SHORT COMMUNICATION

Time-dependent observer errors in pulmonary nodule detection

[1]D MANNING, PhD, FInstP, [1]S C BARKER-MILL, PhD, [1]T DONOVAN, MSc and [2]T CRAWFORD, PhD

[1]School of Medical Imaging Sciences, St Martin's College, Lancaster LA1 3JD and [2]Department of Psychology, Lancaster University, Lancaster, UK

ABSTRACT. The work was carried out to investigate differences in visual search characteristics between groups of observers with different levels of experience in the task of pulmonary nodule detection in chest radiology and we report here on these differences in respect of time related decisions. Volunteer observers were divided into three groups depending on their level of expertise. There were eight radiologists, eight radiographers and eight novices. Their task was to detect pulmonary nodules in a test bank of 120 digitized posteroanterior (PA) chest radiographs. Five of the eight radiographers were tested twice: once before and once after a 6-month training programme in interpretation of the adult chest radiograph. During each test session the observers' eye movements were tracked. Data on the observers' decisions through Alternate Free Response Operating Characteristic (AFROC) methodology were correlated to their eye-movement and fixation patterns. True negative decisions from all observers were associated with shorter fixation times than false negative decisions. No correct negative decisions were made after fixations exceeding 3 s.

Received 2 March 2005
Revised 24 August 2005
Accepted 30 September 2005

DOI: 10.1259/bjr/13453920

In previously reported studies we have investigated observer experience [1, 2] and the effects of lesion conspicuity [3] on performance in nodule detection from plain chest radiology. Comparisons of diagnostic performance through alternate free response receiver operating characteristic (AFROC) [4] showed that after training and extensive caseload experience radiographers' detection rate improved to approach that of the experts. AFROC is a variation on receiver operating characteristic methodology that takes into account the location in the image of the observer decision. Through this technique the observer must indicate not only his decision on the disease status of the image (positive or negative for the presence of a pulmonary nodule) but also the location of the lesion.

The observer groups were as follows:

- eight first-year student radiographers considered as novices to chest interpretation tasks;
- eight experienced clinical radiographers before training in chest interpretation;
- five of the eight radiographers after their training in chest interpretation; and
- eight radiologists.

Eye-tracking these observers gave insight into differences between the groups in terms of their visual search strategies and we concluded that, amongst other things, the experts were more economical in their patterns of search, carried out fewer fixations and spent less time on the task. After a training period that included a minimum of 500 cases but no specific instruction in search patterns, the radiographers developed spontaneously similar search strategies to those of the radiologists. But both radiologists and trained radiographers had a false negative error-rate in excess of 40% [2]. We accept that the requirements of AFROC are stringent (a false negative is defined by a missed lesion rather than an incorrect case decision); but this is still a significant miss-rate and in this paper we report our observations on the relationship between the experience of the observer, the duration of visual attention and the probability of a false negative error in the task.

Eye movement, visual attention and visual scan paths

The field of view in humans extends over 180° but only the centre of the field provides sharp details. Consequently, we move our eyes (and heads) to bring interesting features into the centre. The pause over the point of interest is known as a foveal fixation and has duration of about 100–300 ms before we move our interest to a new location in a fast jump called a saccade. Fixations are characterized by multiple (clustering) nature when we dwell extensively on a location because the eyes do not remain stationary for long before losing sensitivity. Saccades are too fast for information to be gathered during their operation and so eye-tracking experiments assume that fixations represent the location of conscious attention. Viewers are not conscious of the pattern of their saccades during an observation of a scene, however, and it has been shown that scanpaths for an individual are modified by the task presented [5]. The path is the sequence in which image details are scrutinized but fixation point clusters illustrate the parts of the scene the observer find most interesting as opposed to the parts that receive no attention at all. It has been noted that the same details in an image can be accessed using quite different scanpaths but similar fixation clusters [6]. As a result, analysis of fixation cluster location and dwell time gives a better opportunity for performance comparisons than the scanpath sequence if the research question is focused on whether image features are being hit. However, scanpath data can give fascinating insights into how individuals prioritise locations of semantic interest although the analysis of these paths is difficult and sometimes inconclusive.

Errors in radiology are known to be multifactorial but a sizable proportion are known to be perceptual in origin [7, 8]. This has led to interest in the possible sources of perceptual error and the visual strategies of radiologists in their task [9]. Early work on the visual dwell behaviour of radiologists during film reading was carried out by Kundel et al [10–12] who went on to later describe survival curve methods for analysing time-related decision-making in those observers [13, 14]. Our aim here is to report on the timing of decisions, their accuracy and their relationship to experience by the survival analysis of decisions from eye-tracking data.

Aim

The aim of this study was to report an observed relationship between the experience of the observer, the duration of visual attention and the probability of incorrect decisions in pulmonary nodule detection.

Materials and methods

Observers

Eight volunteers were recruited from a cohort of first year radiography students who acted as the novice group of film readers. Eight experienced radiographers volunteered from a post-graduate course on chest image interpretation at the commencement of their course. Five of these radiographers then offered themselves for repeat testing at the end of their 6 months training and 500 cases of experience. Eight radiologists with extensive and current experience in chest radiology volunteered to act as the most experienced group of observers. All observers gave their consent and all performance data were made anonymous. Recruitment of all the observers and the conduct of the experiments followed the ethical guidelines for experiments involving volunteer human subjects at St Martin's College and Lancaster University. At the time when the data for these experiments were collected (in 2001) there was no requirement of COREC approval for NHS employees who were acting as volunteer participants in this type of research.

Detectable nodules

Each observer viewed a bank of 120 digitized chest images of adults. The images contained 81 pulmonary nodules agreed as significant in pathological appearance from confirmed radiological reports. Nodules were roughly circular and ranged in diameter size from 5 mm to 20 mm with varying degrees of measured conspicuity [3]. Nine films contained more than one nodule and 30 nodules were located in these multi-nodule films. Normal films were included in the observer task and the complete test bank was divided into three sets of 40 images with prevalence-rates of 12%, 50% and 82%.

Observer performance measurement

Alternate free response operating characteristic methodology (AFROC) was used [4]. This required observers to indicate a location to a decision for a lesion and to assign a score between 1 and 4 on their level of confidence in that decision. A zero score was allocated to all decisions of ''no nodule present''. In AFROC methodology false positive decisions are treated in the following way: the highest scoring false positive decision is the only one recorded per image which avoids the possibility of infinite values in summing false positive responses.

Observer test sessions were never longer than 1 h to avoid the effects of fatigue on performance. There was a minimum 6-month interval between the before- and after-training observer tests on the radiographers to give an effective case-memory washout period.

Parameters

All eye-tracking data comparing the performance of the observer groups were processed through the ASL (Applied Science Laboratories, Bedford, MA), software EYENAL®. The measured parameter from the eye tracking for use in the present study was the accumulated dwell time at each decision point.

A fixation

We defined a fixation (visual dwell) as a point of gaze remaining continuously within a 1° area for at least 100 ms. A 1° angle subtended from the eye to the image at a viewing distance of 40 cm approximates to a circle the size of a 2.5 cm diameter disc at arms length. Refixations were summed to give cumulative fixation times to clusters when there was overlap of the 1° areas up to a 5° area. These definitions are ones that are commonly used in work of this kind [9, 14]. The observers were allowed to search freely and no limit was imposed on the duration of inspection for any given image.

A true negative decision was defined as a timed fixation of a lesion-free zone of the chest image that elicited a zero response from the observers on the AFROC scale.

The dwell-time data for all fixations related to positive and negative decisions were analysed through the statistical package SPSS® to provide information on the percentage survival of decisions over time.

These data allowed us to characterize the observers' decisions in greater detail than true and false negative and true and false positive outcome, giving an opportunity to identify time-related features of decision outcomes. The time related information on decisions gave an opportunity to analyse time differences between correct and incorrect decisions whether positive or negative.

Results

Time related decisions: survival analysis

We investigated how the four possible decision outcomes of true and false positive (TP FP) and true and false negative (TN FN) related to the duration of gaze through a survival analysis of the fixation data. The results are presented in Tables 1–4 and Figures 1–4. Survival analysis is used in this context as a means of showing the proportions of decisions that are completed for each category (TP FP TN FN) at increasing accumulated time intervals of visual attention. It is analogous to the cell survival curves used in radiobiology to indicate the proportion of cells surviving increasing radiation doses [15]. So in Table 1 and Figure 1 for example, 50% of all true positive decisions made by the radiologists survived 2200 ms of accumulated visual attention on the tumour location but none of their true positive decisions took more than 5000 ms of visual scrutiny.

In the tables we have highlighted the time values for the completion of all the observer decisions in their outcome categories to draw attention to the differences in duration between the true and false decisions in each case. For all observers the false negative decisions took them significantly longer than true negatives. For positive decisions, correct decisions were made faster than incorrect ones except in the case of the novices who dwelt longer on genuine lesions before deciding they were pathologies.

The survival curves show that for all observers 50% of all their true negative decisions were made within 1000 ms of gaze duration and the false negative decisions have dwell-times somewhere between the positives (true and false) and the true negatives. This is consistent with the findings of others who have carried out this type of analysis [14] but our results indicate extension along the time axis with decreasing levels of experience.

Table 1. Radiologists

Survival duration (ms) for each decision group				
TN	FN	TP	FP	% surviving
0	0	0	0	100
200	220	400	220	95
220	240	455	250	90
450	500	1000	750	80
680	700	1680	1700	70
750	800	1800	1800	60
790	1000	2200	2200	50
820	1250	2500	2700	40
900	1600	3000	3500	30
1010	2000	3500	4000	20
1800	3000	4500	4700	10
2250	4750	5000	5500	0

TN, true negative; TP, true positive; FN, false negative; FP, false positive.

Table 2. Radiographers post training in chest interpretation

Survival duration (ms) for each decision group				
TN	FN	TP	FP	% surviving
0	0	0	0	100
200	220	300	500	95
300	330	800	600	90
450	500	900	800	80
500	500	1200	1000	70
500	550	1850	2000	60
550	700	2500	4000	50
600	1100	3000	5000	40
800	2000	4400	5500	30
1000	2500	5000	6000	20
1500	4000	6000	7000	10
3000	5000	6600	8000	0

TN, true negative; TP, true positive; FN, false negative; FP, false positive.

Discussion

The aim of this work was to analyse eye-tracking data to classify the false negative errors made when observer groups with different levels of experience are asked to detect pulmonary nodules. The different levels of experience in the groups give some insight into how, if at all, errors vary with expertise. The time related data provided measurable differences in the fixation times associated with observers making correct or incorrect decisions.

Survival analysis

Observers' decisions are made over time periods that can be related to the duration of visual fixations from eye-tracking data. Mean or median dwell times over trials and between readers can be calculated but dwell times are not normally distributed, making statistical comparisons difficult to interpret. The technique of survival analysis has been found useful in these circumstances [13, 14] and this operation on the data demonstrated some consistent findings for categories of decisions.

Figures 1 and 2 show the family of decisions survival curves for radiologists and radiographers with chest interpretation training derived from the data in Tables 1 and 2. Virtually all the true negative decisions made by these, the most experienced observers, were made within 2 s of cumulative fixation time on an image feature. False negative decisions were characterized by longer dwell times but no negative decisions (either true or false) were made by radiologists after 4.75 s fixation. Figures 3 and 4 and Tables 3 and 4 show similar findings for the low experience groups but there is a tendency towards longer fixation times for incorrect negative decisions with decreasing levels of experience. In the case of the novices (Table 4) lesions were missed after a cumulative gaze duration of up to 8 s. The time difference in the proportions of true and false negative decisions made by these observers can be summarized by saying that in

Table 3. Radiographers pre training in chest interpretation

Survival duration (ms) for each decision group				
TN	FN	TP	FP	% surviving
0	0	0	0	100
300	300	350	350	95
330	400	800	820	90
330	500	1100	1200	80
400	600	2000	2000	70
560	1200	3000	3100	60
700	1600	3300	3800	50
850	2000	3800	4800	40
1000	2400	4200	5500	30
1200	3000	5100	6500	20
1600	4500	6500	8000	10
2400	6500	8200	9000	0

TN, true negative; TP, true positive; FN, false negative; FP, false positive.

Table 4. Novices

Survival duration (ms) for each decision group				
TN	FN	TP	FP	% surviving
0	0	0	0	100
300	300	500	400	95
400	400	1000	500	90
500	500	1700	800	80
550	600	2000	1200	70
600	1000	2600	1500	60
680	1500	3800	2000	50
800	1800	4500	3000	40
900	2200	5500	4500	30
1500	3500	6500	5600	20
2200	5000	9000	8000	10
3000	8000	12000	10000	0

TN, true negative; TP, true positive; FN, false negative; FP, false positive.

this experiment, all negative decisions made after gaze duration of 3 s were incorrect.

It seems that correct negative decisions (correct decisions that normal features are not nodules) tend to be made rapidly after fixation occurs and that this is a feature of expert performance. Conversely, incorrect negative decisions are characterized by extended gaze duration especially in novice readers. When areas of the images hold their attention for several seconds of accumulated fixation time, observers show a semantic interest in the appearances. This suggests that they are suspicious of the feature and they are operating on the information at a perceptual/cognitive level. These errors are not failures of detection but of recognition and decision and can be explained partly by the visual ambiguities of the image and partly by the level of experience of the reader. The finding may be important because of: (a) its potential for reducing false negative error if visual dwell information is fed back to the reader; and (b) its potential use in training schemes for radiologists and monitoring the effects of caseload experience.

If these results are reproducible in other settings and our interpretation of their meaning is correct there are several ways that they may help in the improvement of observer performance. A simple expedient is to inform film readers

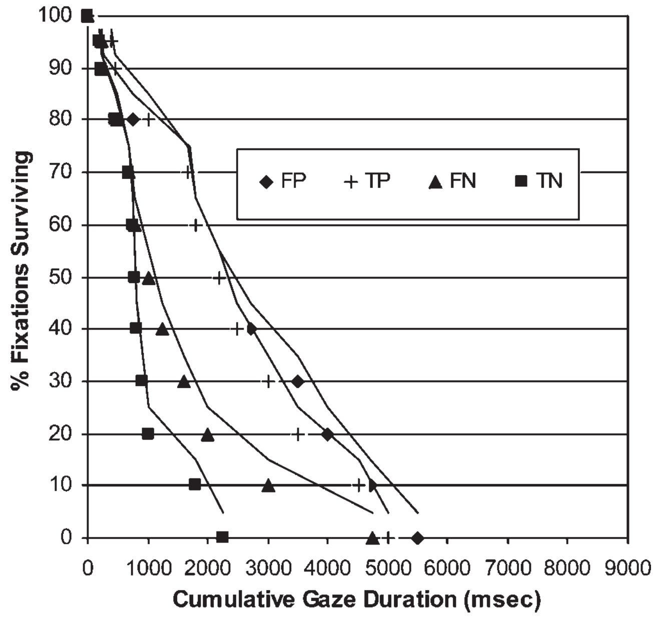

Figure 1. Time-related decisions for radiologists. Compared with less experienced observers the radiologists made their decisions faster. The separation between the positive and negative decision pair curves is closer than for all other groups.

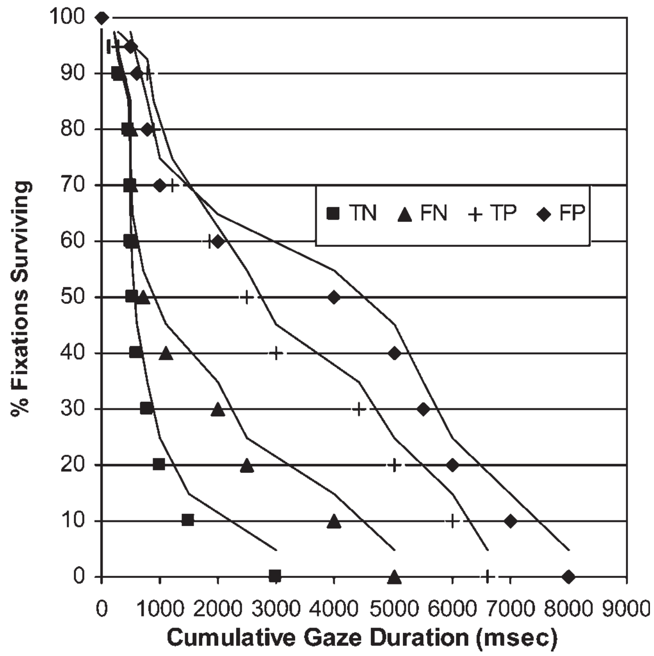

Figure 2. Time-related decisions for radiographers post-training. After experience and training over 6 months and 500 cases the radiographers have speeded their decision making and markedly reduced the separation of their true and false negative curves. Compare with Figure 3.

that their negative decisions are more likely to be incorrect when they are made after a period of indecision over a particular image feature. In short, we suggest informing observers that if a feature looks suspicious enough to warrant more than 2 s of their attention it is probably not innocent. More sophisticated aids linked to these findings

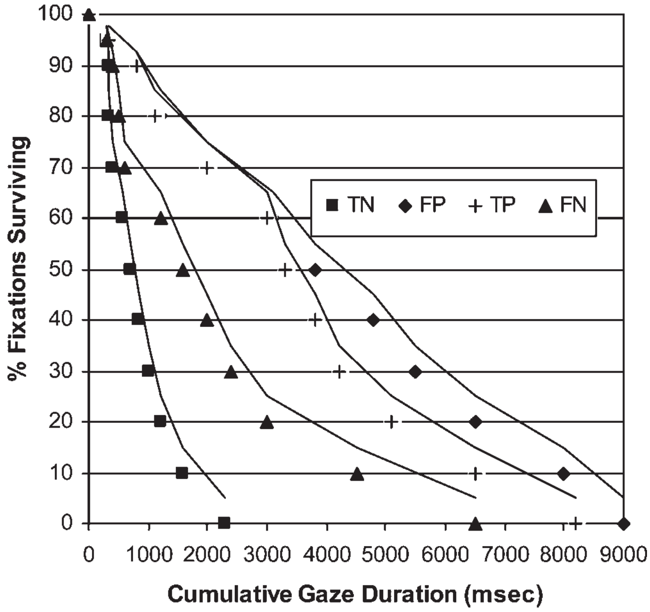

Figure 3. Time-related decisions for radiographers pre-training.

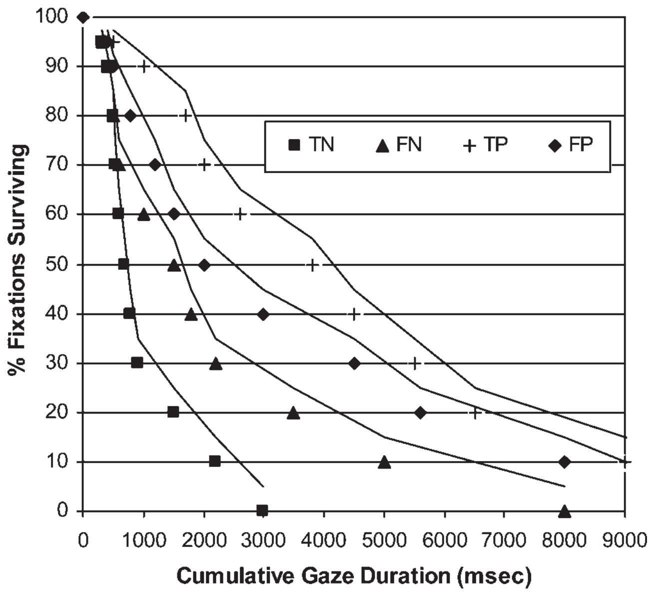

Figure 4. Time–related decisions for novices. Novices show a similar trend to the experienced observers although their positive decision curves are reversed and more widely separated. All their decisions extend further along the time axis.

might involve computer aided feedback to observers in real-time to give indication of the gaze duration for individual locations in the image.

Conclusion

The data and their analysis from an eye-tracking experiment have given insights into the errors made by observers with different levels of experience. The most notable outcomes of this are that the duration of fixation on a feature in the chest image may be an effective discriminator in predicting whether a negative decision will be correct or incorrect.

We consider this observation to be potentially important for feedback strategies for education purposes and to aid performance.

Acknowledgment

We wish to thank the radiologists, radiographers and students, all of whom are associated with the School of Medical Imaging Sciences at St Martin's College, Lancaster, UK, who kindly agreed to take part in this study. Eye-tracking equipment for this research was supplied through a grant from the Peter Barker-Mill Memorial Trust.

References

1. Manning DJ, Ethell S, Crawford T. An eye-tracking AFROC study of the influence of experience and training on chest x-ray interpretation. Medical imaging 2003; Image Perception Observer Performance and Technology Assessment. Editors Dev Chakraborty and Elizabeth Krupinski, Proc SPIE 2003;5034:257–266.
2. Manning DJ, Ethell SC, Crawford T, Donovan T. How do radiologists do it? The influence of experience and training on searching for chest nodules. Radiography Available online 23 March 2005. Item 49, www.sciencedirect.com/science/journal/10788174.
3. Manning DJ, Ethell SC, Donovan T. Detection or decision errors? Missed lung cancer from posteroanterior chest radiograph. Br J Radiol 2004;77:231–5.
4. Chakraborty DP. Statistical power in observer performance studies: comparison of the receiver operating characteristic and free response method in tasks involving localization. Acad Radiol 2002;9:147–56.
5. Yarbus AL. Eye movements and vision. New York, NY: Plenum Press, 1967.
6. Kundel HL. Reader error, object recognition and visual search. In: Chakraborty D, Eckstein M, editors. Medical imaging 2004: image perception, observer performance and technology assessment. Proceedings of SPIE 2004;5372: 1–9.
7. Fitzgerald R. Error in radiology. Clin Radiol 2001;56:938–46.
8. Chief Medical Officer. Learning from failure: evidence and experience. An Organisation with a Memory. London: Stationery Office, 2000:1–7.
9. Gale AG. Human response to visual stimuli. In: Hendee WR, Wells PNT, editors. The perception of visual information, 2nd edition. New York, NY: Springer, 2000:134.
10. Kundel H, Nodine C, Carmody D. Visual scanning, pattern recognition and decision making in pulmonary nodule detection. Invest Radiol 1978;13:175–81.
11. Kundel H, Nodine C, Krupinski E. Searching for lung nodules: visual dwell indicates locations of false positiveand false negative decisions. Invest Radiol 1989;24:472–8.
12. Kundel H, Nodine C, Toto L. Searching for lung nodules: the guidance of visual scanning. Invest Radiol 1991;26: 777–81.
13. Anderson S, Auquier A, Hauk WW, Oakes D, Vandaele W, Weisberg HI. Statistical methods for comparative studies. New York, NY: John Wiley & Sons, 1980.
14. Kundel H. Visual search in medical images. In: Beutel J, Kundel H, Van Metter R, editors. Handbook of medical imaging Vol 1 Physics and psychophysics. Bellingham WA: SPIE, 2000:837–58.
15. Motulsky H. Intuitive biostatistics. Oxford: Oxford University Press Inc., 1995:Chapter 6.