Prompting in mammography: Computer-aided Detection or Computer-aided Diagnosis?

Mark Hartswood\textsuperscript{a}, Rob Procter\textsuperscript{a} and Linda J. Williams\textsuperscript{b}

\textsuperscript{a}Department of Computer Science, Edinburgh University, Edinburgh, EH9 3JZ
\textsuperscript{b}Department of Public Health Sciences, Edinburgh University, Edinburgh, EH8 9AG.

Abstract. This paper addresses radiologists’ use of Computer-Aided Detection systems in screening mammography. Our focus is on how radiologists interpret prompting information and how this interpretation subsequently affects their decision making. Generally a distinction is made between systems designed to assist radiologists make a more complete examination of a mammogram (detection aids) and those that assist a radiologist to distinguish between benign and malignant lesions (diagnostic aids). We present evidence to show that it is difficult for radiologists to maintain this distinction in practice. We suggest that radiologists are inclined to use prompts as evidence to support diagnostic decisions in cases where they are uncertain about the interpretation of a lesion. It is possible that this mode of use may have a detrimental effect on performance.

1 Introduction

The goal of a computer-aided detection system like PROMAM (PROmpting for MAMmography) is to reduce errors by drawing radiologists’ attention to possible abnormalities. PROMAM is not intended to be used as a computer-aided diagnosis tool: the decision as to whether a feature is of clinical significance remains with the radiologist [1, 2].

In practice, however, the distinction between detection and diagnosis may be blurred. One study has indicated that, for subtle microcalcification clusters, subjects’ confidence that a cluster was present was increased if the cluster was prompted, and decreased if the cluster was unprompted [3]. Another study reported that prompting can entail an increase in False Positive (FP) decisions without necessarily having an overall effect on confidence levels [4]. The first study would seem to indicate that radiologists’ confidence with respect to the detection task is affected by prompting, but that their diagnostic decision making remains largely unaffected. The second study, however, raises doubts regarding the latter conclusion.

We have recently completed a small-scale trial of PROMAM and have used this opportunity to explore further the effect of prompting on radiologists’ recall decisions under clinical, rather than laboratory conditions. Our results suggest that radiologists are inclined to use the information supplied by a detection system as evidence to support diagnostic decisions in cases where there is some ambiguity about the interpretation of a lesion.

2 Procedure

Five subjects were recruited from radiologists at a Scottish breast screening centre. Two thousand archive cases (including 102 pathology proven cancers) were digitised and analysed by the PROMAM system. The system performance was as follows: microcalcification sensitivity 93.8%, FP rate of 0.54 cases prompted; mass sensitivity 72.9%, FP rate of 0.66 cases prompted [5]. The films were then divided into twenty sets of approximately one hundred films each and double read, once by a subject in a prompted condition and once by a subject unprompted. Constraints on subject availability meant that it was impossible to ensure that subjects read the same number of prompted as unprompted conditions. In the prompted conditions, subjects were asked to first examine the films, then examine the prompt sheet, and then to record their decision i.e. recall or normal.

Subjects were trained in the use of PROMAM prior to participating in the trial [6]. In particular, they were instructed that they should not use prompts as contributory evidence in their recall/normal decisions.

In addition to subjects’ recall/normal decisions, data was also collected through post-session interviews to explore how subjects used the prompts, and pre- and post-trial questionnaires.
3 Results and Discussion

In each of the post-prompted session interviews, subjects were asked if the prompts had some influence on their recall decisions. Out of a total of sixteen interviews held after prompted sessions, subjects indicated that their recall decisions had been affected one or more times in a total of eleven of those sessions.

3.1 Aiding detection

In ten interviews subjects reported that on one or more occasions during that session their attention had been drawn to features that they had overlooked. These events fall into two subcategories: (1) features that subjects had failed to detect, which they then decided were normal, and (2) features that subjects had failed to detect, which they then decided to recall. There were several reported occurrences of category (1) events. For example:

“Yes, there were a couple of cases, I think they were calcs and they were unaltered from previous.”
(Subject A)

The incidence of category (1) events might seem low given that the majority of missed features brought to the radiologists’ attention are likely to be of this type. However, these events might be under-represented as they are possibly ‘less interesting’ to subjects than missed features that resulted in a recall. There were also several reported occurrences of events in category (2). For example:

“Yes, one, on micro-calcifications … that I didn’t see and then I brought back.” (Subject E)

Apart from drawing attention to features that may have been missed, prompts may influence radiologists’ visual search patterns by encouraging them to take another look at prompted features. In the post-session interviews, several instances of this were noted by subjects. For example:

“There were cases where it made me look again, I don’t think it actually made me change my mind. But it did make me look back again.” (Subject B)

3.2 Aiding diagnosis

Despite the instructions given in pre-trial training, both questionnaire data and responses given in post-session interviews indicate that subjects were inclined to use prompts to aid diagnosis. Subjects referred to occasions where they had found the absence of a prompt ‘reassuring’. For example:

“Yes, yes, I think that that is reassuring. It might just be falsely reassuring sometimes.” (Subject B)

The quotes above indicate that the absence of a prompt is viewed as ‘reassuring’ only, merely confirming a decision that has already been made. However, subjects also reported cases where the presence of a prompt had seemingly made them more inclined to recall. For example:

“There was one where I was undecided, and it was prompted… ‘I will bring it back, yes’ … otherwise I probably would have said ‘oh, forget it’, whether that’s right or not I don’t know.” (Subject B)

Overall, subjects’ comments suggest that the presence or absence of a prompt is most likely to influence a decision when the evidence available from the image alone is ambiguous. It is possible that in these situations radiologists will attempt to use whatever evidence that is to hand, including prompts, to resolve any ambiguity:

“Maybe it was highlighting something that I wasn’t seeing in a dense breast, so that’s why it needed confirmed. Erm … I (?) with it you go with the prompt.” (Subject E)

One subject drew an analogy between heightened suspicion when another radiologist asks her to examine a case, and when a case is prompted by a computer system:

“…it’s like when someone shows sets of mammogram and they’ll say, you know, it’s always nice for someone not to say, point out what they are worried about, because if you do, then immediately you heightened suspicion because someone else is suspicious about it.” (Subject E)

In pre- and post-trial questionnaires subjects were asked to rate their agreement with the following statements: (a) the presence of a prompt will make me more likely to recommend recall; (b) the absence of a prompt makes me less likely to recommend recall on a five point scale (‘Strongly agree’, ‘Agree’, ‘Uncertain’, ‘Disagree’, ‘Strongly disagree’). The results are shown in Figure 1 (a) and (b) respectively.
Figure 1. (a) the presence of a prompt will make me more inclined to recommend recall (b) the absence of a prompt will make me less likely to recommend recall.

Responses to the first statement show little difference between subjects’ pre- and post-trial opinions, with only one subject changing their opinion from ‘Uncertain’ to ‘Agree’. This is perhaps not very remarkable — if there is uncertainty in diagnosis, it might be expected that the default position would be to recall. Responses to the second statement indicates that there is a change of opinion post-trial, with subjects being more likely to believe that the absence of a prompt might influence their recall decisions. This is not consistent with the assumption that, under uncertainty, a recall decision is likely by default. However, responses to both statements are consistent with the conclusion that prompts are being used to aid diagnosis.

The reliability of data based upon self-reporting assumes that subjects are aware of their thought processes. This is most likely in instances where the prompts had caused — or had been used to inform — conscious deliberation about the status of some feature. The most obvious examples of this would be if a subject had overlooked a feature that the prompt subsequently brought to their attention, or if the presence (or absence) of a prompt had otherwise made some significant contribution to their decision to recall. However, it is also possible that the prompts may affect decision making in ways that are not available to introspection, and therefore in ways that might go unreported in response to questions posed during interviews. In addition, the accuracy of subjects’ responses to interview questions will depend on their ability to take a dispassionate and objective view of their own behaviour. Subjects might be inclined to underrate the effect of the prompts if they believe that any effect is at odds with the integrity of the objective application of their skill. Conversely, they might be inclined to overrate the effects of the prompts if they believe that this outcome is of particular interest to the person conducting the interview.

By comparing unprompted and prompted recalls, it is possible to gain a more objective view of the influence of prompts on subjects’ recalls. In prompted conditions in the trial, subjects had been asked to record if a correct prompt was given for the significant feature in each case they recalled. This information was not available for those cases recalled by the unprompted reader alone, so a follow-up exercise was devised to determine which of these recalls had actually been correctly prompted.

Prompt sheets for unprompted reader alone recalls were initially examined by a member of the PROMAM team, and 43 cases that clearly had not been correctly prompted were eliminated. Eliminations included cases where there was no prompt on the side the recall had been made for, or where the prompt was quite obviously for a different feature, or in a completely different region of the breast. The remaining 53 cases were examined by a radiologist to determine the accuracy of the prompts.

Table 1 shows that for the case $P^+U^-$ 50.7% of recalls were correctly prompted, whereas only 32.3% of recalls were correctly prompted for the case $P^-U^+$. A Chi-squared test indicates that this result would not be expected.
if exposure to the system and the proportion of correctly prompted recalls were independent ($p=0.017$). Thus there is a greater level of agreement between subjects and PROMAM after the subjects were exposed to prompting information — implying that the prompts have had influence on decision making. This influence could be due to the prompted condition leading to the detection of a greater number of significant features that would have otherwise been overlooked. However, it is also consistent with our earlier conclusion that radiologists’ diagnostic decisions are being influenced by the presence or absence of prompts.

4 Summary and Conclusions

The aim of prompting systems is to draw attention to evidence that an observer may have overlooked. From our results, however, we conclude that prompts also influence radiologists’ recall decisions. Though only two subjects stated explicitly that they were using prompts to aid diagnosis, others hinted that this might be the case in answer to specific questions in the post-session questionnaires, and analysis of the correlation between prompts and recalls provided further corroboration for our conclusion. We argue that this is because the presence or absence of a prompt has a subtle effect on a radiologist’s confidence threshold when making a diagnosis, and that radiologists are not necessarily always aware of this influence.

The prevailing view is that systems that aid detection are designed to address a different problem than those that aid diagnosis [7]. However, our data suggests that it is difficult to draw such a clear distinction between detection and diagnosis aids: when radiologists are faced with a difficult diagnosis, they can be influenced by, or may make use of, whatever evidence is available. If radiologists are being influenced involuntarily this would make the task of correcting their behaviour more difficult. As this study demonstrates, simply instructing radiologists that they should not use prompting information to aid diagnosis is not in itself sufficient.

One way of reducing dependance on prompts for diagnosis would be to change reading practice so that decision to recall made before examining the prompts will automatically stand. This should effectively prevent the absence of a prompt from influencing a radiologist’s recall decision, thus mitigating the worst effects of using a detection aid to aid diagnosis. While seeming a relatively simple solution, problems of administration and compliance should not, however, be underestimated. Another approach would involve training to ensure that radiologists develop the best strategy for interpreting the prompts. Since it is possible that radiologists may be involuntary users of prompting information for diagnosis, a systematic approach to training is required. This would possibly involve evaluated reading sessions so they might be assisted in recognising the particular circumstances where the diagnostic influence of prompts is likely.

It is possible that the effects observed in our study may have only transient significance. Though our study was performed in realistic clinical conditions, its duration still falls far short of the time periods that would probably be necessary to observe user learning effects. For example, with access to pathology and interval data, radiologists may be able to adapt their behaviour over time to maximise the value of prompting systems.

References