Eur Respir J 1999; 13: 781–786
Printed in UK – all rights reserved

European Respiratory Journal
ISSN 0903-1936

Subjective differentiation of normal and pathological bronchi on thin-section CT: impact of observer training

A.A. Bankier*, D. Fleischmann*, V. De Maertelaer**, M. Kontrus*, T. Zontsich*, K. Hittmair*, R. Mallek*

Subjective differentiation of normal and pathological bronchi on thin-section CT: impact of observer training. A.A. Bankier, D. Fleischmann, V. De Maertelaer, M. Kontrus, T. Zontsich, K. Hittmair, R. Mallek. ©ERS Journals Ltd 1999.

ABSTRACT: The effect of observer training on sensitivity, specificity and interobserver agreement in the differentiation between normal and pathological bronchi on computed tomography (CT) was studied.

The wall thickness of bronchi with normal walls and with pathologically thickened walls were subjectively scored by three independent observers before and after a training period of 2 weeks. Sensitivity, specificity and interobserver agreement were calculated for reading sessions before and after training. Increase and decrease in agreement after training were determined.

There was a statistically significant difference (p=0.001) between objectively measured wall thickness of normal and pathological bronchi, both for reference bronchi and for bronchi used for reading sessions. While training increased interobserver agreement, it had no effect on sensitivity (0.46 *versus* 0.44 after training) and specificity (0.71 *versus* 0.72 after training) in detecting pathological bronchi. Increased agreement after training was significantly (p=0.001) more frequent than decreased agreement.

There is a discrepancy between the effect of training on interobserver agreement and on sensitivity and specificity in the subjective differentiation between normal and pathological bronchi. Interobserver agreement alone is not a reliable indicator of a beneficial effect of training in the evaluation of this parameter.

Eur Respir J 1999; 13: 781–786.

*Dept of Radiology, University of Vienna, Austria and **Institute for Interdisciplinary Research, Faculty of Medicine, Free University of Brussels, Belgium.

Correspondence: A.A. Bankier
Dept of Radiology
University of Vienna
Waehringer Guertel 18–20
A-1090 Vienna, Austria
Fax: 43 1404004894

Keywords: Bronchi
observer performance
thin section computed tomography

Received: April 24 1998
Accepted after revision December 8 1998

Supported in part by the Ludwig Boltzmann Institute for Physical and Radiological Tumour Diagnosis.

Bronchial wall thickening has been considered clinically relevant in a number of thin-section computed tomography (CT) studies [1–10]. Because there are currently no established objective criteria for the characterization of this parameter [3, 6, 8], its assessment relies on the judgement of the individual reader. The resultant subjectivity has been considered a major limitation in the assessment of bronchial wall thickness [10, 11]. Although preliminary guidelines for the appropriate evaluation of bronchial wall thickness were elaborated to attenuate the subjectivity of the individual reader [10, 12, 13], these experimentally determined guidelines have not yet been fully implemented and validated in routine imaging. Inadequate evaluation of bronchial wall thickness, may, however, have potentially serious clinical implications, notably when correlated with the results of functional examinations [10]. Because of these implications, there is a resurgence of interest in the qualitative outcome of subjective assessment of bronchial wall thickness, particularly in terms of its accuracy [7, 14].

Two recent thin-section CT studies [7, 14] found reasonably good interobserver agreement when bronchial wall thickness was subjectively evaluated in patients with asthma and bronchiectasis. Both studies, however, focused on patients with clinically and/or radiologically confirmed disease, and the relevant distinction between normal and pathological bronchi was only marginally addressed. Also, one of the authors [14] speculated that training of observers would increase interobserver agreement in the subjective assessment of bronchial wall thickness. Therefore, the aim of the present study was to analyse the diagnostic potential of the subjective evaluation of bronchial wall thickness on thin-section CT in the differentiation between normal and pathological bronchi, and to test whether observer training has a beneficial effect on sensitivity, specificity, and interobserver agreement in the evaluation of this parameter.

Materials and methods

Bronchi

The selection of bronchi used in this study was coordinated by a radiologist who was not involved in the later film reading. The selection of bronchi was based on: 1) histopathological work-up, and 2) absolute objective measurements of bronchial wall thickness on CT. For the reading sessions, 20 normal (10 segmental and 10 subsegmental bronchi) and 20 pathological bronchi (10 segmental and 10 subsegmental bronchi) were used. Apart from those bronchi used for the reading sessions, 23 reference bronchi were used for the consensus and training sessions. All bronchi were required to be of uniform wall

thickness over the visible part of their circumference. To avoid misleading enlargement of the bronchial walls, the CT scan plane was oriented as perpendicularly as possible to the bronchi [12]. Also, the chosen bronchi had continuous external and internal perimeters [15]. The selection of bronchi in the various subgroups described below was mainly determined by the clinical settings and the general patient selection in the hospital. Therefore, very severely thickened bronchial walls, for example, as seen in patients with advanced bronchiectasis or cystic fibrosis, were not included in the study.

Histopathological work-up reported the bronchi to be of normal wall thickness or to have pathological wall thickening. Objective measurements of bronchial wall thickness were performed by two independent observers who were not involved in the later film reading. Measurements were performed on the CT console after postprocessing of the original images. The display window settings for postprocessing were identical to those used for the photographic images, as described later in the *Computed tomography* section. Bronchial wall thickness was defined as the distance between the outer and the inner border of the bronchial wall, as measured with a calliper. Three calliper measurements were performed by each of the two observers, and the mean of these measurements used for later analysis. The calliper technique used for these measurements has been described in previous studies [10, 11, 13] and was found to be adequate for the objective evaluation of bronchial wall thickness.

Bronchi used for film reading. Normal bronchi. The 20 normal bronchi were selected on the basis of the histopathological work-up of transbronchial biopsies performed in 14 nonsmoking individuals screened for suspected interstitial lung disease. All bronchial biopsies included representative segments of the bronchial wall that would warrant a reliable diagnosis. Bronchial wall thickness was reported normal in all 14 individuals. The chosen CT slices were matched with the site of the biopsy using anatomical landmarks such as vascular or bronchial structures and their respective segmentation. Objective measurements were performed as described above.

Pathological bronchi. The 20 pathological bronchi were selected on the basis of the histopathological work-up of postoperative specimens. The specimens were resected in 12 patients with severe emphysema who had undergone lung volume reduction surgery. Bronchial walls were reported pathologically thickened owing to chronic inflammatory infiltrates in all 12 patients. The chosen CT slices were matched with the levels of the histopathological cuts by use of anatomical landmarks. CT slices were also chosen such that bronchi were surrounded by relatively undestroyed areas of lung parenchyma. Objective measurements were performed as described above.

For both groups of bronchi used for later film reading, an equal number of bronchi were chosen from the upper (n=7) and lower lobes (n=13).

Reference bronchi used for consensus sessions and training. The reference bronchi used for training and the consensus session were selected on the basis of the histopathological work-up of postoperative specimens. The specimens were obtained from 23 patients after volume reduction surgery (pathological bronchial wall thickening; n=14) or because of solitary pulmonary nodules (normal bronchial wall; n=9). CT slices were matched with the levels of the histopathological cuts by use of anatomical landmarks. Objective measurements were performed as described above. None of the bronchi used for the consensus session or training were included in the later film reading.

Computed tomography

For both image acquisition and postprocessing of the data, a Tomoscan SR 7000 (Philips Medical Systems, Best, The Netherlands) was used. Initial thin-section CT examinations were performed at maximal inspiration with the subject in the supine position, with a slice thickness of 1.5 mm and with section intervals of 10 mm, from the lung apices to the lung bases, at 140 kV and 175 mA, using a 512×512 matrix. A high spatial frequency algorithm was applied for image reconstruction.

Raw data from all CT images were postprocessed at the console. All images were reconstructed in a field of view of 15 cm and at a magnification of 1.5 factor, with the selected bronchus in the centre of the image. One single bronchus was identified for evaluation on each image. The bronchi used for the film reading sessions were photographed, with each of the 40 chosen bronchi being imaged 11 times at window settings within the range recommended in the literature for appropriate imaging of bronchial wall thickness (window centre, range -350–-550 Hounsfield units (HU); window width, range 1200–1500 HU) [10, 12, 13, 15]. This multiple imaging prevented easy recognition of the bronchi in the second reading session. The potential influence of multiple imaging on the statistical analysis is discussed below. Reference bronchi used for consensus and training sessions were photographed at the same display window settings.

The above mentioned protocol produced a total of 440 thin-section CT images for use in the reading sessions (220 images of normal bronchi (110 segmental, 110 subsegmental) and 220 images of pathological bronchi (110 segmental, 110 subsegmental)). All images obtained after postprocessing of the data were photographed, 12 per film (35.6×43.2 cm), twice in random order for the two readings before and after training. Display window settings and patient names were obscured on the images obtained.

Film reading

All three readers involved in this study were experienced chest radiologists, familiar with reporting in the same manner as part of their daily routine, and with the same level of experience. None of the readers were informed about the selection criteria of the bronchi chosen for the reading sessions.

Scores. For all reading sessions, bronchial wall thickness was subjectively scored according to a three-point scale (3 for bronchi with obviously thickened walls, 2 for bronchi with normal wall thickness and 1 for bronchi with walls appearing thinner than normal). The score of

1 was introduced because, according to the authors' experience, walls of normal bronchi may sometimes appear thinner than expected on thin-section CT scans.

Consensus session. As recommended in the literature [16–19], a consensus session was performed before the first reading session. The aim of this session was to familiarize the readers with the scoring system to be applied, and to compare the scores with the gold standard as defined for this investigation (histopathological analysis and objective CT measurements). During the consensus session, images of the reference bronchi were evaluated twice. First, all reference images were read without knowledge of the results of the histopathological work-up and objective measurements. Then the results of histopathological work-up and objective measurements were revealed to the readers, and the images reread. During both readings special attention was given to the differentiation between normal and pathological bronchi compared to the defined gold standard.

Reading sessions. The definitive set of images was read independently by the three radiologists in six separate sessions (three before training, three after training) who classified the bronchi according to the three-point scale described above. The three reading sessions before training took place on the day following the consensus session, and the three reading sessions after training took place 2 weeks after the first reading session. For the second reading session, the readers were not informed about the results of the first reading session.

Training

The training sessions were performed at the end of each working day for the 2 weeks between the first and second reading (n=10). During these sessions, all routine thin-section CT examinations performed at the authors' institution on the respective day were reread by the three observers involved, and under the supervision of a fourth radiologist who had been in charge of bronchus selection and image postprocessing. By the end of day 10, a series of 38 thin-section CT examinations had been reread. The training sessions focused on the differentiation between normal and thickened bronchial walls. The bronchi used during the consensus session therefore served as references. Each segmental and subsegmental bronchus depicted by one of the 38 thin-section CT examinations was categorized as either normal or pathological and scored according to the three-point scale by each of the three readers. In cases of divergent results, the subjective criteria of each reader for categorizing a bronchus as either normal or pathological were discussed until unanimity was reached.

Statistical analysis

For both the reference bronchi and the bronchi used for training and the consensus session, the means and standard deviations of the objective measurements were calculated. The statistical significance of differences between the groups of normal bronchi, between normal and pathological bronchi, between segmental and subsegmental bronchi and between observers were studied using analyses of variance (ANOVA) with two repeated measures between factors, each at two levels (normal/pathological; segmental/subsegmental), and with one repeated measure (observer) to assess interobserver variability.

For the subjective differentiation between normal and pathological bronchi, sensitivity and specificity were calculated according to the guidelines of FISHER and BELLE [20]. Scores 1 and 2 (not thickened) were grouped and compared to score 3 (pathologically thickened). Subjective scores were then compared to the combined findings of the histopathological work-up and objective measurements.

The effect of training was evaluated as follows. Firstly, the agreement between observers was compared before and after training. This was performed by comparing the Cohen's kappa coefficients [21–22] calculated for each pair of observers before and after training. Each kappa coefficient was based on the 440 CT images obtained from the 11 slices of the 40 bronchi. If a bias on the kappa coefficient resulted from the fact that sets of 11 slices came from the same bronchus, this bias was similar both before and after training, leaving the comparison still valid [21]. Secondly, based on the results of the evaluations before and after training, five categories of images were determined: 1) total agreement between the three evaluations before and after training; 2) one observer disagreed before training and the three observers agreed after training; 3) the three observers agreed before training and one observer disagreed after training; 4) one disagreement before training and one disagreement after training; and 5) other situations.

The images frequency categororized as above was calculated for both the overall and predefined subgroups of bronchi. Comparisons between the scores before and after training were performed using the McNemar test [20]. Calculations were performed using the SPSS statistical software [23], with statistical significance set at the $p<0.05$ level.

Results

Results of the objective measurements of bronchial wall thickness are shown in table 1. For the objective measurements used for the consensus and the training sessions, the mean bronchial wall thickness for pathological

Table 1. – Results of objective measurements of bronchial wall thickness

	n	Bronchial wall thickness mm
Bronchi used for training and consensus session		
Normal segmental bronchi	5	0.86±0.04
Normal subsegmental bronchi	4	0.47±0.03
Pathological segmental bronchi	7	1.87±0.04
Pathological subsegmental bronchi	7	0.95±0.03
Bronchi used for reading sessions		
Normal segmental bronchi	10	1.14±0.03
Normal subsegmental bronchi	10	0.46±0.03
Pathological segmental bronchi	10	1.77±0.05
Pathological subsegmental bronchi	10	0.95±0.04

Data are presented as mean±SD.

bronchi was significantly higher than for normal bronchi (p=0.001), both for segmental and subsegmental bronchi. No statistical significance (p=0.47) was observed between the normal bronchi used for training/consensus sessions, or between those used in the reading sessions. Also, there was no statistically significant difference (p=0.603) between the objective measurements of the two observers. The same conclusions were reached for the bronchi used for the reading sessions (p=0.001), with no statistically significant difference (p=0.94) between the measurements of the two observers.

The sensitivity and specificity for the subjective differentiation between normal and pathological bronchi before and after training are shown in table 2. For both the overall assessment and the detailed assessment of segmental and subsegmental bronchi, the differences between the two reading sessions were negligible for both sensitivity and specificity.

Table 3 displays the Cohen′s kappa coefficients between the observers compared pairwise, before and after training. An obvious increase in the kappa coefficient is seen after training (kappa=0: total independence, kappa=1: perfect agreement between the observers).

Table 4 shows the absolute numbers and the corresponding percentages of the images classed according to the categories defined to evaluate the effect of training. The discrepancy between the scores before and after training never exceeded 1 and the fifth category was empty. Among the images that did not show total agreement before and after training, there was no statistically significant difference between the percentage of normal (63/109) and pathological (81/126) bronchi (χ^2= 1.04, p=0.348) for which an increased observer agreement was observed after training. Also, among the images which did not show perfect agreement before and after training, there was no statistically significant difference between the percentage of segmental (71/121) and subsegmental (73/114) bronchi (χ^2=0.71, p=0.424). The percentage of images with increased agreement after training was significantly superior to those with decreased agreement in each of the four predefined categories of images (p=0.001).

Discussion

The introduction of thin-section CT has substantially increased the number of detectable pathologies in the lung, and, consequently, has considerably widened the diagnostic spectrum of thoracic imaging. There is, however, little evidence to suggest that improved technology alone reduces the diagnostic error rates of individual readers within the expanding group of potentially detectable lesions [24]. This is amplified by the fact that on thin-section CT a series of findings is assessed on a subjective basis, owing to the lack of objective criteria for the differentiation between normal and pathological findings [1–6, 8, 25]. The diagnosis of these findings may, therefore, be hampered in terms of sensitivity, specificity and accuracy [14]. Bronchial wall thickness has been given particular attention with regard to its diagnostic implications, with a concomitant elaboration of the technical parameters necessary for the appropriate assessment of its meaning for diagnostic purposes [10, 12, 13, 15]. In addition to optimizing technical parameters, observer training has been suggested as another effective method of improving the outcome of the subjective evaluation of bronchial wall thickness on thin-section CT [14]. The present study, designed to test this hypothesis, could only partly confirm the beneficial influence of observer training on the assessment of bronchial wall thickness. While training substantially increased agreement between observers, it had virtually no effect on sensitivity and specificity in the subjective differentiation between normal and pathological bronchi.

The positive impact of training on the agreement between the observers in the present study was obvious. The interobserver agreement seen after training was comparable to data reported in previous investigations [7, 14]. The comparatively lower rates for interobserver agreement before training were probably attributable to the selection of bronchi for analysis. While prior studies [7, 14] assessed bronchial wall thickness in individuals with known disease who, consequently, had an increased likelihood of pathological wall thickening, the bronchi analysed in the present study were either normal or pathological, and the readers were kept unaware of this fact. Although this lack of bias resulted in decreased interobserver agreement before training, the increase of agreement after training suggested a beneficial effect of training leading to a substantial overall improvement in observer agreement.

Unfortunately, this positive effect of training on interobserver agreement was attenuated by its poor effect on both sensitivity and specificity in the differentiation between normal and pathological bronchi. The results revealed that, for all predefined categories of bronchi, sensitivity and specificity remained virtually unchanged before and after training. This dissociation between the effect of training on interobserver agreement and on sensitivity and specificity indicates that the apparently positive influence of training on interobserver agreement should be interpreted with caution. In light of the virtually unmodified rates for sensitivity and specificity before and

Table 2. – Sensitivity and specificity in the subjective differentiation of normal and pathological bronchi before and after observer training

	First reading (before training)		Second reading (after training)	
	Sensitivity	Specificity	Sensitivity	Specificity
Overall	0.46	0.71	0.44	0.72
Segmental bronchi	0.46	0.65	0.45	0.65
Subsegmental bronchi	0.44	0.75	0.45	0.78

The overall rates and the rates obtained for segmental and subsegmental bronchi are shown.

Table 3. – Cohen′s kappa coefficients with their asymptotic standard error in a pairwise comparison.

kappa±SEM	Observers 1&2	Observers 1&3	Observers 2&3
Before training	0.317±0.044	0.358±0.045	0.178±0.045
After training	0.698±0.033	0.731±0.033	0.717±0.033

Table 4. – Interobserver agreement and disagreement before and after training

	Total agreement before and after training	One disagrees before training, 3 agree after training	3 agree before training 1disagrees after training	No increase or decrease of agreement after training
Overall	205 (46%)	144 (33%)	5 (1%)	86 (20%)
Normal bronchi	111 (50%)	63 (28%)	4 (3%)	42 (19%)
Pathological bronchi	94 (43%)	81 (37%)	1 (0%)	44 (20%)
Segmental bronchi	99 (45%)	71 (33%)	3 (1%)	47 (21%)
Subsegmental bronchi	106 (48%)	73 (33%)	2 (1%)	39 (18%)

Both the overall rates and rates obtained for the defined subcategories of results obtained from the 440 CT images are shown, as number of images with percentage of defined subcategory in parentheses.

after training, the improved interobserver agreement after training reflects only the fact that reading errors after training appeared more systematically than before training, rather than indicating an overall decrease in reading errors after training. Thus, the results show that good interobserver agreement alone is not a reliable indicator of the validity of a predefined diagnostic approach [7, 14, 26].

This ambiguous influence of observer training on sensitivity, specificity and interobserver agreement in the subjective assessment of a predefined parameter, as investigated, has been widely discussed [16–19, 24, 27–30]. While some studies [16, 27, 28, 30] reported that observer training had a positive impact on their results throughout, others [19, 29] could not confirm any impact of training on their results. A third group of studies [17, 18] even found that observer training decreased the quality of their results. Those divergent conclusions might be attributable to the variety of cognitive tasks that were studied in each of the investigations [16–19, 24, 27–30]. As there is no generally approved protocol for the performance of observer training, individual training protocols must be tailored to meet specific demands. Most studies, however, agree that, independently of these specific demands, any basic protocol should include the supervised and repetitive performance of training skills, accompanied by the verification of the obtained results against a gold standard [17, 18, 24, 27, 28]. Although those basic requirements were fulfilled by the present study design, factors such as the period of time between the readings before and after training, the group of readers, the particular setting of the readings and the choice of material used for the training may vary from study to study and, as in other investigations, is likely to have influenced the present results.

In this study design, the length of the time period between the readings before and after training was an important consideration. Ideally, this period of time should be as short as possible to minimize any change in the circumstances of the reading sessions. On the other hand, sufficient time should have elapsed to preclude individuals remembering particular images from the reading session before training. In the context of the present study, a gap of 2 weeks was considered to adequately fulfil both requirements. The literature, however, reflects a wide spectrum of different opinions on this issue and proposes time intervals before and after training ranging from several days [27], to several months [16, 30]. Nevertheless, studies with a design comparable to our investigation [17, 28] also used intervals of approximately 2 weeks between the reading sessions. By repetitive analysis of sequential reading results these studies also confirmed that a time interval of 2 weeks had no systematic influence on the qualitative outcome of the reading sessions [28].

As study designs like the present require expert readers, those readers, by definition, cannot be untrained. Also, the consensus session might have been equivalent to a brief-training session. Therefore, the use of a control group of readers who read the images twice without training and/or consensus session could have strengthened the results of the present study. Nevertheless, reader-order effect is presumed to result in an improvement of diagnostic accuracy in the reading after training [17] rather than in a lack of effect on sensitivity and specificity of the reading. Since the latter was observed in the present study, the results may appear biased toward underestimating the true effect of observer training.

Also, the setting used in the present study was not necessarily comparable to diagnostic decision making in daily routine [17, 19, 24, 29]. It is conceivable that, under such specific circumstances, readers are more anxious than usual and/or more concerned about missing or underdiagnosing abnormal cases. This would suggest that, in "test" conditions, participants are more likely to err on the side of caution [24, 29]. The fact that the present results yielded no substantial differences between sensitivity and specificity before and after training nevertheless seems to show that observer training neither amplified nor attenuated the potential tendency of the readers to err on the side of caution. Moreover, the results confirmed that the particular setting of the reading sessions was not an apparent obstacle to an overall improvement in interobserver agreement, even if, as mentioned above, this improvement should be interpreted with care.

In conclusion, in the subjective assessment of bronchial wall thickness on thin-section computed tomography, the effect of observer training is ambiguous. While interobserver agreement was obviously improved after training, training did not substantially affect the ability of observers to discriminate between normal and pathological bronchi. From this perspective, improved observer agreement after training probably reflects a systematization of reading errors as the effect of training rather than a decrease in reading errors after training. This also suggests that improvement of interobserver agreement after training alone does not reliably indicate an overall beneficial effect of training on the subjective evaluation of bronchial wall thickness. Despite the fact that clinical assessment of bronchial wall thickness might remain subjective in the near future, these drawbacks nevertheless emphasize the future need for routinely available and user-friendly computer tools designed to objectively quantify this parameter.

References

1. Hartman TE, Primack SL, Lee KS, Swensen SJ, Müller NL. CT of bronchial and bronchiolar diseases. *Radiographics* 1994; 14: 991–1003.
2. Müller NL, Miller R. Diseases of the bronchioles: CT and histopathologic findings. *Radiology* 1995; 196: 3–12.
3. Naidich DP, Lee JJ, Garay SM, McCauley DI, Aranda CP, Boyd AD. Comparison of CT and fiberoptic bronchoscopy in the evaluation of bronchial disease. *Am J Roentgenol* 1987; 148: 1–7.
4. Naidich DP, McCauley DI, Khouri NF, Stitik FP, Siegelman SS. Computed tomography of bronchiectasis. *J Comput Assist Tomogr* 1982; 6: 437–444.
5. McGuinness G, Naidich DP, Leitman BS, McCauley DI. Bronchiectasis: CT evaluation. *Am J Roentgenol* 1993; 160: 253–259.
6. Grenier P, Maurice F, Musset D, Menu Y, Nahum H. Bronchiectasis: assessment by thin-section CT. *Radiology* 1986; 161: 95–99.
7. Grenier P, Mourey-Gerosa I, Benali K, *et al*. Abnormality of the airways and lung parenchyma in asthmatics: CT observations in 50 patients and inter- and intraobserver variability. *Eur J Radiol* 1996; 6: 199–206.
8. Lenique F, Brauner MW, Grenier P, Battesi JP, Loiseau A, Valeyre D. CT assessment of bronchi in sarcoidosis: endoscopic and pathologic correlations. *Radiology* 1995; 194: 419–423.
9. Rémy-Jardin M, Rémy J, Boulenguez C, Sobaszek A, Edme JL, Furon D. Morphologic effects of cigarette smoking on airways and pulmonary parenchyma in healthy adult volunteers: CT evaluation and correlation with pulmonary function tests. *Radiology* 1993; 186: 107–115.
10. Seneterre E, Paganin F, Bruel JM, Michel FB, Bousquet J. Measurement of the internal size of bronchi using high resolution computed tomography. *Eur Respir J* 1994; 7: 596–600.
11. Desai SR, Wells AU, Cheah FK, Cole PJ, Hansell DM. The reproducibility of bronchial circumference measurement using computed tomography. *Br J Radiol* 1994; 67: 257–262.
12. Webb WR, Gamsu G, Wall SD, Cann CE, Proctor E. CT of a bronchial phantom: factors affecting appearence and size measurements. *Invest Radiol* 1984: 19: 394–398.
13. Bankier AA, Fleischmann D, Mallek R, *et al*. Bronchial wall thickness: appropriate window settings for thin-section CT and radiologic–anatomic correlation. *Radiology* 1996; 199: 831–836.
14. Diederich S, Jurrians E, Flower CDR. Interobserver variation in the diagnosis of bronchiectasis on high-resolution computed tomography. *Eur J Radiol* 1996; 6: 801–806.
15. McNamara AE, Müller NL, Okazawa M, Arntorp J, Wiggs BR, Paré PD. Airway narrowing in excised canine lungs measured by high-resolution computed tomography. *J Appl Physiol* 1992; 73: 307–316.
16. Jones S, Thomas GDH, Williamson P. Observer variation in the assessment of adequacy and neoplasia in cervical pathology. *Acta Cytol* 1996; 40: 226–234.
17. Gur D, Rockette HE, Good WF, *et al*. Effect of observer instruction on ROC study of chest images. *Invest Radiol* 1990; 25: 230–234.
18. Gjorhup T, Kelbaek H, Nielsen D, Kreiner S, Godtfredsen. Interpretation of the electrocardiogram in suspected myocardial infarction: a randomized controlled study of the effect of a training programme to reduce interobserver variation. *J Intern Med* 1992; 231: 407–412.
19. Robbins P, Pinder S, de Klerk N, *et al*. Histological grading of breast carcinomas: a study of interobserver agreement. *Hum Pathol* 1995; 26: 873–879.
20. Fisher LD, van Belle G. Biostatistics. A Methodology for the Health Sciences. New York, Wiley, 1993.
21. Agresti A. An Introduction to Categorical Data Analysis. New York, Wiley, 1996.
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
23. Norusis MJ. SPSS for windows. Chicago, USA, SPSS Inc., 1992.
24. Rackow PL, Spitzer VM, Hendee WR. Detection of low-contrast signals: a comparison of observers with and without radiology training. *Invest Radiol* 1987; 22: 311–314.
25. Grenier P, Cordeau MP, Beigelman C. High-resolution computed tomography of the airways. *J Thorac Imaging* 1993; 8: 213–229.
26. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. *Am J Roentgenol* 1996; 167: 851–854.
27. Zoli M, Merkel C, Sabba C, *et al*. Interobserver and interequipment variability of echo-doppler sonographic evaluation of the superior mesenteric artery. *J Ultrasound Med* 1996; 14: 99–106.
28. Sabba C, Merkel C, Zoli M, *et al*. Interobserver and interequipment variability of echo-doppler examination of the portal vein: effect of a comparative training programme. *Hepatology* 1995; 21: 428–433.
29. Brooks KW, Trueblood JH, Kearfott KJ. Subjective evaluations of mammographic accreditation phantom images by three observers groups. *Invest Radiol* 1994; 29: 42–47.
30. Loughran CF. Reporting of fracture radiographs by radiographers: the impact of a training programme. *Br J Radiol* 1994; 67: 945–950.