Prompting in practice: How can we ensure radiologists make best use of Computer-Aided Detection Systems?

Citation for published version (APA):
Procter, R., & Karssemeijer, N. (Ed.) (1998). Prompting in practice: How can we ensure radiologists make best use of Computer-Aided Detection Systems? In N. Karssemeijer (Ed.), *Proceedings of the Fourth International Workshop on Digital Mammography* (pp. 363-370). Kluwer Academic Publishers.

Published in:
Proceedings of the Fourth International Workshop on Digital Mammography

PROMPTING IN PRACTICE: HOW CAN WE ENSURE RADIOLOGISTS MAKE BEST USE OF COMPUTER-AIDED DETECTION SYSTEMS IN SCREENING MAMMOGRAPHY?

M. HARTSWOOD, R. PROCTER, L. J. WILLIAMS[1]
Department of Computer Science,
[1] *Department of Public Health Sciences,*
Edinburgh University,
Scotland.

1. Introduction

PROMAM is a prompting system for mammography which aims to improve radiologists' detection performance by drawing their attention to possible ill-defined lesions and micro-calcification clusters.

Various approaches such as ROC methodology or McNemar's test (a paired binary response statistic) have been used to quantify the performance gains that might be achieved through the radiologist's use of such a prompting system [5, 6]. However, they tell us little about radiologists' understanding of the system, nor about how radiologists use the prompts to inform their decision-making. Our earlier studies of PROMAM's use have demonstrated that these factors may be critical to its effectiveness [2, 3]. In particular, we believe that it is important to:

1. ensure that the radiologists develop a correct understanding of the system's scope and function,
2. ensure that prompting information is being used appropriately, and
3. understand how radiologists' use of the system changes over time as they learn about its behaviour and adapt their reading procedures.

The goal of computer-aided detection systems like PROMAM is to reduce errors by drawing radiologists' attention to possible abnormalities. In operation, a prompting system delivers locational information for features it considers to be suspicious to be used as attention cues by radiologists. This view of what information is available to a radiologist from a prompting system — and how, in practice, radiologists use that information — may be overly simplistic. For example, in extended use radiologists are able to make an assessment of the system's abilities based on an appraisal of its performance [3].

In a recent small scale clinical evaluation of PROMAM's performance we collected interview and questionnaire data to address these issues further [4]. The results suggest that radiologists use prompting information not only as attention cues, but also to inform their decision-making where there is uncertainty in the interpretation of a lesion. Furthermore, we found that radiologists developed strategies to economise on the effort required to dismiss false positive prompts: (a) by anticipating where prompts were likely to appear, and (b) by making a judgement on the value of a prompt based on information in the prompt itself, rather than on the image content of the prompted region.

2. Methods

Five subjects were recruited from radiologists at a Scottish breast screening centre. Two thousand and two 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%, with 54% of cases falsely prompted; mass sensitivity 72.9%, with 66% of cases falsely prompted [6]. The films were then divided into twenty sets of approximately one hundred films 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.

Data collection methods included observation of all the experimental sessions. Subjects were interviewed and asked to complete a questionnaire immediately following the prompted sessions; the interviews were tape recorded and subsequently transcribed. Further questionnaires were administered prior to starting the experiment, and after each subject had completed all their allocated sessions.

3. Training

Our previous studies revealed that users of a prompting system assumed a level of interpretive sophistication similar to their own, and thus either misjudged the operational scope of the system, or were confused by apparent inconsistencies in the system's performance [3]. For example, one radiologist found it confusing that the system would only prompt one or two locations in cases where there was widespread benign calcification — a confusion that could have easily been avoided with a little knowledge of the clustering rules used by the algorithm.

In preparation for this trial we devised a prototype training package that included a description of algorithm function. The aim was to give radiologists an understanding of situations where the algorithm would produce true positive (TP) and false positive (FP) prompts. An explanation was also given of categories of lesion that the system might fail to detect — e.g., because of lesion size, appearance or location. The explanations were illustrated with a series of example cases.

As part of the training we also presented a model of 'best practice' for using the prompt information. In particular, we emphasised that prompts should be used only as cues to examine the prompted region, and that any decision as to a feature's clinical significance should be made solely on the evidence available from the film itself.

4. Impact on decision-making

In each of the post-prompted session interviews, subjects were asked if the prompts had had some influence on their recall decision. 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. Subjects reported a number of occasions where the prompts had drawn significant features to their attention which they had overlooked, sometimes resulting in a recall decision.

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 give assistance with classification decisions. 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 quote above indicates 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 uncertainty:

> "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 questions: (a) the presence of a prompt will make you more likely to recommend recall? (b) the absence of a prompt makes you less likely to recommend recall? on a five point scale ('Strongly agree', 'Agree', 'Uncertain', 'Disagree', 'Strongly disagree'). The results are shown in Figures 1(a) and 1(b) respectively.

Both Figure 1(a) and Figure 1(b) show that subjects' belief that the presence or absence of a prompt influenced their decisions to recall or not recall respectively, and is consistent with their interview comments.

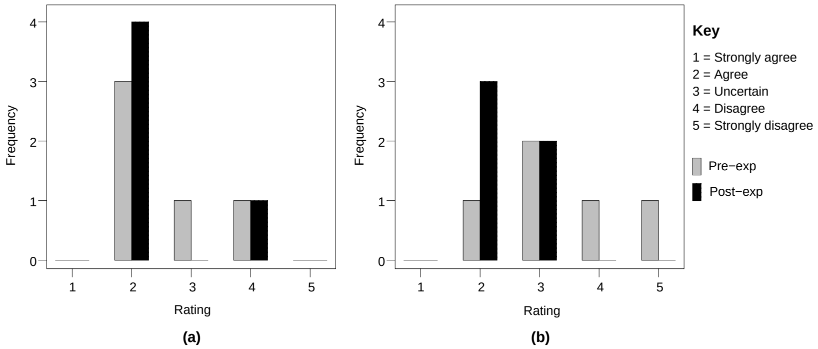

Figure 1. (a) the presence of a prompt will make me more inclined to recommend recall; (b) the absence of a prompt will make me less likely to recommend recall.

Data based upon self-reporting may be subject to various unconscious biases. By comparing unprompted and prompted recalls, it is possible to gain a more objective view of the influence of prompts on subjects' recalls. In the prompted conditions, 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 cases recalled only by the unprompted reader, so a follow-up exercise was devised to determine which of these recalls had been correctly prompted.

Prompt sheets for cases recalled only by the unprompted reader were initially examined by a member of the PROMAM team, and 43 cases that clearly had not been correctly prompted were eliminated. These included cases where there was no prompt, 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.

Recalled By		Correctly Prompted?		Total
Prompted Reader	Unprompted Reader	Yes	No	
Yes	No	35	34	69
No	Yes	31	65	96

TABLE 1. Correctly prompted recalls made by prompted and unprompted readers.

Table 1 shows that 50.7% of recalls in the prompted condition were correctly prompted, system, where as only 32.3% of the unprompted recalls had correct prompts. 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 when the subjects were exposed to prompting information, which implies that the prompts did have an influence on decision-making. This influence could be due to the detection of a greater number of significant features that would have otherwise been overlooked, but it is also consistent with the interview data showing that prompts influence classification decisions.

5. Dismissing prompts

Prompting systems typically have a poor specificity when compared with that of radiologists: effective system use depends on a radiologist's ability to easily recognise and dismiss FP prompts. The majority of the effort required to use a prompting system will be accounted for by this type of activity. Ideally, radiologists should give all prompts equal consideration, and only dismiss prompts after careful examination of the prompted region on the mammogram. However, interview data indicates that subjects develop strategies to determine the significance of system information based on an *a priori* assessment of the prompt sheet.

For example, subject D indicated that — under certain circumstances — the shape of prompts for vascular calcifications, and the location of prompts for ill-defined lesions, give a clue as to their cause:

> "I think now you'll start dismissing masses at the back, you're dismissing the calcification at the back and maybe you don't look as (. . . ?) carefully as maybe — you do look carefully but maybe not to the same degree when you clearly see that it is vascular calcification it's prompting on." (Subject D)

When asked if she was able to recognise what the prompts are for from her examination of the prompt sheet alone, subject B gave a similar response:

> "Yes, I mean, if it's the one particularly along the edge of the pectoral and the bottom, lower, inner aspects, yes . . . then the vascular calcification is one (. . . ?) those are very obvious, yes."

Subject E was also able to identify prompts for film artifacts in this way:

> ". . . the ones that happen so frequently at the bottom at the edge of the film, I was thinking that it would be awful if there was a lesion there one day because sometimes it's crying wolf at that point all the time . . . Because sometimes you don't even bother looking — you have a quick glance down . . . "

These comments indicate that subjects learnt to recognise patterns in shape, frequency and location that characterise FP prompts, and used this to determine how much effort they invest in further scrutiny of the mammogram. In such cases, consideration of possible explanations is not deferred until all the evidence has been gathered [1]. Subjects D and E, for example, indicated that they might not look back as carefully — or at all — depending on their initial assessment. While this lessens the overall burden of assessing FP prompts, there is a danger (as subject E remarked) of 'premature closure' — i.e., that TPs might go unnoticed if they happen to correspond with regions or prompt types that radiologists might learn to habitually dismiss.

6. Anticipating prompts

Subjects reported that they were often able to anticipate which features in the mammogram would be prompted, and that these predictions could be used to reduce the number of occasions that the mammogram had to be re-examined for FP prompts. Subjects seemed able to develop this skill relatively quickly, even after just one prompted session:

> "I think that I'm beginning to get so that I can guess what's going to be prompted for." (Subject C)
>
> "I sometimes look at the films and say 'I bet it's going to prompt for that'. . . " (Subject B)

In a later session, subject E volunteered an explanation of how this predictability is of use:

> "At times I'm definitely anticipating that that's going to be prompted. And sort of already decide I'm not going to look at it again almost, you know, you're kind of expecting prompts on certain things so I think you sort of, . . . very quickly dismiss it as (harmless?) without looking again."

Although the degree of predictability exhibited by the system was found to be useful, subjects stated that prompts were surprising as often as they were predictable. For example, subject D stated:

> "Sometimes you will actually be surprised what it is prompting, sometimes then actually you're surprised that it hasn't prompted something. There were one or two bits where I thought that it would have several prompts, (for?) masses, and it didn't actually, . . . getting zero, zero . . . But overall actually I think that you can anticipate some of the prompts, yes."

Subject B believed her predictions to be correct approximately 50% of the time:

> "I find myself sometimes thinking 'well, I bet it's going to prompt for that'. Erm, and that actually makes it easier, if the prompt is there then I can forget about that straight away. But sometimes, when it prompts something out of the blue, then there is nothing you can do . . . [I think I know what it's going to prompt for] about 50% of the time."

There is a cognitive cost associated with this strategy as it requires that radiologists must form a more accurate model of system behaviour. However, checking whether system output meets with expectations appears to be an intuitive reaction for radiologists, and probably essential for establishing and maintaining trust in system performance. We would argue also that anticipation is the better strategy because it implies that the radiologist has actually made an assessment based on the evidence in the mammogram.

The success of anticipation is dependent upon consistency of the prompting system as *perceived* by the radiologist. Image analysis algorithms can be sensitive to variations in appearance which are too subtle for the radiologist to appreciate without close examination — if at all. Though system behaviour may be *strictly* deterministic, it may not be *observably* deterministic if it doesn't respond in the same way to features that radiologists would classify as being similar.

7. Summary and conclusions

The goal of the training package developed for this experiment was to provide a useful account of how system function relates to mammographic appearance, and in particular to highlight circumstances where system behaviour might be counter-intuitive to radiologists. In this respect we believe that we were relatively successful. Our evidence suggests that subjects were able to use the training material to explain some of the prompts. There were also some unexpected outcomes, however, which suggest that training could be enhanced in a number of respects.

Subjects discovered categories of FP prompts that were not accounted for in training. This suggests that the training package be redesigned to provide not only a resource for initial familiarisation, but also to support the continued learning of clinicians and evolving practices. For instance, computer-based tools could be provided to enable radiologists to update and extend the training package with relevant cases drawn from their experience of using PROMAM.

Our investigations also show that radiologists used prompts in ways which were partly informed by training — and partly improvised — to economise on the effort required to deal with FP prompts. Future training must address this issue. In particular, an appropriate balance needs to be sought between making an *a priori* assessment of prompt significance, and carefully examining each prompted region. Our results indicate that analysis of a prompted area may sometimes begin with an interpretation suggested by some property of the prompts, rather than one suggested by some property of the image. In the training material we highlighted the value of attributes (e.g., location) for identifying some FP types (e.g., film artifacts). Our intention was to orientate radiologists to the task of interpretation by cueing candidate explanations. We did not anticipate that radiologists would use these properties to make *a priori* assessments.

In contrast, we believe that training should encourage the use of anticipation as a means of reducing effort since it motivates radiologists' to learn about system behaviour. In turn, these recommendations for use suggest goals for system enhancement: (a) FP types with regular characteristics should be targeted for elimination, and (b) more attention should be paid to the issue of observably deterministic behaviour — e.g., sensitivity to subtle variations in image properties. The latter would help radiologists to develop a more consistent model of system behaviour, and so enhance their ability to anticipate FP prompts.

The training package attempted to reflect our current understanding of best practice for prompted mammography: i.e., prompts should be used solely to aid detection, and not as evidence for interpretation. In this, it was less successful. Our results show that simply asking radiologists not to use prompts to assist with classification decisions is insufficient. One observed effect was the absence of a prompt being used to confirm a decision not to recall. It is possible that this use of prompts is involuntary, which suggests that a more systematic approach to training is required. This might take the form of evaluated reading sessions designed to encourage radiologists to recognise the circumstances in which this particular bias is likely to occur.

A much more rarely observed effect was the presence of a prompt alone being used as sufficient evidence to recall. This indicates that the scope of the system relative to radiologists' own abilities should be made clearer. The value of a prompting system is its perceptual thoroughness, rather than perceptual acuity — i.e., we have no evidence that it has the capacity to detect features that are beyond the perceptual capabilities of the radiologist.

The conclusions we have drawn from this small scale clinical evaluation are necessarily very provisional. Much has yet to be learnt about what constitutes best practice in using systems like PROMAM. So far, it has been system developers who have been cast in the role of experts, and instructing radiologists in PROMAM behaviour and use. Over time, however, as radiologists acquire greater observation-based knowledge of PROMAM behaviour, however, this balance of expertise will shift. As a result, radiologists may feel justified in departing from present notions of best practice: in clinical use, it is the radiologist community which must assume responsibility for its definition. We believe, however, that it is important that radiologists' observations should continue to be grounded in functional accounts of system behaviour. Continued close collaboration between radiologists and system developers is therefore essential to ensure that training materials evolve in line with practical experience.

8. Acknowledgements

We wish to thank the PROMAM team and staff at the South East Scotland Breast Screening Centre for their support and cooperation. This work is supported by Scottish Office Home and Health Department grant K/CSO/1/327.

References

[1] Gale, A. G. (1995) Human Response to Visual Stimuli. In Hendee, W. and Wells. P. (Eds.) The Perception of Visual Information. Springer-Verlag.

[2] Hartswood, M., Procter, R., Williams, L. and Prescott, R. (1997) Subjective Reaction to Prompting in Screening Mammography. In Taylor, C. et al. (Eds.) Proceedings of the First Medical Image Analysis and Understanding Workshop. Oxford, July.

[3] Hartswood, M., Procter, R., Williams, L., Prescott, R. and Dixon, P. (1997) Drawing the line between perception and interpretation in computer-aided mammography. In Bannon, L. et al. (Eds.) Proceedings of the First International Conference on Allocation of Functions. Galway, October. IEA Press, p. 275-291.

[4] Hartswood, M., Procter, R. and Williams, L. (1998) Prompting in mammography: Computer-aided Detection or Computer-aided Diagnosis? Submitted to the Second Medical Image Analysis and Understanding Workshop. Leeds, July.

[5] Hutt, I. (1996) The Computer-Aided Detection of Abnormalities in Digital Mammograms. Unpublished Ph.D. Thesis, Manchester University.

[6] Williams, L., Prescott, R. and Hartswood, M. (1998) Computer-aided cancer detection and the UK National breast screening programme. To be published in Karssemeijer, N. (Ed.) Proceedings of the Fourth International Workshop on Digital Mammography. Nijmegen, June.