Subjective responses to prompting in screening mammography

Subjective Responses to Prompting in Screening Mammography

Mark Hartswood[1*], Rob Procter[1], Linda Williams[2],Robin Prescott[2], Pat Dixon[1]

[1]Department of Computer Science, Edinburgh University, Edinburgh, EH9 3JZ
[2]Department of Public Health Sciences, Edinburgh University, Edinburgh, EH8 9AG

Abstract. We present the result of an experiment that examines the subjective responses of radiologists to a prompting system designed to assist with screening mammography. The results suggest that we should re-conceive our notions about the value of False Positive (FP) prompts. We conclude that the effectiveness of a prompting system operating at a given sensitivity is a function of the *types* of FP prompts produced.

1 Introduction

We are developing a computer-based system to analyse mammograms for signs of specific features associated with the early stages of breast cancer. For each one found, a prompt is produced and presented when the mammogram is subsequently read by a radiologist.

Experimental evidence suggests that prompting can improve human performance in visual search tasks by directing attention towards potential targets, but it was found that if the false positive (FP) prompt rate is more than 1.5 times the True Positive (TP) rate, then prompting ceased to be effective [3]. Since in screening mammography, the underlying cancer rate is approximately 0.5%, then given 90% sensitivity a prompting system would only be allowed 0.68 FP prompts per 100 cases, a combination of specificity and sensitivity far superior to a radiologist.

However, there are problems with extrapolating directly from these earlier results to the clinical setting. First, the test set was biased with respect to TP cases. Second, it is unclear whether the FP prompts were representative of the types of FP that a detection algorithm might actually produce. It is difficult to conclude whether the observed effect was due to the FT:TP ratio, or to overall prompting rates.

2 The Experiment

An experiment was designed to examine the properties of a prompting system under more realistic conditions, with the goal of determining an upper limit to the acceptable FP rate. Realistic reading conditions were simulated, including use of standard reporting forms and attaching reporting forms and prompt sheets to a film bag. Outputs from two feature detection algorithms being developed at the Royal Observatory at Edinburgh were used to generate prompts for microcalcification clusters [2] and ill-defined lesions [4]. Representative film sets were selected at random from four average days' screening at one clinic and balanced with respect to number of recalled cases, density of breast tissue and nodularity. There were two pathology proven malignancies in the set, treated as recalled cases for the purposes of randomisation.

The low proportion of malignancies, inevitable given the use of representative film sets, precluded the possibility of assessing the impact of prompting on radiologists' detection performance. The goal of this study was to investigate recall rates and radiologists' subjective assessment of the system under different prompting rates. The principal hypothesis was that radiologists' recall rates would not be influenced by the system prompt rate.

Prompt sheets consisted of a hard-copy, low resolution image of the mammogram pair with prompt information superimposed [5]. Prompts for ill-defined lesions consisted of an ellipse surrounding the suspect region, and for microcalcifications an irregular outline of the potential cluster (Figure 1). Prompt sheets were attached to reporting forms via a paper clip in such a way that a subject would have to lift the reporting form to examine the prompt sheet. A prompt sheet was produced for each case irrespective of whether that case was actually prompted or not.

* Author for correspondence, mjh@dcs.ed.ac.uk

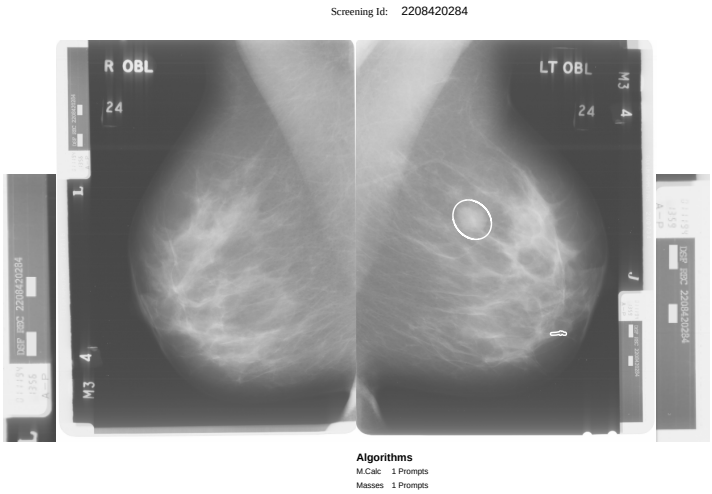

Fig. 1. Example prompt sheet

Sensitivity	Ill-defined lesions		Microcalcifications	
Condition	Prompt rate	Sensitivity	Prompt rate	Sensitivity
High	1/2	62%	1/3	94%
Medium	1/4	37%	1/6	86%
Low	1/8	22%	1/12	76%

Table 1. Average prompt rates for prompted conditions

The subjects were four experienced radiologists. The experiment consisted of four conditions, three were prompted at different rates, one was an unprompted control. Subjects were given an indication of the sensitivity of the algorithms for each condition (High, Medium or Low), they were also told the approximate prompt rate of each algorithm on a number of cases prompted basis (Table 1). Each condition consisted of 116 cases. The first five cases of each condition were used to familiarise the subjects with experimental procedure. The remaining cases were read in two sessions consisting of 56 and 55 cases respectively. There was a 15 minute break between these sessions. A Graeco-Latin square design was used to enable effects due to changes in prompt rate to be isolated from subject effects, session effects, and effects due to differences in the test sets. Each subject read each condition, but on different film sets.

The data recorded included recall rate and time taken to read each condition. Questionnaires were administered before and after the experiment and after each condition. A 20 point Likert test was used to assess subjects' attitudes to the system after each condition, with the higher the total score the more favourable the assessment.

3 Results

Wald Statistics for type 3 analysis of the recall rate showed no difference between the prompting levels at the 5% significance level ($p=0.061$). The principal hypothesis was therefore confirmed, with there being no increasing trend in recall rate as prompt rate increased. On the other hand, radiologist, reading order and film set were all significant contributors to the variation in the recall rate.

Figure 2 shows the results of the pre/post experiment questionnaire on the perceived value of prompting for particular types of benign feature. Subjects were asked to rate each feature type on a scale of one (useful) to five (distracting). A t-test of the results showed that subjects were significantly more likely to believe that prompting for benign features would be useful after the experiment than they were before it ($p<0.05$). The majority stated that they would prefer a system that was more sensitive (and obviously less specific) than themselves, but without prompts for obviously benign features.

The Likert test results in Figure 3 show that for three of the four subjects, scores increased monotonically, reflecting a more positive assessment of the system with increasing prompt rates. When making a recall decision, subjects were asked to indicate whether the relevant feature had been correctly prompted.

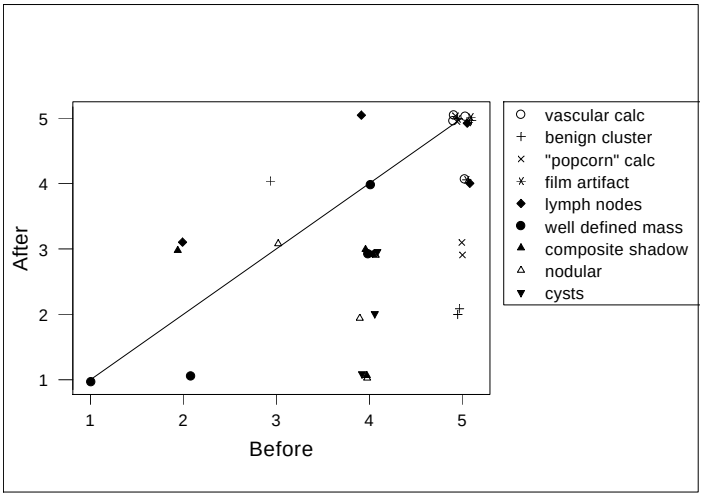

Fig. 2. Subjects' assessment of value of particular types of FP prompt (1 = useful and 5 = distracting) before and after the experiment.

Figure 4 shows the percentage of correctly prompted recalled cases for each condition against the Likert score for that condition. For the majority of subjects, a monotonically increasing Likert score is apparent as the number of correctly prompted cases in the set increases.

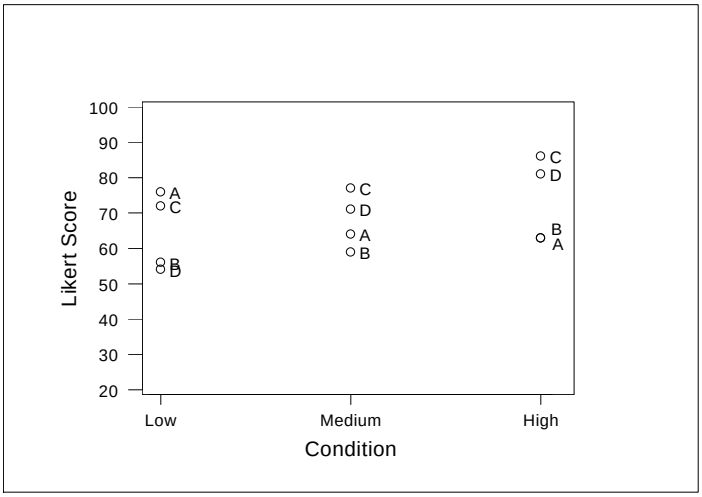

Fig. 3. Likert score against condition for subjects A to D.

4 Discussion

Our results indicate that when tested under realistic conditions, radiologists' tolerance level for FP prompts is appreciably higher than the upper limit established by Hutt for improved detection performance. Of course, positive subjective assessment may not necessarily coincide with objective performance effects, but we argue that our results point to the possibility that earlier work underestimates the FP prompt upper limit.

As there were so few true malignancies in the test sets, subjects were not expected to be able to

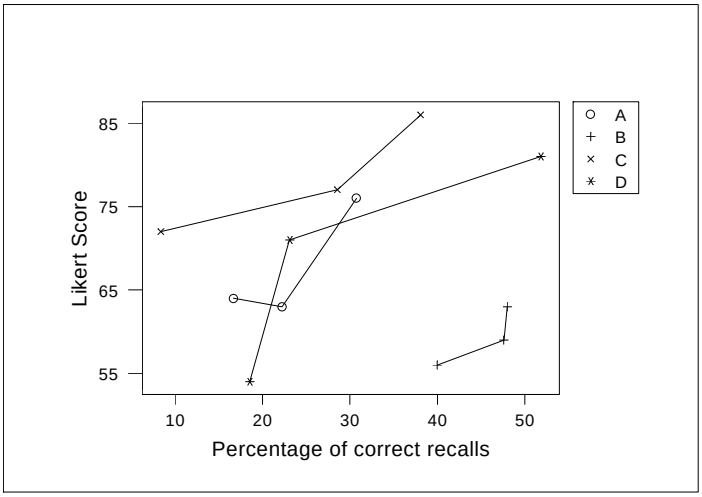

Fig. 4. Percentage of correctly prompted recalls against Likert score for each subject.

form an accurate picture of the system's capabilities. However, comments made both during and after the experiment showed that their assessment of the system's sensitivity was actually very acute. Figure 4 suggests that this judgement was informed by the proportion of recalled cases that were correctly prompted. We argue, therefore, that subjects' tolerance of FP prompts was due to the fact that they were informative of the system's behaviour.

We suggest that the effect of FP prompts will depend on their nature. When reading, radiologists consider a number of *candidate* features for recall, but only a proportion of these features result in recall, and only about 10% of recalled cases actually turn out to be cancers. We suggest that prompts for *candidate* features would be acceptable to radiologists in the clinical setting, whereas prompts for *other* features would not. The latter would be distracting, and contribute to the degradation in performance found in earlier work. In contrast, the former affords learning about — and positive confirmation of — the system's behaviour. It is our belief that this will be important for effective routine clinical use of such a system. In support of this, we have evidence of radiologists doing similar 'articulation work' for each other in double reading [1].

5 Conclusions and Further Work

The results reported here shed further light on the requirements for feature detection algorithms in breast screening. In particular, they suggest that the acceptable FP prompt rate is a function of the types of feature prompted, rather than the FP:TP ratio alone.

To explore this issue further, radiologists will be asked to rate prompts from *useful* through to *distracting* on a five point scale. This will enable us to classify prompts as *candidate*, *recall* or *other* features.

References

1. Hartswood, M., Procter, R., Williams, L. and Prescott, R. Drawing the line between perception and interpretation. To be published in Proceedings of Allocation of Functions Conference: New Perspectives, Galway, Ireland, October, 1997.
2. Hume, A., Thanisch, P., Hartswood, M. and Procter, R. On the evaluation of microcalcification detection algorithms. Proceedings of the Third International Workshop on Digital Mammography, Chicago, 1996.
3. Hutt, I. The Computer-Aided Detection of Abnormalities in Digital Mammograms. Ph.D. Thesis, Manchester University, 1996.
4. Miller, L. and Ramsay, N. The detection of malignant masses by non-linear multiscale analysis. Proceedings of the Third International Workshop on Digital Mammography, Chicago, 1996.
5. Procter, R., Thanisch, P., Astley, S. and Hutt, I. User interface design and data management for digital mass mammography. Proceedings of the Second International Workshop on Digital Mammography, York, 1994.