Categories of observer error from eye-tracking and AFROC data.

David Manning, Susan Ethell, Tim Donovan.
Department of Radiography & Imaging Sciences, St Martin's College, Bowerham Road, Lancaster. LA1 3JD UK. email, d.manning@ucsm.ac.uk

ABSTRACT

Twenty-four volunteer observers were divided into groups of eight radiologists, eight radiographers and eight novices to carry out a pulmonary nodule detection task on a test bank of 120 digitised PA chest radiographs. The eight radiographers were tested twice: before and after a six-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 AFROC methodology were correlated to their eye-movement and fixation patterns. False negative error-rates were recorded as 41% for the radiologists, 45% for the novices, 47% for the radiographers before training and 42% for the radiographers after training. The errors were sub-classified into search, recognition and decision errors depending on the duration of the fixation-time for each faulty response. Errors due to satisfaction of search were determined from images with multiple nodules. Differences between the groups were shown. Errors due to inefficient search were in the minority for all the observer groups and the dominant cause of unreported nodules was incorrect decision-making. 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 three seconds.

Keywords : Eye-tracking, AFROC, fixation duration, false-negative errors, satisfaction of search, experience.

1. INTRODUCTION

In a previously reported study we investigated the performance of observer groups with varying degrees of experience in the specific task of nodule detection in PA chest radiographs[1]. The purpose of that work was to compare expert radiologists with radiographers before and after training in chest x-ray interpretation to evaluate the effectiveness of training and investigate some features of expert performance. Comparisons of their diagnostic performance through AFROC methodology[2] showed that after training and extensive caseload experience the radiographers improved their detection rate in this specific task to approach that of the experts. A group of novice, first year student radiographers were included in the study to act as controls for performance comparisons. The groups were: eight experienced radiologists with current responsibilities in reporting plain chest radiographs, eight radiographers who were enrolled on a postgraduate programme of training in chest interpretation in the adult and eight novice radiography students newly enrolled on their initial training programme.
Eye-tracking the activities of all these observers during the detection of the nodules from a test bank of films 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 their training period, which included a minimum of 5oo cases but no specific instruction in search patterns, the radiographers had developed, spontaneously, similar strategies to those of the radiologists. However, there was a significant (False Negatives) error-rate for both radiographers and the radiologists. Rates in excess of 40% were recorded for all the observer groups. Even taking into account the stringent requirements of AFROC (where a false negative is defined by a missed lesion rather than an incorrect decision on a whole film) this is still a significant miss-rate. Errors in radiology are known to be multi-factorial [3] but broadly they have causes that can be investigated at the personal or individual level, or through an organisational, systems approach. The work presented here is experimental in nature and eliminates many of the organisational components of an error chain that have been described recently in a 'Swiss-cheese model' of accident causation[4]. It is aimed at gaining a better understanding of the perceptual and judgmental causes of radiological errors by a further analysis of eye-tracking data obtained from individuals with four different levels of experience.

1.1 Aim:

To classify sub-types of the false negative errors made by observers in the task of detecting pulmonary nodules.

Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment, edited by Dev P. Chakraborty, Miguel P. Eckstein, Proceedings of SPIE Vol. 5372 (SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.533489

2. MATERIALS and METHODS

2.1 Detectable Nodules

Each of the 24 observers viewed a bank of 120 digitised images. The chest images of adults contained 81 pulmonary nodules agreed as significant in pathological appearance from confirmed radiological reports. Nodules were roughly circular and ranged in size from 5mm to 20mm diameter with varying degrees of conspicuity measured by a method reported by us previously [5,6]. Nine films contained more than one nodule so that 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 to give prevalence-rates of 12% 50% and 82%.

2.2 Observer performance measurement

Alternate free response operating characteristic methodology (AFROC) was used [2]. 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 one hour to avoid the effects of fatigue on performance. There was a minimum six-month interval between the before-and-after-training observer tests on the radiographers to give an effective case-memory washout period.

2.3 Parameters

All eye-tracking data comparing the performance of the observer groups were processed through the ASL (Applied Science Laboratories, Bedford, Mass.), software EYENAL® and parameters measured from the eye tracking were:

- The mean saccadic amplitude per image in degrees subtended from the eye
- The coverage of the image area ,
- The number of fixations per image,
- The accumulated dwell time at each decision point
- The total duration of film scrutiny in seconds.

2.4 A Fixation.

We defined a fixation as a point of gaze remaining continuously within a 1degree area for at least 100milliseconds. The observers were allowed to search freely and no limit was imposed on the duration of inspection for any given image. Time thresholds were set in the data collection to categorise errors into perceptual or cognitive subsets:

- non-fixated nodule errors or search errors (dwell <100msec)
- recognition errors (dwell >100msec <200msec) and
- decision errors (dwell > 200msec)

A true negative decision was defined as a timed fixation of a lesion-free zone of the chest image that elicited a zero response 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 characterise 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 the errors was the basis for their categorisation but also gave an opportunity to see if there were response time differences between correct and incorrect decisions.

2.5 Multi-nodule films.

Decisions made on nodules in images where more than one nodule was present were analysed as a subset. This gave data on the possibility of early termination of search after the successful detection of one of the lesions; the satisfaction of search phenomenon[7].

3. RESULTS

3.1 False Negatives

Table 1 and figures 1 and 2 show the overall miss-rate for the detection of pulmonary nodules for the four subgroups of observers. The error rate of more than 40% is at the high end of the range for unreported lesions in studies of this kind and we think that this reflects the rigorous nature of the scoring system using AFROC methods.
Search errors were defined as those nodules not scored through AFROC and not fixated in the data from eye-tracking. By applying time-threshold criteria to those nodule that were fixated but not scored, we show in Table 1 the proportion of missed lesions that were fixated briefly (recognition errors - dwell >100msec <200msec) compared with those that were scrutinised for longer (decision errors - dwell > 200msec). In all the observer groups, between 54% and 74% of misses were due to decision errors and missed lesions due to faulty search were comparatively rare (4%to 12%).

3.2 Errors from films with multiple nodules

In Table 2 and figure 3 we show the proportion of fixated and non-fixated false negatives in those images with more than one nodule. The proportion of total errors that could be classified in this way ranged from 18% for the radiologists to 25% for the radiographers before training. For all the observers, most of these unreported lesions were fixated for longer than 200msec and, as in the case of errors from chest images with solitary nodules, were therefore classified in our scheme as decision errors.

3.3 Time related decisions – survival analysis.

We were able to investigate how the four possible decision outcomes of True and False positive and True and False negative related to the duration of gaze through a survival analysis of the fixation data. The results are presented in Figures 4 to 7. In these graphs we have inserted a vertical line to indicate a temporal threshold of 2 seconds. This is an arbitrary time limit but is useful as a baseline for comparing the proportions of the four decisions surviving at that point. Two seconds is also of duration long enough for the decisions to be considered cognitive rather than perceptual ones.

4. DISCUSSION

The aim of this work was to analyse data acquired from eye-tracking to classify the false negative errors made when observer groups with different levels of experience are asked to detect pulmonary nodules. An underlying question here was whether lesions are missed because they are not fixated or whether they are seen by the observers but then rejected as candidate nodules because of ambiguities in their appearance. Second, the different levels of experience in the groups would give some insight into how, if at all, these types of error vary with expertise. A third point of interest, made available from the time related data, was whether there were measurable differences in the fixation times associated with observers making correct or incorrect decisions.

4.1 General patterns of error

The overall classification of error rates shown in Table 1 and Figure 1 gives the clear indication that the pattern of errors for the radiographers and the novices is essentially the same as that of the radiologists (although the radiologists' performance is generally better). Figure 2 shows how in all cases, decision errors are the dominant form of mistake in this task. However, the novices showed the lowest error rate due to inadequate search (only 4% of their errors were search errors). This finding fits with our previous report that their mean saccadic amplitude was lower and fixation density in the task was greater than that of the other observer groups even though their overall performance measured as the area under the AFROC curve (A_1) was lower. The novices are, therefore, exhaustive in their visual coverage of the images but are less effective than the more experienced observers in recognising pulmonary nodules or deciding that a detected feature is this type of pathology.
After their specialist training in chest image interpretation the radiographers' error profile became less like that of the novices and took on some of the features of the radiologists'. For example they made rather more search errors but fewer decision errors. This suggests that training and experience provides greater skill and confidence in the identification of pathology [8] and that this is associated with a less pedantic search strategy.

4.2 Satisfaction of search errors?

Satisfaction of search is a source of false negative error where the observer is thought to terminate prematurely the search for further significant pathology on the discovery of a lesion. In cases where multiple pathologies exist in the same image this error can be of concern because a clinically significant lesion may be missed while a relatively trivial disorder may be reported.

This experiment was highly task-specific in requiring only the detection of pulmonary nodules. Some of the clinical images had co-existing abnormalities or normal variants, some were free of nodules but had some other reportable feature and some were normal. But nine of the images contained multiple nodules. A separate analysis of errors made by the 24 observers for these films where one or more nodules were reported but at least one other nodule was missed could be taken as a comment on the satisfaction of search as a component of the false negative errors.

Figure 3 shows that 18% of the lesions missed by the radiologists were in multiple nodule films. Analysis of their eye-tracking data showed that only a small fraction of these errors (12.8 % of the errors so defined) were genuine errors of search and the majority (87%) were fixated for longer than 200msec. A similar pattern is shown in Figure 3 to exist for the other observers. This suggests that the term 'satisfaction of *search'* may not be the most accurate description of what is being observed in this case. The errors appear to be cognitive in nature and may be due to conscious decisions by the observers to terminate further reporting of what is considered by them to be the same clinical condition. It might be a better to term this type of error as 'satisfaction of decision'. Clarification of this issue may be possible by introducing a recorded commentary by observers at the time of decision-making.

4.3 Survival analysis

Observers make decisions that fall into one of four possible categories as outlined in 3.3 above. These decisions are made over measurable periods of time that can be related to the duration of visual fixations from eye-tracking data and an analysis of those data in the experiment demonstrated some interesting and consistent findings for negative decisions.

Figure 4 shows the family of survival curves for true negative (TN), false negative (FN), true positive (TP), and false positive (FP) decisions for the radiologists. Virtually all the true negative decisions made by the radiologists' were made within the arbitrarily chosen threshold of 2 seconds fixation on an image feature. 97.5% of their true negative decisions were made within 2 seconds of visual fixation. But only 80% of their false negative decisions were made within 2 seconds of fixation. This percentage steadily increased up to 4.8 seconds of gaze duration and no negative decisions on the status of a feature were made by radiologists after this length of fixation time. This time difference in the proportions of true and false negative decisions made by the radiologists can be summarised by saying that in this experiment, all negative decisions made after gaze duration of 2.2seconds were incorrect.

This trend was maintained in the remaining survival curves for all the other observers as shown in Figures 5,6 and 7. It seems that a correct negative decision (that a feature is *not* a nodule) tends to be made rapidly after fixation occurs. Conversely, incorrect negative decisions – the false negatives that in this experiment were in excess of 40% of the nodules present – are characterised by extended gaze duration. The semantic interest that observers show for areas of the images that hold their attention for several seconds suggests that they are suspicious of the appearances and they are operating on the information at a cognitive level. These errors are not failures of detection or of perception but of decision and can be explained, at least in part, by the visual ambiguities that characterise noise-limited medical images of complex anatomy.

The finding is important because of its potential for reducing false negative error and we are eager to know if others have found similar results in eye-tracking experiments using complex images such as mammograms, chest radiographs or MR images. If our results are reproducible 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 that their negative decisions are likely to be incorrect when they are made after a period of indecision over a particular image feature. In short, giving feedback to observers that if a feature looks suspicious enough to warrant more than two seconds of their attention it is probably not innocent.

More complicated technological aids linked to these findings might involve computer aided feedback to observers in real-time to give indication of the gaze duration for individual locations in the image.

5. CONCLUSION

The additional data and their reanalysis from the eye-tracking experiment have given new insights into the errors made by observers with different levels of experience. There is some confirmation of the findings we reported in the earlier work based on fewer subjects, that there are behaviours and styles shown by experts that are assumed by less experienced observers as they acquire greater familiarity with the task. The most notable outcomes of this report are:

- Decision errors make the largest contribution to the number of unreported pulmonary nodules in chest radiography irrespective of the level of observer experience,
- Inefficient , faulty or inadequate search makes only a minor contribution to error rate,
- That the 'satisfaction of search' phenomenon may be an imprecise description of the failure to report nodules where multiple examples exist in the same image,
- 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 the last of these observations to be particularly important and believe this to be the first recorded evidence of the effect.

ACKNOWLEDGEMENTS.

We wish to thank the radiologists, radiographers and students, all of whom are associated with the Department of Radiography and 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

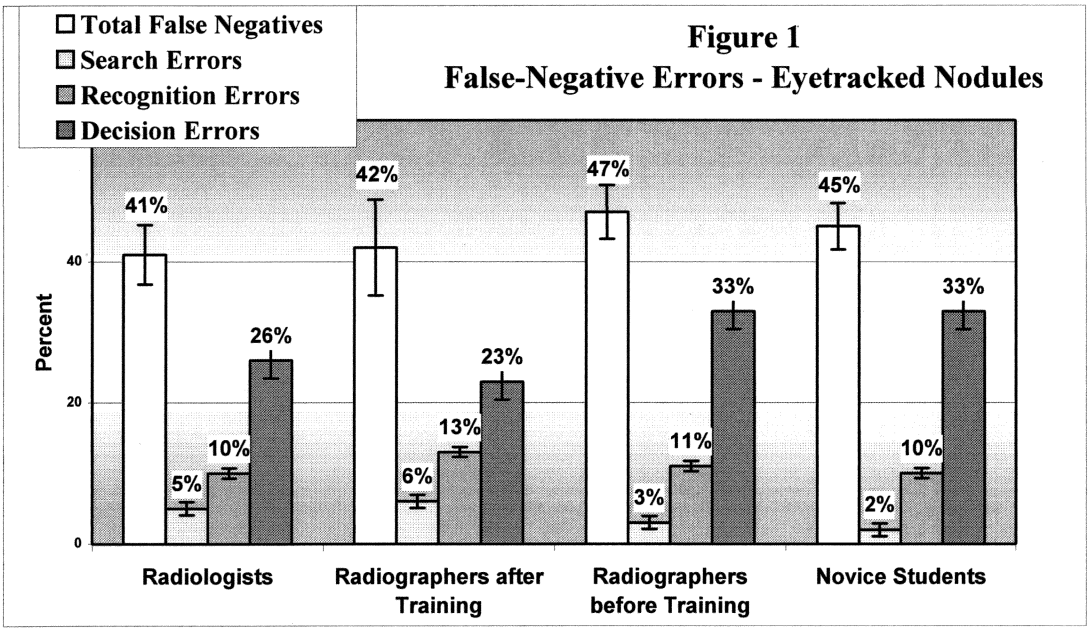
1. Manning D, Ethell S, Crawford T. An eye-tracking AFROC study of the influence of experience and training on chest x-ray interpretation. *Medical imaging 2003; Image Perception Observer Performance and Technology Assessment*. Editors Dev Chakraborty and Elizabeth Krupinski, Proc SPIE 2003 vol. 5034:257-266
2. Chakraborty DP, Winte LHL. Free response methodology: alternate analysis and a new observer performance experiment. *Radiology* 1990; 174: 873-81.
3. Fitzgerald R. Error in Radiology. *Clinical Radiology* 2001; 56: 938-946.
4. Chief Medical Officer. Learning from failure: evidence and experience. *An Organisation with a Memory*. London: Stationery Office, 2000: 1-7.
5. Manning D, Ethell SC. Insights into sources of error in diagnosis of lung cancer from chest radiography. *Proc MIUA 2002*, University of Portsmouth UK. Medical Image Understanding and Analysis 2002.
6. Manning DJ, Ethell SC, Donovan T. Detection or decision errors? Missed lung cancer from posteroanterior chest radiograph. *Br J Radiol* 2004; 77: 1-5.
7. Berbaum KS, Dorfman DD, Franken EA Jr, Caldwell RT. Proper ROC analysis and joint ROC analysis of the satisfaction of search effect in chest radiology. *Acad Radiol* 2000; 7: 945-958.
8. Eng J, Mysko WK, Weller GE, Renard R, Gitlin JN, Bluemke DA, Magid D, Kelen GD, Scott WW Jr. Interpretation of emergency department radiographs: a comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display. *Am J Roentgenol.* 2000; 175: 1233-8.

Table 1

	Radiologists	Radiographers Post-training	Radiographers	Novices
Search Errors	12%	14%	7%	4%
Recognition Errors	24%	32%	23%	22%
Decision Errors	64%	54%	70%	74%
Overall FN Rate	41%	42%	47%	45%

Table 2

	Missed Nodules in Multiple-Nodule Images			(SOS Errors)
	Radiologists	Radiographers after Training	Radiographers before Training	Novice Students
% of Total False Negatives from SOS	18%	23%	25%	21%
Not Fixated	2.30%	1.75%	1.92%	1.70%
Fixated > 200msec	15.66%	21.25%	23.08%	19.30%


Total False Negatives
Search Errors
Recognition Errors
Decision Errors
Figure 1
False-Negative Errors - Eyetracked Nodules
Percent
0
20
40
41%
5%
10%
26%
42%
6%
13%
23%
47%
3%
11%
33%
45%
2%
10%
33%
Radiologists
Radiographers after Training
Radiographers before Training
Novice Students

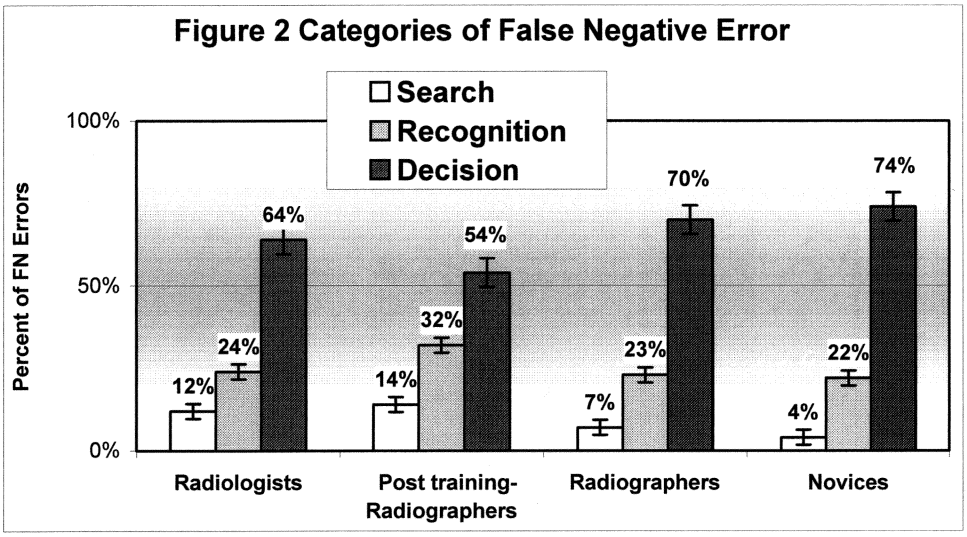
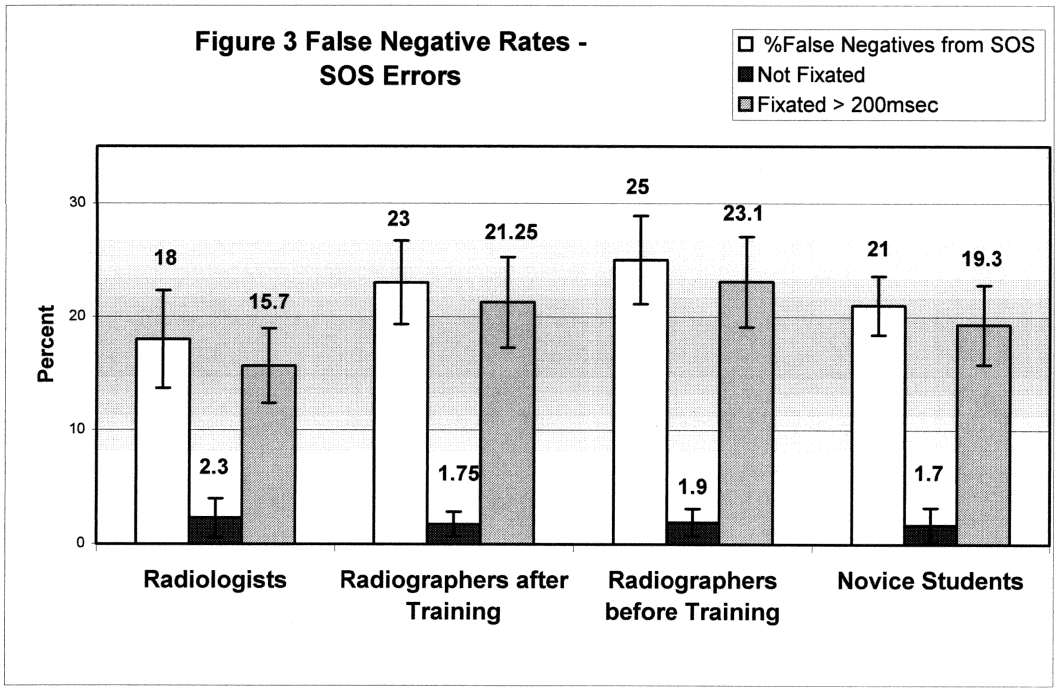


Figure 2 Categories of False Negative Error
Search
Recognition
Decision
Percent of FN Errors
100%
50%
0%
12%
24%
64%
14%
32%
54%
7%
23%
70%
4%
22%
74%
Radiologists
Post training-Radiographers
Radiographers
Novices



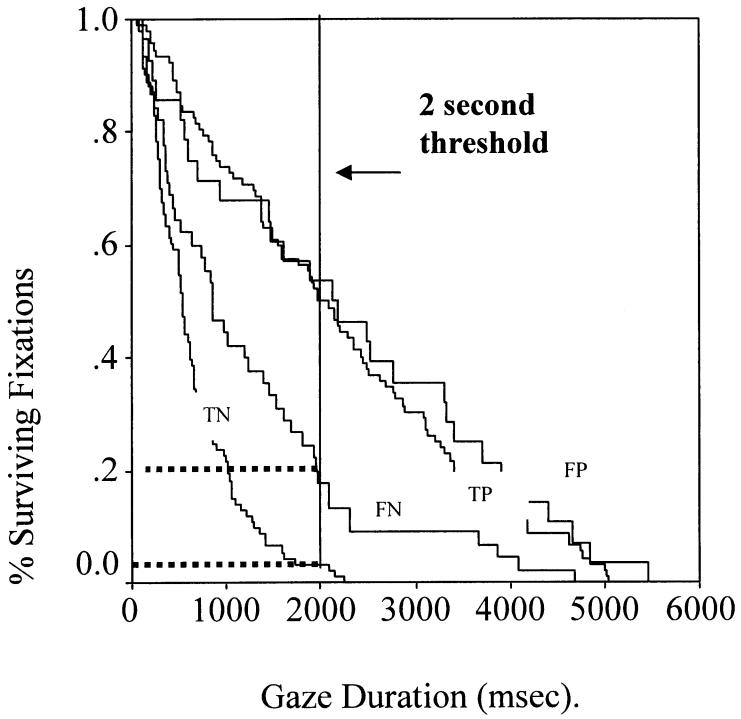


Figure 4 Time-Related Decisions for Radiologists.
After 2 seconds scrutinising a feature a negative decision was more likely to be incorrect than correct. For the radiologists shown here there was a 87.5% chance of such a decision being wrong. These are faulty cognitive decisions concerning genuine lesions and not failure of the observers to detect them during visual search.

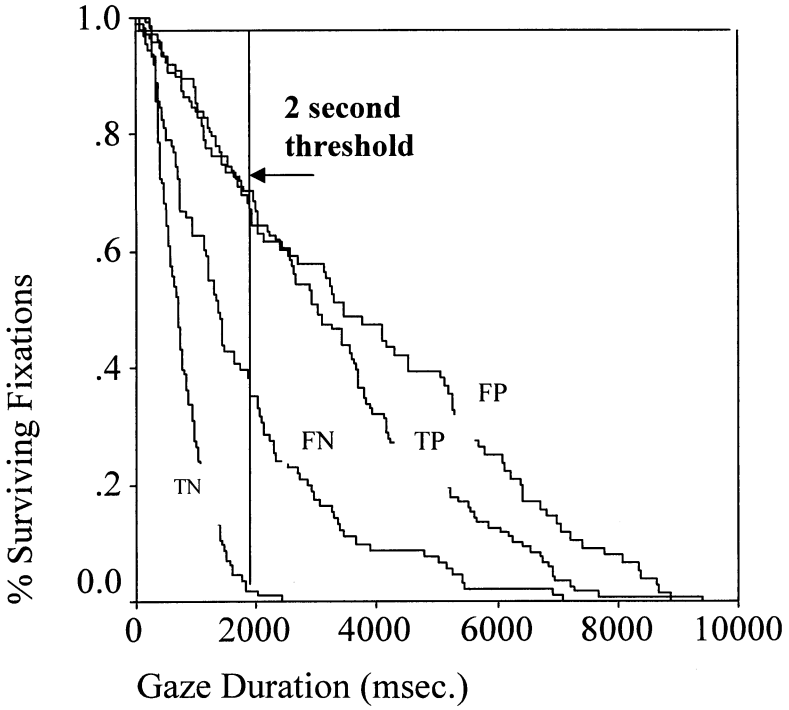


Figure 5 Time-Related Decisions for Radiographers Pre-Training
Time-related decision errors were similar to the radiologists in Figure 4 but for radiographers there was a 95% chance that a negative decision would be incorrect if made after 2-second gaze duration on a feature.

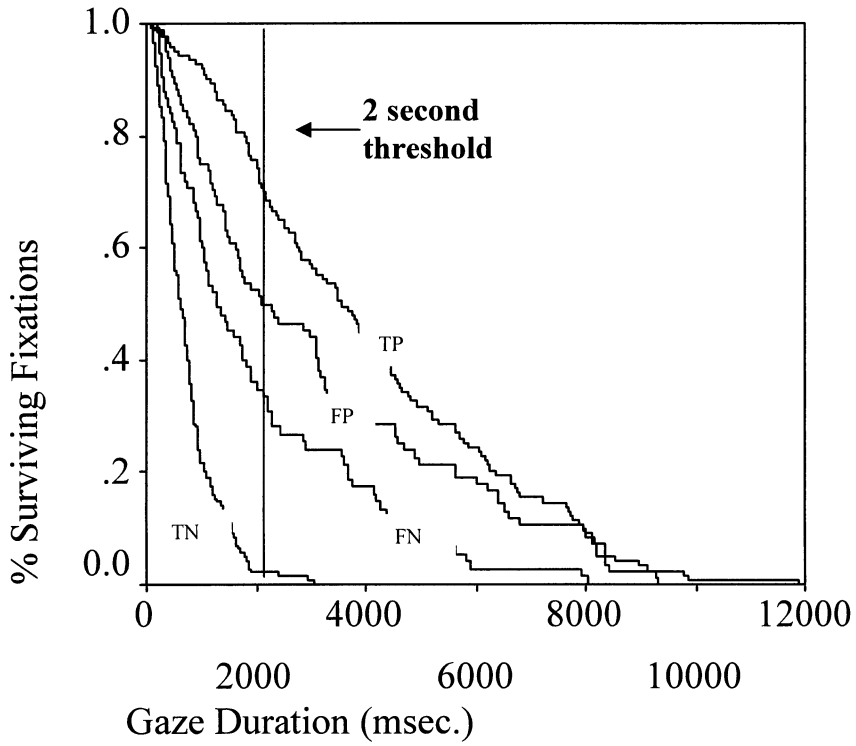


Figure 6 Time–Related Decisions for Novices
Novices show a similar trend to the experienced observers although their positive decision curves are more clearly separated and all their decisions extend further along the time axis. True /False positive curves are reversed compared with experienced subjects. 94% of negative decisions after the 2-second threshold are incorrect

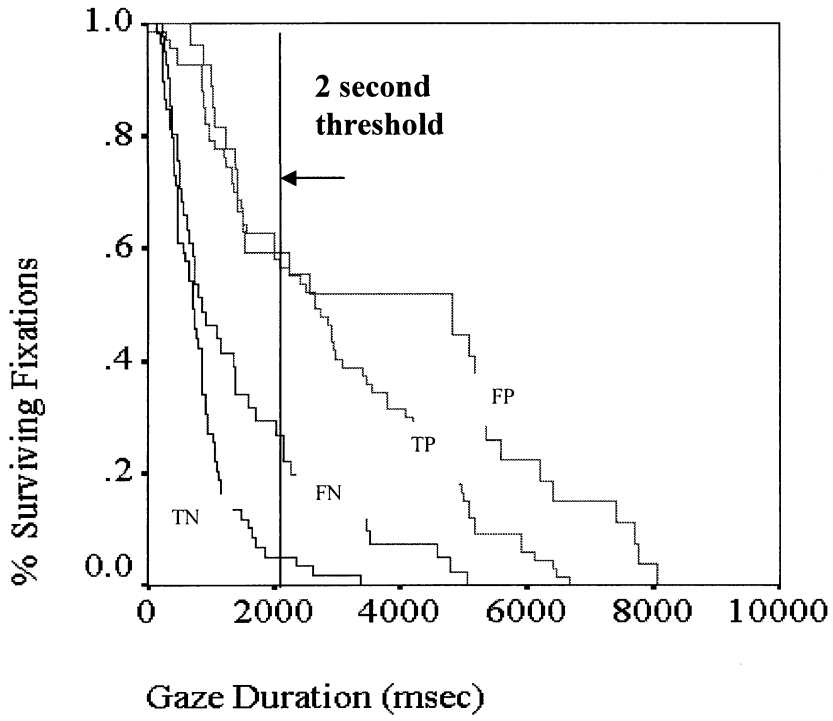


Figure 7 Time-Related Decisions for Radiographers Post-Training
After experience and training over six months and 500 cases the probability that a negative decision would be incorrect after the 2 second dwell threshold reduced to 84%. Compare with Fig 5.

View publication stats