous bronchioloalveolar carcinoma and adenocarcinoma mixed subtype. The κ score depends on the number of classes used for categorisation, on the number of cases examined and their degree of difficulty. In general, κ scores between 0.4 and 0.7 are considered acceptable and higher scores are more difficult to reach. In the two-way classification where the categories were combined to either benign or malignant, an average κ score of 0.81 was reached, demonstrating the possibility of achieving a highly consistent diagnosis at this less specific level. In terms of full agreement, individual pathologists occasionally differed from the consensus, even after discussion.

As in most instances of knowledge acquisition, a clear learning phase was established for diagnoses of early lung cancer. It may be useful, possibly with the help of telepathology, to use these cases as an educational tool, as well as for pathologists not directly involved in such trials and less exposed to this difficult area of diagnosis.

The confidence of the pathologists, *i.e.* whether they were certain or uncertain in their proffered diagnosis, was similar in both series. In 36–38% of the cases there was uncertainty. In general, adequate personal confidence in combination with appropriate referral/consultation with other suitably trained pathologists is highly desirable for adequate diagnosis in surgical pathology, but is not always easy to achieve [5]. This is another argument in favour of the panel approach to diagnosis.

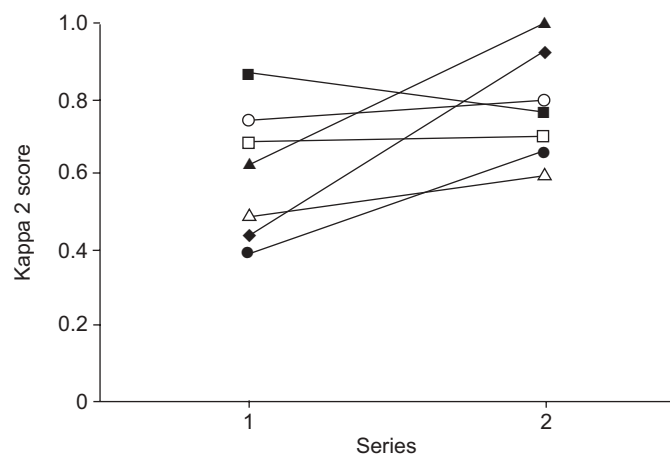


FIGURE 1. Kappa 2 values for the different pathologists for the first and the second series.

The coincidence rates for agreement in this study are similar to those reported in a recent Japanese study [6]. In addition, the authors of that study found an improvement in diagnostic accuracy and consensus through a learning phase [6]. However, the selection of cases in that study was based on small adenocarcinomas and bronchioloalveolar carcinomas only and used a consensus of three pathologists as standard. In the current study, consensus was not reached on all cases, even after in-depth discussion during the review process. Furthermore, the present study's definition of “consensus” was more demanding than in the Japanese study. Importantly, the difficulty in diagnosing these lesions is thus far considered as a two-tier system, either benign or malignant. However, the biological lesions under study may well represent a continuum, in which the growth rate [7] may increase in time for AAH, BAC and invasive adenocarcinoma with BAC features [5, 8, 9]. Making distinct categories in a continuous spectrum will inherently cause some variation around the borders of these categories. These variations are more likely to be

controlled during discussion amongst the panel members, rather than from individual judgements alone.

Results from the pathological review of resected carcinomas obtained through the Early Lung Cancer Action Program (ELCAP), further highlight difficulties in distinguishing reactive processes and AAH from BAC, and BAC from invasive adenocarcinoma, mixed subtype, and support the use of a panel review of trial patients [10]. Of 92 lesions, one case submitted as parenchymal scar with reactive atypia was reclassified as BAC, six cases submitted as AAH were reclassified as either BAC or invasive adenocarcinoma, mixed subtype and 18 cases with submitted diagnoses of BAC were reclassified as invasive adenocarcinomas, mixed subtype.

In conclusion, this study affirms the hypothesis that many of those small glandular proliferations likely to be encountered in low-density high-resolution computed tomography screening programmes are diagnostically difficult, even for pulmonary pathologists and that certainty about diagnosis may not always be reached. These findings emphasise the importance of establishing a central pathology review of all screen-detected resection specimens.

REFERENCES

- 1 International Association for the study of Lung Cancer. www.iaslc.org/EuUs.asp. Date last updated: August 1, 2005.
- 2 Travis WD, Brambilla E, Muller-Mermelink HK, Harris CW, eds. WHO classification; Pathology and Genetics: Tumors of the Lung, Pleura, Thymus and Heart. Geneva, IARC Press, 2004.
- 3 Flieder DB. Screen-detected adenocarcinoma of the lung. Practical points for surgical pathologists. *Am J Clin Pathol* 2003; 119: Suppl. S39–S57.
- 4 Silcocks P. Measuring repeatability and validity of histological diagnosis - a brief review with some practical examples. *J Clin Pathol* 1983; 36: 1269–1275.
- 5 Travis WD, Garg K, Franklin WA, et al. Evolving concepts in the pathology and computed tomography imaging of lung adenocarcinoma and bronchioloalveolar carcinoma. *J Clin Oncol* 2005; 23: 3279–3287.
- 6 Noguchi M, Minami Y, Iijima T, Matsuno Y. Reproducibility of the diagnosis of small adenocarcinoma of the lung and usefulness of an educational program for the diagnostic criteria. *Pathol Int* 2005; 55: 8–13.
- 7 Thunnissen FB, Schuurbiens OC, den Bakker MA. A critical appraisal of prognostic and predictive factors for common lung cancers. *Histopathology* 2006; 48: 779–786.
- 8 Lopez JI, Colby TV, Gazdar AF. Current status of small peripheral adenocarcinomas of the lung and their importance to pathologists. *Ann Diagn Pathol* 2005; 9: 115–122.
- 9 Noguchi M, Morikawa A, Kawasaki M, et al. Small adenocarcinoma of the lung. Histologic characteristics and prognosis. *Cancer* 1995; 75: 2844–2852.
- 10 Flieder DB, Vazquez M, Carter D, et al. Pathologic findings of lung tumors diagnosed on baseline CT screening. *Am J Surg Pathol* 2006; 30: 606–613.