The British Journal of Radiology, 77 (2004), 231–235 © 2004 The British Institute of Radiology
DOI: 10.1259/bjr/28883951

Short communication

Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph

D J MANNING, PhD, FInstP, S C ETHELL, BSc and T DONOVAN, MSc

Department of Radiography and Imaging Sciences, St Martin's College, Lancaster LA1 3JD, UK

Abstract. A test bank of verified chest radiographs was compiled for visual search experiments to investigate radiology performance in the detection of early lung cancer. A measure of the physical characteristics of the lesions was derived to determine the conspicuity (χ) of the nodules and to investigate possible causes of failed detection. Observer performance was measured by alternate free response operating characteristic (AFROC) methodology and was supplemented with visual search recording. Correlation of AFROC scores and the χ values was poor but inspection of the visual search recordings showed that most nodules were fixated. Fixations on missed lesions produced average dwell times greater than three times the minimum duration thought to be associated with detection. We conclude that the majority of errors were failures of decision rather than detection and comment on the implications of this for strategies to improve diagnostic effectiveness.

Chest images contain a range of perceptual ambiguities that contribute to a significant error rate in diagnosis [1]. It has been commented that it is not unusual to discover in retrospect significant radiological abnormalities in patients who are later diagnosed with lung cancer [2]. Some confounding features are well understood but a complete picture of the sources and nature of reader error is still to be described. Studies on the detection and recognition of significant lung nodules are a good way of investigating error because the tasks relate to an important pathology and they transfer more generally to a broader class of radiological problems [3]. However, an interesting and unsolved aspect of medical image interpretation is how physical measurements describing the characteristics of images and their diagnostic features relate to observer performance [4, 5]. Medical image interpretation is a noise-limited decision task. To measure the "correctness" of an observer decision we assume a variable that is broadly defined as the "level of apparent abnormality" for a given feature in the image. The variable need not refer to the image itself but to the observer's interpretation of it. So a decision is subject not only to variations in visual features from the image, but also the thresholds of normality held by the observer. However, there is an intuitive assumption that the confidence with which an observer makes a decision will increase as the conspicuity of the target increases. Indeed a great deal of effort in medical imaging is invested in optimizing image quality in respect of contrast, resolution and signal to noise ratio (SNR) — the important contributors to the level of conspicuity of any image feature [5]. We suggest that in complex images, with a search component and multiple targets, the decision task may have too many confounding factors for this assumption to be completely reliable. Even experienced readers may not always register visually obvious lesions.

Received 28 November 2002 and in final form 13 November 2003, accepted 19 November 2003.

Aim

The aim of this work was to determine whether observer error in the detection of early lung cancer from posteroanterior (PA) chest radiographs is due to failure of detection or failure of interpretation. We first developed a measure of conspicuity for the pulmonary nodules to compare with their detection rate in an observer study. We used alternate free response operating characteristic (AFROC) [6] techniques for the observer performance measurements because, unlike conventional ROC methods, this requires the observer to give location information for each target. In large images requiring significant search activity this is an important experimental consideration. Eye tracking during the procedure was used to reveal whether unreported lesions were visually fixated. These are lengthy experimental procedures for both observers and experimenters and we therefore were limited to contributions from seven experienced radiologists in the study and of the seven, four gave additional time for eye-tracking.

Methods

Nodules

120 digital chest images contained 81 pulmonary nodules distributed in a variety of locations in the lung-fields as shown in Figure 1. The nodules were agreed as significant in pathological appearance from confirmed radiological reports. Nodules were roughly circular and ranged in size from 5 mm to 20 mm diameter with a mode value of 10 mm. 65 normal films were included in three test banks of 40 images.

Measurement of SNR/contrast, lesion size and edge gradient

For each nodule, four profiles were taken through its centre and extended for one nodule dimension beyond the opposing edges of the lesion to sample pixel information in

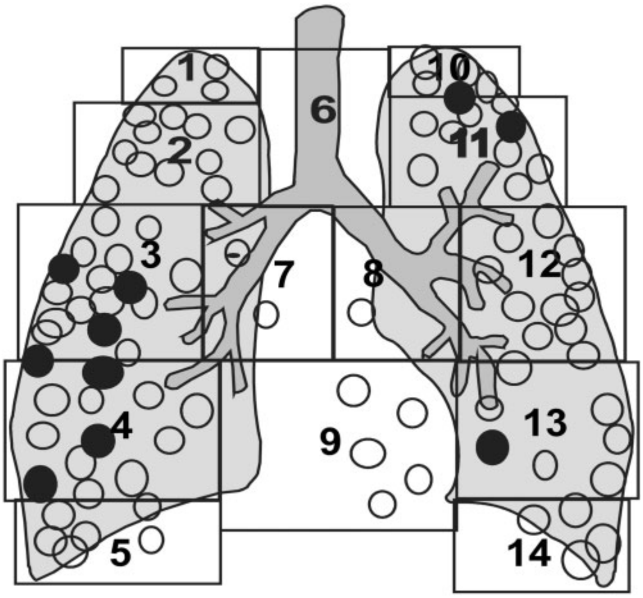

Figure 1. The lesion distribution in the test bank of images. The nodule locations marked ● were those used in the eye-tracking procedure. All locations shown were used for the alternate free response operating characteristic scoring.

its immediate background. An illustration of the arrangement of the profiles is shown in Figure 2.

The edge of a lesion was determined by eye from the image then confirmed by the profiles that allowed an estimate of the percentage reduction in grey level from peak to edge. Figures 3 and 4 show an example of the graphing procedure.

The index of conspicuity for each nodule was calculated from the relationship,

$$\chi = d\ \tan[\theta - 1]\Delta \mathrm{GL}/\sqrt{\sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{n}}^2}$$

where:
θ is the maximum slope angle to the edge of the lesion profile in degrees;
d is the dimension in centimetres of the lesion measured along the longest profile axis;
μ is the mean grey level value in units taken from an 8 bit (256) scale from the four profile lines;
GL contrast $= \mu_{\mathrm{nodule}} - \mu_{\mathrm{background}}/\mu_{\mathrm{nodule}}$;
$\Delta \mathrm{GL} = \mu_{\mathrm{nodule}} - \mu_{\mathrm{background}}$
SNR for the nodule $= \Delta \mathrm{GL}/\sqrt{\sigma^2_{\mathrm{nodule\ (s)}} + \sigma^2_{\mathrm{background\ (n)}}}$

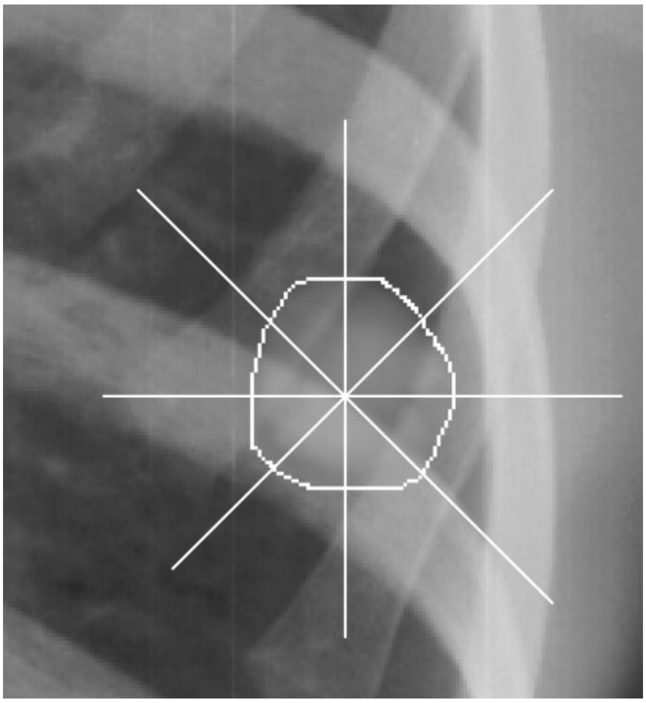

Figure 2. The boundary of each lesion was outlined by eye from the image appearance and this was then confirmed by using the % grey level reduction from its peak value as shown in Figures 3 and 4. The four profiles were taken as a minimum number of samples required to represent heterogeneity of nodule internal structure as well as its background.

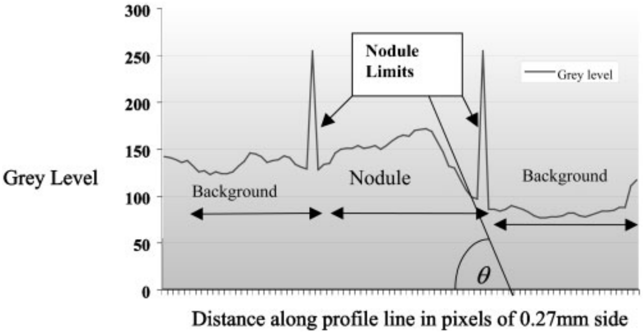

Figure 3. The mean values of grey level for nodule and background and the greatest value of θ from four profiles were used in the calculation of the conspicuity index of each of the 81 nodules.

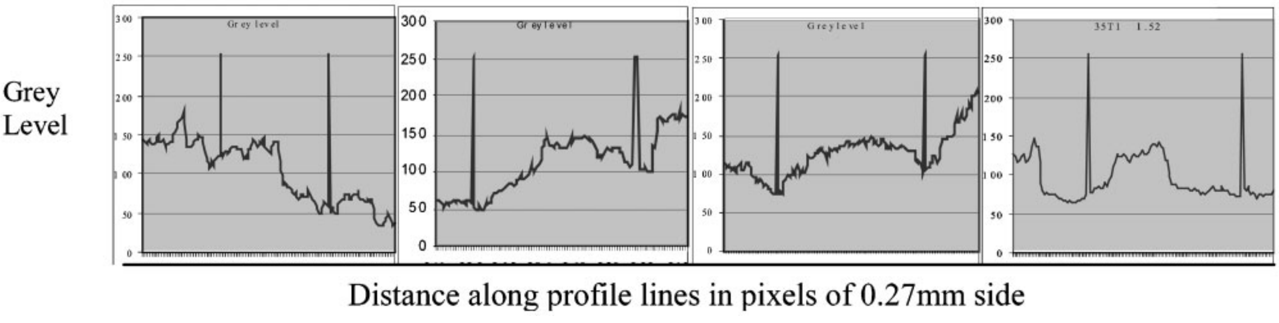

Figure 4. The four profiles for each nodule frequently displayed a wide variation in the edge sharpness.

σ=standard deviation of the grey levels of background (n) and internal structure (s) of lesions. The relationship, $\chi = d\ \tan[\theta-1]\Delta\mathrm{GL}/\sqrt{\sigma_s^2+\sigma_n^2}$ is derived from the definition of SNR and contrast from values of grey level described by Yocky et al [7] for studies of this kind using digital images.

The value of θ was found by taking a linear regression on the profile of the nodules. For each nodule the greatest value of θ was used in order to represent the most conspicuously sharp edge. Tan ($\theta-1$) was used to avoid infinite values. The size of the lesion, d, was represented by the number of pixels within the outlined lesion boundary along the longest profile line. The pixel size is related to the number of lines and pixels in the display and can be calculated from image size and resolution. In the present experiment with display pixel density of 1280×1024 this was 0.27 mm per pixel. The χ definition of conspicuity takes into account the complexity of the surrounding structure through σ_n and the sharpness of nodule outline though the angle of the profile edge slope θ and a representation of its size through d. These quantities are reported as having the strongest influence on the visual impact or "conspicuity" of the features [5].

Test procedure

Observers were informed that the chest images might contain single or multiple nodules or no nodules at all and their task was to decide on the presence or absence of pulmonary nodules; incidental findings were not to be called. 40 min per test bank was the maximum time permitted to avoid effects of fatigue. Explanation of the AFROC rating scale was given to the observers so that they would ascribe a location and a certainty value between (1 and 4) to all identified nodules. In AFROC methodology the observer ascribes a value to each detected lesion on a 1 to 4 scale of increasing confidence in the lesion being significant and a score of zero is given for a decision of no lesion present. Chakraborty [8] gives a comprehensive guide to the methodology of AFROCs and other variants of ROC analysis.

Eye tracking

The eye movements of four radiologists were tracked using a remote, infra-red pupil-corneal reflection device manufactured by ASL (Applied Science Labs, Bedford, MA). Comparisons were made between features of the eye movement, the AFROC decision score, and the χ values of the nodules. The eye movement parameter measured for this experiment was the mean value of visual dwell time over the lesion in seconds. 45 of the images were selected for eye tracking. These were selected to represent normal chest images and a range of conspicuity values of 33 of the nodules distributed in the range of chest zones shown in Figure 1. The images were embedded in the middle of each test bank to allow the observers to accommodate to the detection task and scoring method.

Analysis

Results were processed using ROCFIT® software (University of Chicago, Chicago, IL). A true positive (TP) result was accepted when the identified location of a nodule was within the radius of that nodule. Fixations, areas of interest (AOIs) and fixation duration from eye movement data were determined through the ASL software program EYENAL®.

Results

χ values and observer performance

The AFROC scores for the observers were pooled and compared with the χ values calculated for each nodule. The results (Figure 5) show a weak positive correlation (R^2=0.0054). Variation between the individual overall performances of the seven observers measured as the areas under their AFROC curves (A_1) gave a standard deviation of 0.08.

The mean values of χ for lesions missed by the observers were compared with those that were detected. There was no significant difference in the two groups (p=0.78). Figure 6 illustrates the way in which missed and detected lesions compare in terms of the characteristics used in the calculation of χ, as well as in their derived values of χ; and Figure 7 shows the differences in visual attention given to the lesions by the observers.

Figure 8 is an analysis of the surviving number of nodule decisions under visual fixation up to a limit of 15 s. The curves represent decisions falling into the four possible categories of true positive, false positive, true negative and false negative. The upper time limit for fixations that are associated with detection without recognition is generally considered to be 1 s. Fixation times longer than 1 s are thought to be associated with the cognitive processes of recognition and identification [9]. The survival curves show that 80% of the true negative (TN) decisions were made within 1 s but that nearly 80% of all positive decisions, both true (TP) and false (FP), were given the longer visual scrutiny associated with cognition. The false negative decisions (FN) were fixated for greater than 1000 ms in

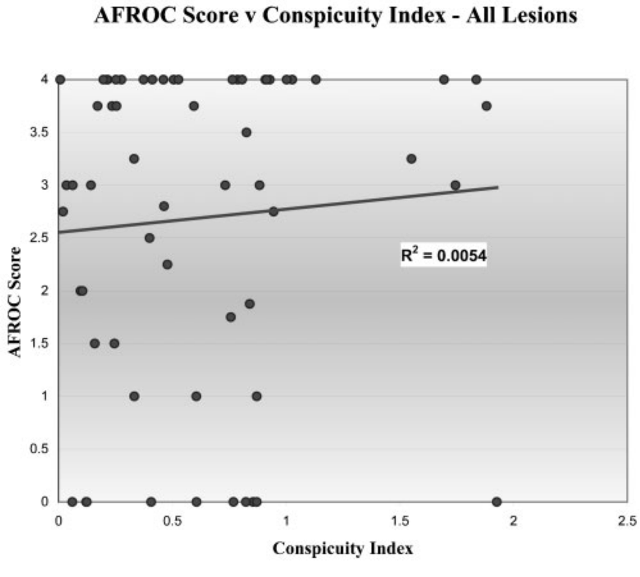

Figure 5. The correlation between the conspicuity index of the nodules and the probability of them being called diagnostically significant in the alternate free response operating characteristic (AFROC) scoring was weak (R^2=0.0054).

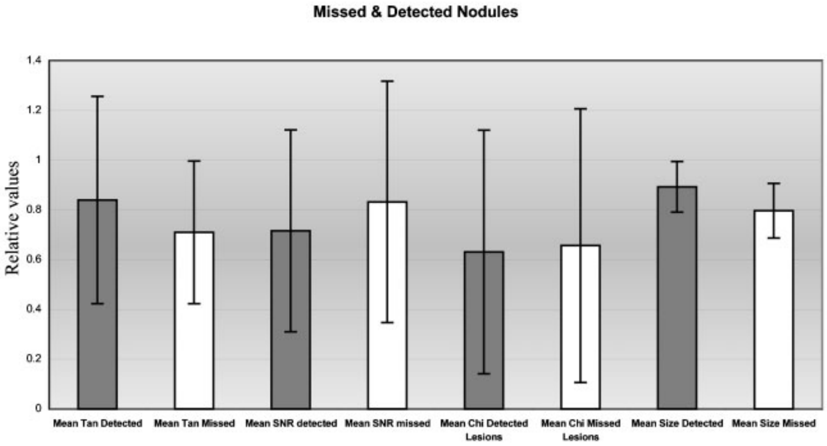

Figure 6. The seven radiologists detected 59 (72.8%) of the 81 lesions. There was no significant difference in the physical characteristics of the missed and detected lesions measured by their edge sharpness as the mean value of tan ($\theta-1$), signal to noise ratio (SNR), length in cross section of derived conspicuity index ($p=0.78$).

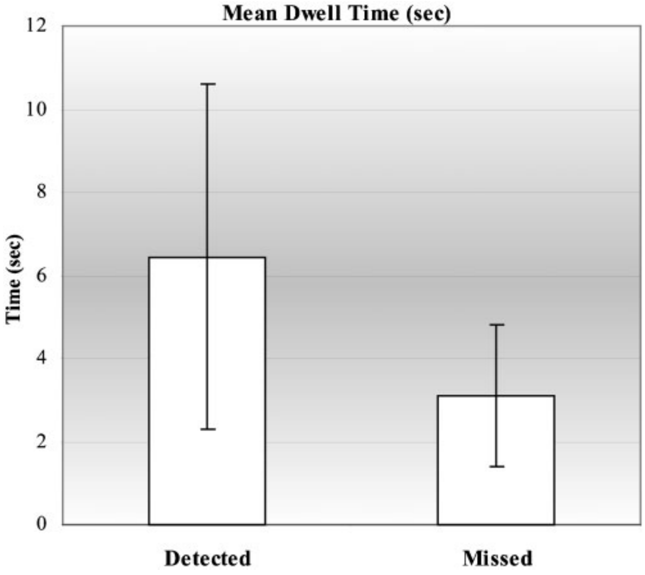

Figure 7. The eye movements of four of the radiologists were tracked. Their mean dwell-time of gaze duration for detected nodules was twice that of those missed. However, the missed lesions attracted average gaze duration of 3.1 s.

65% of cases. This suggests that the majority (65%) of false negative decisions were not due to failures of detection but of interpretation.

Discussion and conclusions

The error rate for all the 81 nodules reported here in terms of missed lesions was 27.2%. This is broadly in line with the findings of others in this area of work [1] but may have been elevated a little by the rigorous requirements of the AFROC methods we used that demand precise location information in the decision responses. From the eye tracking data, AFROC scores on positive decisions correlated well to visual dwell time ($R^2=0.692$) indicating that the levels of confidence in those positive decisions are linked to the time spent fixating the feature. This is consistent with a report on changes in time-related decision-making for radiologists during their training [9] and we suspect the finding would be more marked in experiments with less experienced observers. Eye tracking showed that all missed lesions were visually fixated and were dwelt on for an average time of 3.1 s. This was half the average dwell time for detected nodules but still well in excess of the 0.9 s dwell time accepted as the minimum period required for detection to occur [10]. Low correlation of conspicuity with AFROC does not necessarily imply that the χ measure or the AFROC methodology is invalid, although we accept that the conclusion is possible. The range of lesion sizes and contrasts in the test set of images was limited. The inclusion of a greater number of smaller and more subtle nodules might show a stronger correlation with correct observer decisions if we repeated the experiment but this would almost certainly increase the false negative rate. Limiting the number of profiles to four for measurement of the conspicuity was, to some extent, a compromise and clearly a larger sample of profiles will give better spatial information on the surrounding environment of the lesions. An alternative method would be to take the grey scale histogram data from within a lesion and from an annulus surrounding it to measure signal-to-surround information. The profile data we acquired, however, gave a good measure of the most acute slope to the nodule edge in each case, and the mean grey level taken from the profiles with its standard

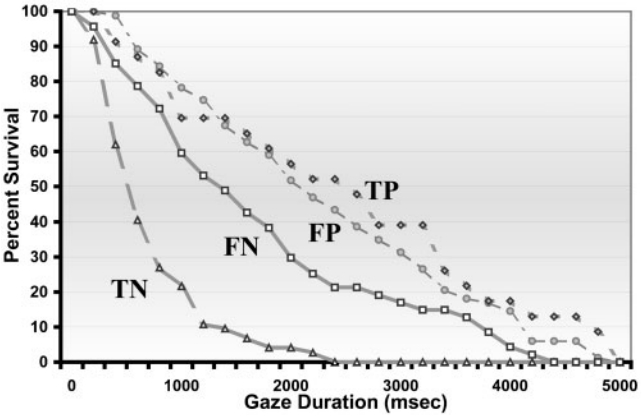

Figure 8. The percentage of nodules holding visual attention over a 15 s time interval for the four possible decision outcomes of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Note that nearly 70% of the missed lesions (FN) held a gaze duration of greater than 1 s. Deciding correctly that a chest zone contained no nodules was rapid (TN curve) but all positive decisions followed extended periods of attention (TP, FP).

deviation has been used to good effect by others [8]. Our interpretation of the findings relating to the χ measure of conspicuity and perceptual performance measured by AFROC is that the observers were making recognition (cognitive) errors because although they visually detected nodules they sometimes misinterpreted their significance [11]. This is supported by the survival analysis shown in Figure 8. Our conclusions support the idea that the complexity of the visual information in chest imaging makes it difficult for observers to discriminate between normal anatomical structures and nodular pathological features, even when such features have been made visually obvious by the imaging process. Such difficulties do not imply reader incompetence but suggest that perceptual rather than imaging limits may be the fundamental problem in some image interpretation tasks. The visual cognitive templates held in the mind for nodular pathological features may be so similar to templates for normal or inconsequential structures that one cannot distinguish reliably between the two. What is more, improving the resolution of the image may not help in the decision-making process if the anatomical background noise is enhanced to the same extent as the lesion. Certain other tasks, such as detection of micro-calcification, may not suffer this problem to the same extent because of the dissimilarity of those features to normal surrounding structures. However, our conclusions from this study may underline a need for aided decision-making in the diagnosis of lung cancer by chest radiology. Such aids might take the form of artificial intelligence, double reporting or some form of educational feedback on performance. For example it may be worth informing observers that the longer they deliberate over a negative decision the more likely it is to be incorrect (Figure 8). In a future paper we plan to discuss more extensively the implications of our finding that some decision errors appear to be related to the duration of visual attention. Generally we feel that our findings question the assumption that technological changes to improve contrast, image resolution and detective quantum efficiency will naturally and always result in an improved diagnostic outcome. We note carefully the observation made [12] that variations in observer performance are significantly larger than the technical variability of the diagnostic images they read.

References

1. Samuel S, Kundel HL, Nodine CF, Toto LC. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. Radiology 1995;194:895–902.
2. Turkington PM, Kennan N, Greenstone MA. Misinterpretation of the chest x-ray as a factor in delayed diagnosis of lung cancer. Postgrad Med J 2002;78:158–60.
3. Kundel HL, Nodine CF, Krupinski EA. Searching for lung nodules. Visual dwell indicates location of false-positive and false negative decisions. Invest Radiol 1989;6:472–8.
4. Kimme-Smith C, Hart EM, Goldin JG, Johnson TD, Terwilliger R, Aberle DR. Detection of simulated lung nodules with computed radiography: effects of nodule size, local optical density, global object thickness, and exposure. Acad Radiol 1996;3:735–41.
5. Kundel HL, Revesz G. Lesion conspicuity, structured noise, and film reader error. AJR Am J Roentgenol 1976; 126:1233–8.
6. Chakraborty DP, Winter LHL. Free response methodology: alternate analysis and a new observer performance experiment. Radiology 1990;174:873–81.
7. Yocky D, Seely GW, Ovitt TW, Roehrig H, Dalls W. Computer simulated lung nodules in digital chest radiographs for detection studies. Invest Radiol 1990;8:902–7.
8. Chakraborty DP. The FROC, AFROC and DROC variants of the ROC analysis. In: Beutel J, Kundel HL, Van Metter RL, editors. Handbook of medical imaging. Bellingham, WA: Society of Photo-optical Instrumentation Engineers, 2000: 771–96.
9. Krupinski E, editor. The past and future of radiologic errors. Proceedings of Medical Imaging 1999—Image Perception and Performance; 1999 February 24–25; San Diego. Bellingham, WA: Society of Photo-optical Instrumentation Engineers, 1999.
10. Nodine C, Kundel H. A visual dwell algorithm can aid search and recognition of missed lung nodules in chest radiographs. In: Brogan D, editor. Visual search. London: Taylor Francis, 1990:399–405.
11. Manning DJ, Leach J, Bunting S. A comparison of expert and novice performance in the detection of simulated pulmonary nodules. Radiography 2000;6:111–6.
12. Krupinski E, editor. Assessing the value of diagnostic imaging: the role of perception. Proceedings of Medical Imaging 2000; Image Perception and Performance; 2000 February 16–17; San Diego. Bellingham, WA: Society of Photo-optical Instrumentation Engineers, 2000.