Computer-Aided Mammography: A Case Study of Error Management in a Skilled Decision-making Task

Journal of Topics in Health Information Management, vol. 20(4), May, 2000, p. 38--54.

Computer-Aided Mammography: A Case Study of Error Management in a Skilled Decision-making Task

Mark Hartswood and Rob Procter, Institute for Communicating and Collaborating Systems,
University of Edinburgh, Scotland.

Keywords: breast screening, prompting aids, error management, ethnographic studies

Abstract

The practice of breast screening calls for radiologists to exercise a combination of perceptual skills to find what may be faint and small features in a complex visual environment, and interpretative skills to classify them appropriately -- i.e., as benign or suspicious. Radiologists make errors, however, and evidence suggests that these can be reduced by employing computer-based image analysis techniques to find, and prompt for, target features.

Computer-aided breast screening provides an interesting case study of error management issues raised by the introduction of computer support for a skilled decision-making task. Of course, since the rationale for computer-aided screening is that humans make errors, it is essential that system designers understand the nature of these errors. Equally, since the image analysis techniques themselves are not error free, it is important that radiologists understand the nature of the errors that the prompting system makes. More generally, this requires that radiologists be able to account for the behaviour of a technically complex system.

To investigate these issues, we have performed an ethnographically based study of clinic and reading practices, and controlled studies of the effects of prompting on radiologists' performance. These studies have enabled us both to understand better how radiologists approach the problem of error management in current practice, and their needs with respect to error management when using computer aids. Our results also show that the ways in which a computer-based tool actually gets used may be quite different from what was originally envisaged by its designers.

Finally, we outline what may be required in terms of training and practice to ensure the safe and effective use of computer prompting systems in screening applications.

Introduction

Breast cancer is the commonest form of cancer in the UK. Each year there are about 24,000 new cases and 15,000 deaths from the disease, accounting for one-fifth of deaths among women from all forms of cancer. Mammography (radiological imaging of the breast) remains the only method of detecting early stages of breast cancer, and preventative breast or mammography screening programmes operate in many countries.

We have been working as members of a team which is developing PROMAM (Prompting for MAMmography), a computer-aided mammography system designed for use in the UK breast screening programme (refs. 9-11, 20, 26). PROMAM is a prompting system which aims to improve radiologists' detection performance by drawing their attention to possible ill-defined lesions and micro-calcification clusters, and so reduce errors.

Prompting systems are designed to improve observer performance in visual search tasks by employing image analysis techniques to highlight areas that the observer should examine. In principle, observer errors arising from inattention, fatigue, etc. can be reduced. However, the practical realisation of improved observer performance is not easy. First, prompting systems are not infallible, so observers must be able to recognise system errors if their performance is not to be adversely affected. Second, the introduction of such a system in e.g., breast screening, may involve changes in work practices which reduce the overall effectiveness of error management procedures in the workplace. Investigating these issues has been a key part of our contribution to the PROMAM project.

We begin with an overview of the UK breast screening programme, followed by a review of the literature on visual search errors and the role of prompting as an aid to improved observer performance. Next, we present the results of field studies of breast screening work. Our focus here is on how radiologists manage their work both individually and collaboratively, and formally and informally, so as to minimise errors in their performance. We then present the results of a study of radiologists using the PROMAM prompting system, with a special emphasis on how they make sense of its behaviour and the role that this sense-making has in the management of system errors. Finally, we consider the implications for prompting systems on the informal mechanisms for error management.

Screening Mammography

The basic principle behind mammography as a screening test is that signs of malignancy are sufficiently distinct to achieve a reasonable sensitivity and specificity. Types of feature that are indicators of malignancy include:

> **Microcalcifications** are small deposits of calcium visible on a mammogram as tiny bright specks. They can be due to benign processes: e.g., it is common for vessels to calcify, giving a characteristic 'tram line' appearance on the mammogram. Small clusters of calcification can be indicative of early breast disease. Typically, the number, shape and distribution of calcifications within a cluster are used to determine the likelihood that they are the result of a malignant process.
>
> **Ill-defined lesions** are areas of radiographically-dense tissue appearing as a 'bright patch' on the mammogram that might indicate a developing tumour. Typically, lesions that are well-defined are the result of benign processes: e.g., they may be cystic. Lesions that do not have a well-defined edge are considered suspicious.
>
> **Stellate lesions** are visible as a radiating structure with ill-defined borders. The radiating components (or spicules) are the result of malignant processes infiltrating the breast tissue.
>
> **Architectural distortion** may be visible when breast tissue around the site of a developing tumour contracts. In the absence of other signs this might give a subtle clue to the presence of a tumour.
>
> **Asymmetry** between left and right mammograms may be the only visible sign of some hard to detect features. Asymmetry can be difficult to interpret as there is often a natural asymmetry in the distribution of breast tissue.

The UK Breast Screening Program (UKBSP) is a national service with a regional organisation. Each region is served by a number of screening clinics, each with two or more radiologists. The initial screening test is by mammography, where one or more X-ray films (mammograms) are taken of each breast by a radiographer. Each mammogram is examined for evidence of abnormality by two experienced radiologists. Women between the ages of 50 and 64 are invited to attend a clinic for screening every three years.

The goal of screening is to achieve a reliable and controlled cancer detection rate. Two performance parameters are particularly important: specificity and sensitivity. A high specificity -- i.e., low false positive (FP) rate -- means that few women will be recalled for further tests unnecessarily; a high sensitivity -- i.e., a high true positive (TP) rate -- means that few cancers will be missed. Achieving high specificity *and* high sensitivity is difficult.

The radiologists' task is a difficult one, not least because the small number of cancers is hidden amongst a large number of normal cases. It is a task which demands a high level of perceptual and interpretative skill: under certain circumstances normal tissue can have an abnormal appearance -- and vice versa. Figures from the UKBSP show that in the prevalent round (first screening visit) 6.4% of women screened are recalled for assessment, and in the incident round (later screening visits) this figure falls to 3.0%. More cancers are detected in the incident round -- 6.3 per thousand, compared with 3.4 per thousand in the prevalent round (ref. 2). FPs are not life-threatening errors, but they do cause stress and anxiety for those women who are recalled unnecessarily, and they waste resources. Far more serious are interval cancers, i.e., false negative (FN) errors, which are life-threatening.

The UKBSP is continually investigating ways of reducing FP and FN errors. For example, in many clinics, current practice involves each mammogram being 'double read' (i.e., examined independently by two radiologists). In the past five years, interest has grown in the possibility of employing computer-based image analysis techniques to enable a single reader (radiologist) to achieve performance equal to that achieved by double reading. The goal of the PROMAM project is to develop and apply these techniques to a prompting system.

Errors in mammography: Psychological approaches to decision-making have been widely used to provide a conceptual framework for understanding observer performance in radiological search tasks such as screening mammography. There are two basic approaches. One examines the nature of perceptual skill, the other is concerned with cognitive reasoning.

Studies of perceptual skill as applied to radiological tasks often involve eye-tracking studies or ROC (receiver operator characteristic) methodologies. Eye tracking involves using a mechanical apparatus to determine where a reader's gaze is directed when examining an image. Using this approach, Kundel, Nodine and Carmody (ref. 14) classified three types of error that can result in a FN decision for radiological search tasks. These are search errors, detection errors and classification errors. Search errors occur if the lesion does not enter the radiologist's 'useful field of view'. Detection errors are said to occur if the visual dwell time for an unreported lesion falls below some empirically determined threshold, where as classification errors occur if visual dwell time exceeds this threshold. Savage, Gale, Pawley and Wilson (ref. 21) define search and detection errors as occurring where a radiologist fails to report the presence of a lesion, and classification errors where the lesion has been reported, but inappropriately acted upon.

Studies of cognitive processes involved in radiological search tasks tend to be concerned with the radiologist's orientation to, and interpretation of, an image. Gale (ref. 8) posits a conceptual model whereby an observer selects from a pool of possible hypotheses according to their expectations about the mammogram. Each considered hypothesis is confirmed, or replaced in response to the information gathered by a visual inspection. Similarly, Lesgold, Glaser, Rubinson, Knopfer, Feltovich and Wang (ref. 15) equate the accurate interpretation of radiological images with the selection of appropriate schemata by drawing a distinction between the approach of novice and expert radiologists. They suggest that errors of interpretation made by novices can be explained in terms of both inefficient schemata selection, and the inappropriate testing of selected schemata.

Computer-aided Mammography

The prevailing view is that computer-based prompting aids are designed to address a different problem from those that assist with classification. The goal of the former is to improve sensitivity by cueing readers' attention to features that they may overlook, while the latter are designed to improve specificity by assisting the decision-making process (ref. 7).

The goal of a prompting system is to reduce the occurrence of search or detection errors. It is not intended that radiologists should attach any clinical significance to the presence -- or absence -- of a prompt. In contrast, classification aids offer an interpretation (either in terms of a probability value, or some explicit reasoning) designed to support a radiologist's judgement about the significance of a lesion already detected.

There are number criteria that have to be satisfied for a prompting system to be acceptable in the role of an attention cue in a screening environment. Current mammography image analysis techniques have a poor specificity compared with a radiologist -- prompting systems will therefore generate a relatively large number of FP prompts. The value of a prompting system lies in the fact that it can offer a complementary synthesis of system and radiologists' strengths: the former is more consistent in its visual search performance, the latter has interpretative skills which the system cannot match. If prompt system specificity is too low, however, radiologists will have to attend to too many FP prompts, and the costs of using the system may be perceived as outweighing its benefits. So, a prompting system must have sufficient specificity to maintain FP prompts at a manageable level. To achieve this, a prompting system must necessarily have some capacity to discriminate between features, if not to classify them.

Mammography image analysis techniques do not achieve 100% sensitivity, so prompting systems will also generate FN prompts (i.e., fail to prompt for some cancers). It is important that system and radiologists' FNs are independent. A correlation between system and radiologist FNs would preclude any improvement in overall sensitivity.

Finally, the prompting system approach assumes that prompting does not have an adverse effect on reader performance, i.e., that in attending to prompting information, no systematic bias is introduced into the radiologist's decision-making.

The PROMAM system: PROMAM's prompts consist of a hard copy, low resolution image of the mammogram pair with prompt information superimposed (ref. 20). The choice of this very 'low tech' prompt interface design was determined after an initial requirements investigation. Radiologists liked the simplicity of paper prompting interfaces which have, in addition, the virtue of fitting in easily with current reading practices; paper is handled routinely during the reading session.

The PROMAM system is capable of detecting microcalcifications and ill-defined lesions (refs. 12, 16). Prompts for microcalcifications consist of an irregular outline of the potential cluster. Prompts for ill-defined lesions consist of an ellipse surrounding the suspect region (see Figure 1).

An Overview of Breast Screening Work

An ethnographic style investigation of work practices in two Scottish and four English breast screening centres was conducted over a six month period. The centres are referred to here by the letters A through to F to preserve anonymity. Both observational and interview data were collected during a 2 month period of investigation at centre F, and during one week period in each of the other five centres. Each of the six centres studied had agreed to participate in clinical trials of PROMAM, and access for the purposes of this study was negotiated as a contribution to the on-going development of the PROMAM system. Thus centre selection was governed by suitability for clinical trials, rather than representativeness of screening practice.

Where data is presented, the mode of data collection is indicated (e.g., interview, field notes). Observations of reading sessions were conducted by asking the reader to indicate and explain their reasoning when they encountered something 'interesting' while reading. Where a comment or observation is attributed to the statement or activity of a reader, the reader is identified by a number and the screening centre by a letter (A-F). Thus FR1-C refers uniquely to a particular reader in centre C.

Breast screening practice in the UK has undergone important changes since the programme's inception. The autonomy given to individual screening centres has enabled innovation to proceed by the independent adoption of new practice. Practice innovations may be regularised (with the concomitant national resourcing implications) in the light of formal studies demonstrating clinical and/or cost effectiveness. An example of this is the adoption of two view mammography for incident round screens. The practice was initially adopted 'unofficially' by a number of screening centres in England, and a prospective study followed that demonstrated its effectiveness in reducing both FN and FP errors (ref. 23). This precipitated the adoption of two view mammography as standard practice by the UKBSP.

At the time of this investigation, the clinics studied were involved in more or less formal studies of practice innovations. For example, clinic E was examining the logistics of performing two view mammography in the second and third round by dedicating one of its mobile units to this procedure, and had just completed its involvement in a trial to study the efficacy of reducing the screening interval.

The practice of double reading represents an innovation of uncertain status. It has been adopted as standard practice in Scottish screening centres (ref. 4), and by a number of English centres (ref. 25). This situation is partly due to methodological problems in ascertaining performance gains attributable to double reading, resulting in widely varying estimates for its effectiveness (ref. 27), and also because of local shortages of trained readers (ref. 25).

Differences in practice may also emerge as a response to variations in local circumstances. For example, clinic A changed from a system of worst opinion recalls to a system of third reader arbitration because their recall rate became unmanageable. In contrast, centre D discontinued a policy of discussion of recall disagreements because individual readers held out for their own recall decisions. In centre E, only around 50% of cases are double read. This is due, in part, to the fact that two of its readers 'got used' to single reading during a period when they were the only two readers available. Centre E is also the only centre to operate a system of blinded double reading (where the opinion of the first reader is not available to the second reader).

In contrast with inter-clinic practice differences, intra-clinic practice is surprisingly homogeneous. This can be explained in part by how reading is organised. Many of the activities involving readers require co-ordination with other clinic staff (e.g., supervising clinics, attending meetings etc.), and so are difficult to re-schedule or to interrupt. In contrast, reading demands less commitment, and so can be more flexibly attended to. In consequence, reading tends to be organised around other activities; it is often

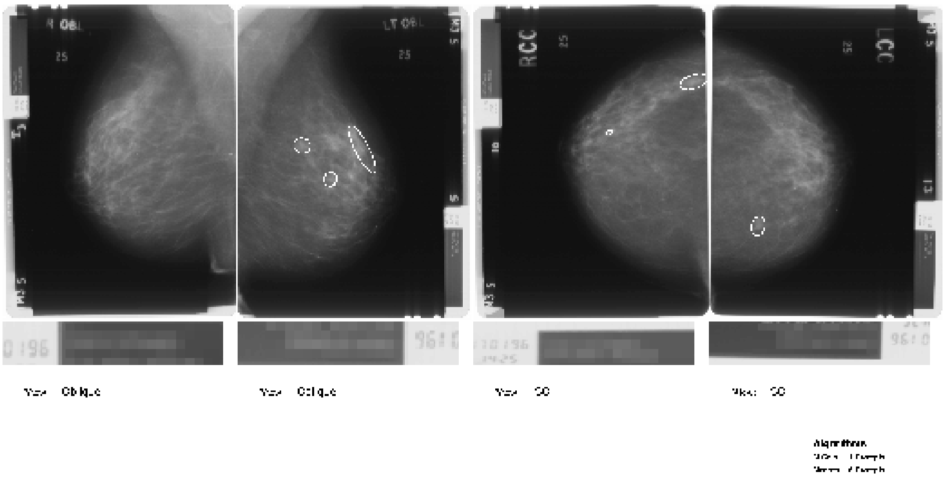

Figure 1. An example prompt sheet.

done ‘in a quiet moment’ or at the beginning or end of the working day. The availability a given reader at a given time often cannot be guaranteed, nor can the time they have available for reading be predicted. One consequence of this is that the artefacts used in reading have to be arranged in a way that is neutral with respect to who will be reading the films, leaving little scope for tailoring the selection and organisation of artefacts to suit individual preferences.

Of course, reading is not entirely commitment free. Obligations associated with reading include ensuring that women are informed of the outcome of screening in a timely fashion. The level of commitment associated with reading may therefore increase if there is a backlog of cases to be read. For example, in centre B cases may be single read if there is more than a week’s delay in reading.

Thus a combination of local circumstances and innovation can lead to inter-clinic variations, but the nature of screening work usually demands that within clinics homogeneous practice is established by consensus between readers.

Reading practices: Each mammogram is examined by at least one medically qualified reader who typically is also a trained radiologist. However, members of other medical specialities may also be employed as film readers if given the appropriate training, so in this paper the generic term ‘reader’ is used. On average, between 50 and 150 mammograms may be read in a single session. A reader will work through the cases on the viewer and mark his/her decision on the screening form. The decision of a reader may be one of:

> **Return to routine recall** When the reader decides that the case is normal the woman will be invited again for screening after a three year interval.
> **Recall for assessment** When a possible abnormality has been detected the reader will recommend attending an assessment clinic for further tests.
> **Technical recall** When the reader decides that a diagnosis may be inaccurate because of imperfect mammography then repeat screening films may be requested.

The degree of certainty about whether a feature indicates malignancy can vary considerably. Some are unequivocally malignant, whereas others might be only mildly suspicious. There are also various natural processes in the breast that can give the appearance of malignancy to varying degrees, and there are malignancies that are mammographically ‘occult’, i.e., they do not appear at all on the mammogram. It is common practice for reader to classify the features they find according to the probability that they indicate malignancy. For instance, at one clinic readers use a five point classification scale: C1 (normal), C2 (benign), C3 (equivocal), C4 (suspicious), and C5 (malignant), and set the recall threshold at C3.

Some readers mark each decision as it is made on the screening form, others defer marking decisions until they reach a case they wish to recall. In the latter case, intervening normal decisions are then marked as a 'batch'. The number of cases examined consecutively in this way will vary as recalled cases are randomly distributed. Some film readers might 'batch up' an arbitrary number of cases, rather than waiting for the next recalled case. If the cases are being double read, it falls to the second reader to ensure that the recalled cases are removed from the viewer before the normal cases are taken down.

Double reading involves the separate examination of each case by two readers, who each give their opinion. A final decision to either recall the woman for further tests, or to return the case to routine screening is made later by combining the decisions of the individual readers. In the clinics studied, three different strategies were employed for deciding the final outcome. Clinics B, D, E and F use a system of 'worst opinion recalls'. By this method, if either, or both, readers recommend that the case requires assessment, then the case is recalled. Centre A uses a system of 'third reader arbitration', where a reader not involved in the initial reading decides cases where there is disagreement. In centre C, disagreements are resolved by discussion between the two readers.

Where screening decisions are recorded on a paper screening form, double reading is generally done without any formal blinding of the second reader to the first reader's decision. This is because the screening form is routinely attended to as a source of various types of evidence, (e.g., HRT status, radiographers comments etc.) and so that the second reader may record his/her own decision. An exception to this is centre C, where the first reader's decisions are written on the back of the 'batch slip'. Although this method of recording the first reader's decision could easily serve as a blinding mechanism, a second reader was observed to examine the first reader decisions before reading the batch themselves, so apparently it is not always used in this way. However, double reading in centre C is viewed as serving a particular purpose, and this is discussed in more detail below. In centre E reading is blinded. This is facilitated by the use of an electronic system for recording screening decisions.

Comments made by readers suggest that they are alert to the possibility that access to the first reader's decision may bias the decision of the second reader:

> "Sometimes the second reader suppresses a potential recall on the basis that the first reader thought it was nothing. Therefore recalls go up with blinding." (Comment made while reading: field notes FR1-D)

Reader FR1-E stated that in a double reading team there is a tendency for readers to converge in their performance characteristics, maybe due to personality dominance. One reader suggested that the degree of influence that access to first reader decisions has might be a function of experience:

> "... and the fact that X had already first read it normal, you see that should not make a difference ... I think that perhaps did when I first started, but in (...?) I've been at it for a while so I've never thought it would, it shouldn't do." (Comments made while reading: transcript FR1-A)

There is some empirical evidence to suggest that this lack of independence can affect recall decision-making. Data taken from a published double reading study (ref. 27) indicates a strong relationship between the reported sensitivity of readers and the percentage of time they read second. One interpretation is that the second reader is 'prompted' by the first reader and thus picks up cancers that they would otherwise have been overlooked. However, this relationship might also be accounted for by inter-observer variation.

In the absence of a blinding procedure, readers may seek to maintain independence in their decision-making by employing strategies that decrease the accessibility of the first reader's decision. The simplest approach involves "trying not to look" (FR1-A) at the comments made by the first reader before making their own. Another approach is made possible by the practice of 'batching up' cases when reading:

> "When reading first, [I] maybe batch 7 or 8 films before scoring them, when reading second only batch three or four. This is because if the first reader has written something then have to go back and examine the films to see what they were referring to." (Comment made while reading: field notes FR2-A)

In centre D there is a particular incentive for reading first. Assessment clinics are organised so that the radiologist conducting the clinic is usually given cases recalled from batches for which they were the first reader. This provides an opportunity to receive feedback on screening decisions:

> "This makes assessment clinics more interesting. For example, if see something unusual, and don't know what it is, then get a chance for this feedback. [I] may not see anything similar for a number of years." (Comment made while reading: field notes FR1-D)

In contrast, two readers in centre A (FR1-6 and FR4-A) maintain an informal arrangement whereby they contrive to be first and second reader an equal number of times. Furthermore, when second reading one reader (FR4-A) reads the batch in reverse order. This is done under the assumption that readers are likely to be more fatigued towards the end of the reading session and so may be more likely to make mistakes on cases towards the end of the batch.

A double reading system involving discussion of recall decisions demands a greater degree of commitment from readers. Readers acknowledged its logistical difficulties, and also expressed concerns that explicit collaboration could bias decision-making. For example, when a reader in centre E was asked if they ever discussed cases, he replied that this was only done at the review session and at the interdisciplinary meetings. He stated that they were "worried about the effects of dominant personalities" (field notes FR3-E). A reader in centre B expressed similar concerns.

In clinic D, a system of discussing recall decisions had been in place, but this practice was discontinued. One reader suggested why:

> "[Discussion meetings] rapidly became a waste of time as each reader has a particular feature that they are able to detect well (patchy asymmetry, distortion, microcalcs are my own) and would hold out for recalls that they are convinced are something (usually falling into these categories)." (Comment made while reading: field notes FR1-D)

Published studies of double reading are typically concerned with its effects on the sensitivity and specificity of the screening test (refs. 1, 3, 5, 22, 24). Only one makes a passing reference to other potential roles (that it places an emphasis on teamwork (ref. 4)). However, in the centres studied, the practice of double reading appears to fulfil a number of roles that are arguably as important as any direct performance effects. These include:

1. a means of training film readers, and
2. to monitor, provide feedback and reassurance on reader performance.

Although there are acknowledged disadvantages to unblinded double reading, and readers appear to value their autonomy as decision-makers, the unblinded nature of double reading as typically practiced facilitates the achievement of the above roles. We will now discuss them in more detail.

Training: In clinic C double reading is used primarily as a mechanism for training. Typically a trainee will be paired with an experienced reader and disagreements about recall decisions are decided by discussion. For the purposes of training the potential for 'bias' inherent in a system that relies on discussion is actually desirable -- here the aim *is* to influence the decision of the trainee. Use of discussion enables the degree of autonomy given to the novice reader to be actively managed:

> "With the locum reading, the recall rate has gone up ... [I feel] that is important not to always override the decisions of junior readers as this can be a learning experience." (Comment made while reading: field notes FR1-C)

In centre C, an experienced reader (FR1-C) was observed to be reading second following a trainee. The trainee had flagged a case for recall, but had left a comment stating that the case was 'probably OK'. After examining the film the senior reader scribbled the request out, and the case was returned to routine recall (i.e., it was not removed from the viewer for discussion). Thus managing novice decision-making may be effected before the discussion stage is reached.

Centres B and E were also involved in training readers at the time of this study. Both centres employ a system of 'worst opinion recalls' and both have a similar policy of incrementally introducing novices into the reading process. Trainee readers initially attend a recognised training course. They may then spend a period of time reading films in the screening centre and discussing their opinions with experienced readers, but they do not at this stage contribute to recall decisions. Novices are introduced into reading proper as a first reader, and then as either a first or second reader as they gain experience. A number of reasons were suggested for limiting novice readers to reading first initially. These included:

- Providing a learning environment where the novice has to make decisions independently.
- So that any 'unnecessary' recalls can be stopped by the second reader.
- So that the second reader can act as a check and detect any missed abnormalities.

Thus training is organised to take advantage of the structure of double reading to provide a safe an supportive environment where novice readers can be encouraged to make independent decisions. The second reader is able to monitor and manage the novice reader's decision-making, and this also serves to provide a degree of reassurance that any cancers overlooked by the first reader may still be detected. One experienced reader from centre C suggested that she was "particularly careful when reading following a registrar [trainee]" (FR1-C).

Monitoring and feedback: One advantage of double reading is that the responsibility for decision-making is not shouldered entirely by a single reader. Thus, as one reader from centre C suggests:

> "Double reading can take away some pressure. People can have 'off' days." (Response to questionnaire: FR1-C)

Readers are often concerned to ensure their performance is consistent on a day by day basis -- that they are not unduly affected by fatigue, distractions etc. For example, after returning from maternity leave, one reader from centre E asked the clinical director for all the cases that she read to be double read (field notes: FR2-E). She sought to monitor her performance by comparing her cancer detection rate with others. She stated that after returning from maternity leave she 'missed' three that another reader had detected, but has only 'missed' one since then. She stated that double reading provides "reassurance" in these circumstances.

Another reader from centre E uses a similar mechanism to monitor his day to day performance. When reading second, he compares his decisions with those of the first reader to see if he has missed a lesion, or classified one differently. He states that when there is a difference of opinion then 3/4 of the time he has also seen the lesion and has dismissed it, and in the remaining 1/4 he has overlooked the lesion. Where there is disagreement, it is usually over less suspicious features, and that typically there is a large degree of agreement between readers over "actually malignant" features (those with a 4 or 5 classification). He recalled an occasion when he discovered (by checking the first reader recalls) that he had missed an "obvious" spiculated lesion (field notes: FR3-E).

The work of clinic staff is formally monitored through procedures for quality assurance and work documentation. Clinic staff hold regular meetings in which feedback may be given and received:

- multi-disciplinary pathology meetings where radiological appearance and pathology data are compared;
- review of interval cancers which may be evidence of FNs, and
- informal (and at some clinics, formal) discussion about differences in recall opinions.

Such meetings provide an opportunity for readers to articulate aspects of their work which they perform as individuals, such as their reasons for giving a 'recall' or 'no recall' opinion. Our studies suggest, however, that readers use double reading itself as a less formal mechanism for monitoring and feedback (ref. 10).

The first reader's opinion on each of a set of cases effectively provides a standard against which the second reader might compare his or her decisions. Even in centre C, where double reading is seen primarily as having a training function, one experienced reader commented:

> "... the two consultants like to read against each other as well as against the inexperienced radiologists." (Comment made while reading: field notes FR1-C)

Feedback gained by the second reader in this way may fulfil a number of functions. Readers can monitor their performance on a session by session basis and gain some reassurance that intra-observer variations are compensated for. This informal monitoring activity may also have a role to play in maintaining readers' recall thresholds within a manageable range by establishing and reinforcing normative interpretations. In two of the clinics visited it was evident that this informal monitoring has evolved further: first readers sometimes annotate the reporting form in cases they decided are normal (see Figure 2). In effect, this extends the set of cases over and above those recalled for which evidence about the reasoning of the first reader is available:

"Leaving messages for the second reader is useful -- to let them know that you've seen it -- the second reader might want to know whether you've seen it and what your opinion is." (Comment made while reading: field notes FR2-A)

A common (and arguably the simplest) type of annotation is to label a feature on the schematic by writing "Benign" or simply "B". Nothing is said about the reasoning behind the decision, indicating a tacit assumption that this will be readily apparent to other readers. Another common annotation is "BT" (Breast Tissue). Here some interpretation is offered: that the presentation of the feature is ascribed to normal breast tissue, but no reason for this ascription is given. Both these types of annotation appear to suggest that there is little doubt in the reader's mind that his or her opinion is correct, the annotation seems intended to reinforce this opinion and to demonstrate vigilance. On occasions, however, the use of "I think" and "?" is used in association with the description to express, and draw attention to, the reader's uncertainty.

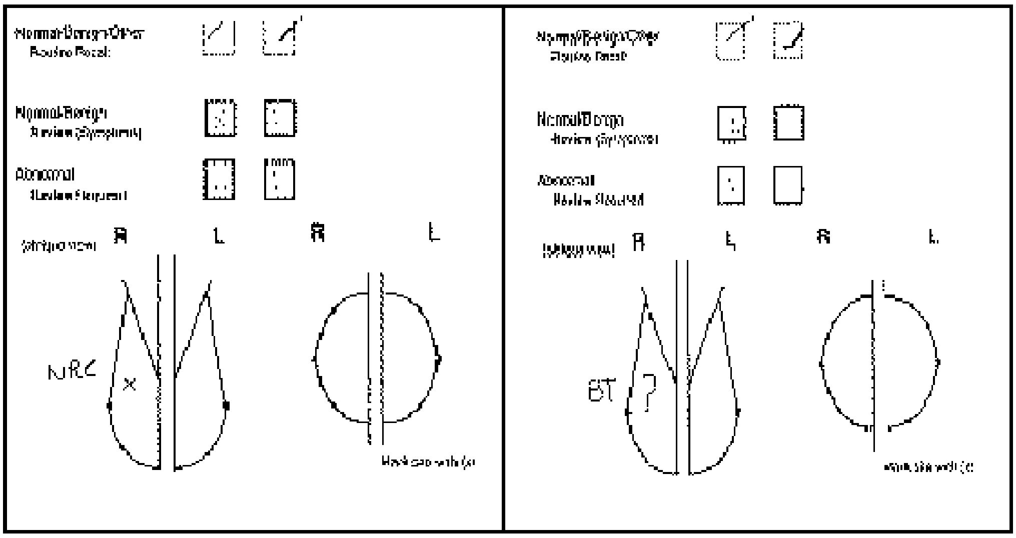

Figure 2. Examples of first reader annotations of benign features.

It is possible that annotation enables the first reader to assert their competence, and the second reader to assess the specificity of their decisions in the context of those made by their colleagues. In addition, annotations may be used to make inferences about the degree of suspicion in particular cases:

> "If the first reader flags a 'composite shadow' and the second reader goes straight past it, then it probably isn't significant." (Comment made while reading: field notes FR2-A)

The act of annotating a feature implies that the feature is worthy of annotation. That is, some characteristic of the feature appears to be sufficiently suspicious to warrant particular attention by the reader so that this suspicion may be discharged. Annotation may thus serve to demonstrate a reader's accountability to the decision-making process.

More complex annotations are also used. These typically make explicit information about a reader's reasoning by referring to the evidence used to mitigate the initial suspicion. Examples include: "Comp CC Ok" -- not visible in the CC view, so is a composite shadow; "NRC" (No Real Change) -- the feature has not changed over time, and thus is less suspicious. Readers were also observed to annotate changes they thought were due to different projections, new microcalcifications clusters that had a benign appearance, calcification clusters that hadn't changed and clusters of benign microcalcifications embedded in a background of vascular calcifications.

Agreement between readers is higher for recalled cases that actually turn out to be cancers, than for recalled cases that turn out to be normal. This 'virtuous' difference between readers' recalls accounts for the performance gains reported for double reading. In fact, double reading serves to compensate for both intra- and inter-observer variations (ref. 16). If these differences are too large then assessment clinics may be overwhelmed and changes in procedure may follow, like changing from a 'worst case' to a 'third reader' arbitration recall decision-making policy. It is possible that feedback from both assessment clinics and from first reader annotations may play a role in maintaining readers' recall thresholds within a manageable range. It is interesting to note that many annotations are for features that fall on the benign side of the recall threshold, that this is the region were most FP and FN decisions are likely to occur, and thus where differences in opinions are likely to have the greatest significance. Annotations are made where there is likely to be some uncertainty -- where decisions can be 'open to interpretation', thus articulation about such cases may serve to communicate and establish norms about the significance of particular kinds of presentation.

Using double reading as an informal or formal (if part of training) mechanism for obtaining information about a reader's performance may be useful, but readers also recognise the possibility that their judgements may be influenced by the first reader's decision. Thus there is a tension between the decision-making and monitoring aspects of a reader's work, where access to a first reader's decision is recognised both as useful as a metric of performance, but also as potentially harmful if it serves then to bias decision-making.

Summary: The results of this study suggest that readers are very aware of the extent and limitations of their expertise and apply these insights routinely in their work. Readers not only possess self-knowledge of performance in terms of measures like sensitivity and specificity in respect of particular feature types and circumstances, they also demonstrate a more general understanding of the psychology of the decision-making process. This understanding relates how particular conclusions are drawn from particular types of evidence, and suggests biases and errors to which a reader may be subject.

The results also show how, through an informal, emergent extension of formal monitoring procedures, readers may use each other to gain reassurance about their performance and to help themselves ensure that it remains within acceptable limits.

Studies of the Effects of Prompting

The key question about prompting in breast screening is what effect will it have on readers' performance? Owing to the low incidence rate of breast cancer, it is impossible to answer this question quantitatively without doing a large scale clinical trial (ref. 26). However, there are several issues that can be addressed using smaller scale investigations. These include: assuming that prompting system performance is not error-free, is there an upper limit to prompting errors -- FPs and FNs -- before a prompting system becomes useless and, if so, what level of error is tolerable by its users? A related question is, are some kinds of errors more tolerable than others? Finally, how do readers make sense of how the prompting system behaves and does this have a bearing on readers' assessment of its usefulness, and on how they use it?

Prompting information represents an additional source of evidence that a reader can draw upon when deciding a case. As with conventional sources of evidence, it is important to examine the possibility that use of prompting information may bias decision-making. It is also important to explore how access and interpretation might be managed to reduce any such effects.

In contrast to a human observer, a computer-based image analysis system will typically make use of only a subset of the available evidence, and will be limited in the ways in which it can access and combine evidence from different sources. Consequently, such a system is unlikely to match the performance of trained human observers in terms of both sensitivity and specificity, and will exhibit behaviours that might be considered naive by human observers. These limitations imply that a reader cannot use their knowledge of the behaviour of readers (as they do to interpret the decisions and annotations of colleagues) to reliably account for the behaviour of the system.

As part of PROMAM's development programme, we conducted three investigations designed to further understanding of how readers would make use of prompts under clinical conditions. The first was designed to elicit readers' subjective responses to system performance. Earlier work had suggested that prompting can improve visual search performance, but only if the FP prompt rate was no more than 1.5 times the TP rate (ref. 13). However, there are problems with extrapolating directly from such studies to the clinical setting. First, heavily biased test sets were employed and so the results may not be directly applicable to the circumstances in which reading is performed in the clinic. Second, it was unclear whether the FP prompts were representative of the types of FP that a prompting system might actually produce. Part of our programme of work has been to investigate these issues further.

The results of our first study indicated that, under realistic reading conditions, readers' tolerance for FP prompts was significantly higher than the upper limit established by Hutt (ref. 9). Subjects commented that they found FP prompts useful as an aid for making sense of the system's behaviour. We concluded that prompts for candidate[1] features may be acceptable in clinical use because they afford the development of their understanding of how the system works, and what its capabilities are (ref. 10).

Second study: To examine in greater detail readers' use of prompts, we devised a second study. All 144 prompted cases from sets 1 and 2 used in the first investigation were employed. In these sets, the ill-defined lesion detection algorithm prompted 155 features, and the microcalcification detection algorithm 188 in a total of 144 cases. A modified prompt form was devised to capture rating information and subjects' classification of each prompted feature. Three screening radiologists were recruited from a Scottish breast screening centre as subjects. Each subject examined the entire series of mammograms individually over two sessions, with 74 cases reported in the first session, and 70 in the second. The appropriate copy films were made available, and arranged sequentially on a Rad X style viewer. Subjects were asked to rate each of the prompted features in the following ways:

1. Indicate whether the prompt would be acceptable in a screening environment (yes/no).
2. Rate the prompt as 'useful' to 'distracting' on a five point scale.
3. To state whether they would recommend recall on the basis of the prompted feature (yes/no).
4. To classify each prompted feature (free text response).
5. To rate the significance of each feature on a five point confidence scale: C1 normal, C2 benign, C3 equivocal, C4 suspicious, C5 malignant.

In addition to supplying details about prompted features, subjects were also asked to annotate and describe additional features in the mammogram if they felt that prompting for that feature would be useful. They were also asked to rate these additional prompts in the same way as the actual prompts. Finally, subjects were encouraged to give a verbal commentary on both their interpretation of the mammograms in the test set and of the system's response using a 'think aloud' protocol. Subjects' commentary was tape recorded and transcribed. In the following, transcript extracts are labelled according to the subject (H, J or R), the session (1 or 2) and the case examined (1-74 or 1-70). Thus the label (H-1.23) identifies the abstract as belonging to subject H reading case 23 in her first session.

The results we present here focus on the question of how subjects used prompts as an error management tool, and on how they made sense of system behaviour. They show how even FP prompts may serve a useful purpose for the reader.

Accountability: Mammograms are information rich artefacts that are examined by trained observers for abnormalities that sometimes have a very subtle presentation. Given the reading workload generated by the breast screening programme, it would be impossible for readers to approach their task by exhaustively examining and analysing each part of each mammogram. Moreover, readers' attention is a limited resource, and human observers are prone to fatigue. Studies have shown that visual search in radiology is often incomplete (ref. 18) and that experienced readers are able to quickly 'zero in' on significant features (ref. 19), attending to each mammogram in a way that is dependent on content.

Mammograms can be more or less difficult to interpret for a number of different reasons, including variations in tissue type, tissue distribution, and the effectiveness of the image acquisition process. Similarly, there can be variation in the degree of difficulty associated

[1] Features that have some properties in common with those readers interpret as suspicious.

with the interpretation of individual features. Some features may be obviously benign or malignant, others may be ambiguous because either they are in the early stages of development, or because they are imperfectly represented within the image. Thus it is not necessary to attend equally to every feature within the mammogram -- some can be cursorily dismissed, others require more protracted thought and examination.

The approach taken by readers involves selectivity in the application of effort to produce an acceptable level of performance under particular resource constraints. Selectivity is mediated by heuristics for deciding what is 'worthy' of examination and in what detail. Readers may expend greater effort in examining dense breasts, and may examine regions in detail with the aid of a magnifying glass. They will examine closely features that 'catch their eye' and, depending on the characteristics of the presentation, may resort to other information sources, such as additional views or previous films, or to particular strategies, such as 'undressing' lesions, or use of a bright light source. They also may pay greater attention to particular regions of the breast known to be sites where cancers can be missed -- the so-called 'danger areas'. In short, readers employ a more strategic approach to render the reading task tractable.

Readers do not have to account for every feature within an image, but they do have to account for features that satisfy generally accepted heuristics for significance. Readers also have to account for features in a particular way -- that is, according to the most appropriate strategy for analysing a given type of presentation. Thus accountability to the process is bound by what an experienced reader might reasonably be expected to notice, the lengths that they might be reasonably expected to go to establish the status of some noticed feature, and by what analytic strategies it might be most reasonable to select given the type of presentation. Accountability demonstrates an approach to the management of selective attention by driving a continual series of reflections about courses of actions available and the certainty of any conclusions. The end results include both a decision and its rationale.

On several occasions subjects appeared to be using PROMAM to maintain their accountability to specific presentations:

> "So it's brought ... for some reason it's decided on that one, but ... I suppose it's valuable in that it makes you look a bit more closely at it. But I think it's breast tissue and I would not be bringing the women back -- I don't think that it's unreasonable to prompt it." (J-1.42)

> "So I think that is useful to prompt. If at the end of the day if we then analyse and say well that's benign that's fair enough. But that's useful to have brought to your attention." (H-2.33)

In each of these cases the prompted feature was judged to be benign and had no influence on the final decision. However, subjects believed the prompt had served a useful function by 'bringing [it] to their attention' and by doing so encouraging a 'closer' or 'proper' inspection. In so doing, it is possible that prompting has a psychological benefit by reducing the anxiety a reader may have about whether a thorough visual search has been made. Prompting may also improve readers' capacity for self-awareness and reflection. When confronted by a prompt it is natural for a reader to reflect on whether they saw the feature, how much attention they gave to it, and the interpretation they reached. One benefit of encouraging reflection might be to prevent the reader from too hastily narrowing the possible interpretations that might be attributed to the mammogram.

Overall, subjects' views on the reasonableness of prompts appear to be highly contingent and dependent on interrelated factors. In addition to the importance of a feature's character as an indicator of suspicion, the context of its presentation also plays a role in determining how much effort should be invested in its investigation. Thus prompts may be judged reasonable because they attend to contextual considerations, sometimes to the extent that they may be judged reasonable even where the feature prompted has little or no significance.

Context: One of the ways readers' orientate themselves to their task involves attending to perceived shortcomings in their abilities. Film readers are aware that there are regions within the breast where lesions are more often missed, the so called 'review areas', and thus may pay greater attention to these regions, may imbue lesions presenting in these regions with a greater degree of significance:

> "I think we probably would recall on that, so it's an asymmetry in a review area --

> therefore it's a bit more sinister, it will probably be nothing but it's an area we would want to see." (H-1.25)

Subjects occasionally judged the reasonableness of prompts against the criteria of location. Subject H was particularly keen for the system to prompt for features in the review areas:

> "What I think it would be useful to prompt is this asymmetry up here in the left. Erm, I'll circle this area up here -- the reason I think that's useful is although you get a lot of normal asymmetries up there, its also a common site, or a relatively common site of cancers." (H-1.3)

In this case, subject H draws attention to her detailed, region specific, knowledge. She demonstrates a sensitivity to the importance of examining a specific region for abnormal presentations, and also to the possibility that in doing so there may be a danger of misinterpretation.

> "Well sometimes we see wee cancers down there, and it's just, it's just you know the sort of thing you just, you (don't?) attend, I'm not saying you don't look, it's the kind of thing you can miss, because it's just at the edge of your field of vision as it were, and I've seen a few missed there just when they've just been at the lower, at the infra-mammary fold. And that's not one, but I mean it's perfectly reasonable to be prompted to have a second look at it." (R-2.7)

Here subject R suggests that it is possible to miss features that present in a particular region of the mammogram. He suggests that the reason for this is not that readers 'don't look', but because it is in an area that they may not be inclined to attend to so readily. Subject R's interpretation of the system's response is particularly interesting because although the prompt is for microcalcifications, it is entirely clear that there are no microcalcifications present. The subject finds the prompt tolerable because of the effect it has in drawing his attention to region of the breast that he believes deserves attention.

One particular difficulty encountered by readers is the interpretation of dense, or patchy, breast tissue:

> "These are a nightmare when I'm doing them, because I think you could hide Moby Dick in there and not know. And these are the ones where we have a high error in that there can be opacities in there which you don't really appreciate, and there can be some micro-calc which you don't appreciate." (J-2.23)

Dense breast tissue complicates the task of interpretation and readers are aware that their judgements may be less reliable. In addition to the anxiety associated with the possibility of misinterpretation, there may be also a wider professional concern that the decision that a breast is normal is made with less confidence, and thus may not have the same significance for a woman with lucent breasts.

Subjects also commented that a prompting system might have a useful role to play in addressing this difficulty:

> "I know that when we read them our sensitivity and specificity goes down the denser the breast, and you would hope that the computer would be able to iron that out and have the same sensitivity." (H-2.23)

It may be that the prompting system's value is perceived as being dependent on specific weaknesses readers identify in their own abilities. For example, more notice may be taken of, and greater significance attributed to, the presence or absence of a prompt in dense breasts as opposed to lucent breasts.

In the next case, subject H attends to breasts rich with features that have suspicious characteristics, and is thus faced with the problem of differentiating between these confusing, attention grabbing, benign presentations and any actual malignancies. If there was only a single presentation of this type, then resources could be efficiently, and less ambiguously, allocated to its consideration. In the case of multiple presentations, additional effort is required to organise how the lesions might be considered:

> "The area that have been prompted erm, are up here -- now that has been quite useful because ... in a breast like that they are difficult to assess because it's so patchy and you can imagine asymmetries all over the place -- what the prompt has made me do is go back and look particularly at that one -- I think that's actually quite useful -- I don't think it's worrying, but out of all the patches that are in front of me it's said look again at

> these two -- and that's quite useful, I think." (H-1.9)

Subject H uses the prompt to focus her analysis in a situation where there are many regions in the breast demanding attention. In doing so, she makes a tacit assumption that prompted features are more likely to be more significant than unprompted features. In the following extract, subject J demonstrates greater tolerance to prompts for vascular calcifications because of the occurrence of a prompt for a suspicious cluster:

> "This is for micro-calc. Oh (...?), this lady's got a cluster, a cluster of micro-calc on the left -- which is A, and on the right -- that's B. And let's see what C is ... Now, I think C is, I'm looking at the diagram, I think C is actually vascular but B is definitely not and none of it is ... or it's not definitely not, it's probably not. And neither is it on the right. So A and B are micro-calc, which actually look ... and I would be recalling. C I think -- probably vascular. And I wouldn't recall for that. Definitely helpful. In this case actually it's not distracting, so it's helpful to look (...?). In the situation where they've got other clusters, then of course this could be another cluster of the same. So I'm going to give it a C4, (...?) C3 (...?)."

The presence of suspicious clusters appears to heighten subject J's alertness towards microcalcifications more generally. She is pleased to have her attention drawn towards other instances of calcification so that they might be accounted for.

<u>Accounting for prompts:</u> For readers to benefit from using a prompting system, they must be confident that it is capable of detecting cancers, and that the system's specificity is sufficient to make its use worthwhile.

> "Now, there's one prompt that's been put all the round the left breast. And there's nothing there -- breast tissue. Deciding that I should look again and make sure there's not a mass, but -- very slightly different projection from the right, and it's breast tissue, and I would not be bringing this lady back." (J-1.40)

In the extract above, subject J makes a point of re-examining the prompted regions to confirm her initial analysis. She takes reasonable steps to ensure that the system has not detected something that the she did not initially apprehend and finally identifies a characteristic of the mammogram as a possible reason for that region being prompted.

> "Now looking at the other bit, that's what caught my eye to start with, it's gone for another area here, sort of (...?) oblique linear. I think that's breast tissue, I would be recalling the women anyway so I will see what it's like." (J-1.61)

Subject J entertains the hypothesis that the prompted feature may be significant, although she doubts this. She maintains her accountability to the prompt by suggesting that she should investigate the region to the limit allowed by current circumstances, and she is initiating an act that will further her understanding of the capabilities of the system.

If a reason for a prompt is not readily apparent, then this can pose problems. Readers are aware that minimal signs of cancer can be overlooked or misinterpreted. Thus using the system is not a simple matter of examining, or re-examining, the prompted region for signs of cancer: the prompt itself demands interpretation. A plausible explanation for the presence of the prompt, in terms of both image properties and system behaviour, has to be sought. The following extracts demonstrate how the interpretation of FP prompts for 'subtle' features can be problematic:

> "Right, so A -- I can't really see -- so well should I be saying, 'Oh, there's calcium there -- recall the patient' -- and obviously I'm overriding this (thing) -- can't see it -- you know, it can't be that worrying." (H-1.65)

The prompt in case 1.65 presents subject H with a dilemma: has the system detected something significant that she cannot herself see? Her discomfort is in part due to the lack of an obvious cause for the prompt that can be used to account for its presence -- there is no good reason for discounting the prompt other than that she cannot see 'what it is for'.

A prompt does not, in itself, indicate what it is for, other than in the broadest sense of being produced by either the microcalcification or for an ill-defined lesion algorithm. It simply highlights a region for examination by a film reader. Thus the onus is on the reader to discover a rationale for the prompt. This process can be time consuming and inconclusive without an understanding of how the feature detection algorithms work -- often a rationale is not obvious from the examination of the prompted region alone. We will return to this issue later.

Influencing interpretation: A prompt should not increase a reader's suspicion of a feature simply because of its presence, but there were a number of occasions where subjects reported that their interpretation of a feature was affected by the presence of a prompt:

> "So it wouldn't be unreasonable at all to bring this woman back and I probably ... with the prompt I probably ... it would make me think 'yeah maybe we should get reviews on this'. That's probably nothing though. So I think that's acceptable and useful." (J-1.36)

Common to subject J's appraisal of these cases is the influence of the prompt on her decision -- the prompt is 'making her think' about recalling, or she is recalling 'because of the prompt', for features that in all probability are 'nothing'. Subject R also reports heightened suspicion due to the presence of a prompt:

> "(...?) That's fair enough to make you look more closely at that particular area, maybe (...?) that's quite useful actually. Would you recall having been prompted to it? I think that once I had been prompted to it I probably would recall it, it's a bit like seeing it as a second reader. If you saw it the first time you might let it go, but if someone has seen it before you wouldn't let it go, so I think we would recall it." (R-2.40)

The features considered for recall by subjects J and R in the above cases all appear to have borderline significance -- they fall on or around the readers' recall thresholds. Readers face a dilemma because they know that some cancers will present minimal signs on the mammogram, but recalling for all features presenting a minor degree of suspicion would overwhelm resources available for assessment clinics. In case 2.40, subject R demonstrates an approach to managing recall decisions where the evidence from the image alone is ambiguous by using the decision made by a first reader as an additional source of evidence. Similarly, he suggests that the presence of a prompt could be used as evidence of abnormality for ambiguous cases. Conversely, subject H suggests that the lack of a prompt can be significant:

> "It hasn't really picked up on the asymmetries but they're not worrying in any way ... would I rather it prompted or didn't? I don't seem to be very consistent do I? Because on the one hand you've got the comfort factor -- oh it's seen it and dismissed it. I think they're not in anyway worrying, if they were more striking asymmetries then perhaps I would want it -- in that case I think I would let them go." (H-1.33)

Here the lack of a prompt is seen as 'comforting', precisely because subject H equates the lack of a prompt as indicating that the system has assessed a region and found it to be benign.

Making sense of PROMAM: In subjects' repeated attempts to understand what a particular prompt 'means' and how it should be 'interpreted', we find evidence of their active engagement in making sense of the system's behaviour. In this way, subjects develop the capacity to assess PROMAM's capabilities -- what it might reliably detect, and what might be overlooked, and also to explain its responses.

Subjects' strategies for making sense of the system included:

1. Comparing the system's responses for similar types of feature.
2. Comparing their ideas of significance with the systems.
3. Assuming purposeful behaviour.
4. Considering what might be indicated by the shape, size and location of prompts.

Often subjects made comparisons between the system's response for similar types of feature, either in the same breast, or between cases. Subjects often found the tendency of the microcalcification detection algorithm not to prompt all benign microcalcifications to be a source of confusion:

> "I'm surprised that it hasn't, that it hasn't picked up on the vascular calcification on the right. (...?) really quite surprised about that, since it's gone for things on the left." (J-1.39)

Similarly, subjects often attempted to make sense of the ill-defined lesion algorithm by making comparisons between prompted and unprompted features:

> "I'm kind of struggling to see what they are prompting for. It's just asymmetrical breast tissue. If it prompted for that -- why did it not prompt for that. So I'm writing 'why not prompted?' because it's more of the same -- (plus)? a bigger area." (H-1.64)

Subjects expect the system to be consistent and so are puzzled when it doesn't prompt for features that are to them (diagnostically) similar to prompted features. This is partly due to their lack of familiarity with the details of system behaviour and also because of the effect on the system of variations in image properties that may seem insignificant, or are difficult for a reader to perceive.

This can be seen most clearly in the operation of the micro-calcification detection algorithm where a simple clustering rule is the system's criteria for prompting. In case 1.39 (above), subject J expresses confusion because of the system's inconsistent response to seemingly similar regions of benign vascular calcification. In its early stages, vascular calcification can be discontinuous or fragmented, and it is this type of presentation that satisfies the algorithm's simple clustering rule. Perceived inconsistencies in prompting may sometimes stimulate a search for more sophisticated explanations of system behaviour:

> "There's a vessel running down there -- and isn't that strange? Well this is it again because we've got other bits of vascular calcification which it hasn't prompted on the same vessel with it coming down here, and that's the bit it's gone and highlighted, I don't know why. So that it is a bit of a cause for concern I think. Just why has it gone for that bit, is it because it's in the bit of black breast ... you know, fat, that's standing out a wee bit more." (J-1.59)

In the above case subject J is able to identify 'low level' differences between prompted and unprompted features -- improved contrast between calcifications and background tissue. In another example, subject J was able to account for the tendency of the ill-defined lesion detection algorithm to produce FP prompts for a "linear increase in density":

> "Now this is another area, it seems to pick up areas like this of linear increase in density which it is calling a mass, I'm sure it's not. It's just the way the breast tissue has involuted. We're left with fibrous strands and just vaguely increased density." (J-1.24)

Much of subjects' sense-making is driven by the assumption that not only is there some reason for the presence of a prompt, but also that there is some *good* (i.e., diagnostically relevant) reason:

> "Why has it prompted that lymph node, and not others, I wonder? [...] Because I mean if it's ... if there's some particular reason it's because if it's margins or something like that, and that's fair enough, just to make sure that it is a lymph node, but if it's going to pick up every lymph node then it's completely unacceptable, it would prompt every second film just about. But it hasn't been doing that, so there must be a reason why it's prompted that, so I'll say that's Ok. But I'm happy (...?) it's a lymph node. Some women will get cancers there as well which, I suppose." (R-2.25)

However, an assumption of purposeful behaviour can be misleading. A striking example of this is where subjects associate ill-defined lesion prompts with asymmetries:

> "... an elliptical prompt there round something that I'm sure is breast tissue -- composite. [break] ... (done?) the same again. But I mean, I suppose, looking at it, there isn't an equivalent area over here, so it's reasonable enough to have prompted that. But I wouldn't have recalled it for that." (J-2.31)

In fact, prompts for asymmetry are chance occurrences. Breasts are naturally asymmetric, and the ill-defined lesion algorithm will tend to produce FPs on denser patches of tissue, which may just happen to correspond to regions of differential brightness or distribution.

Memory for both prompts previously encountered, and candidate explanations for those prompts, accumulate as part of a compiled biography of the system's behaviour that may be used to account for current prompts. This working understanding of the system's behaviour is subject to incremental -- or sometimes radical -- revision as the reader is exposed to more evidence.

The form of the prompts, and their relationship to the prompted feature, were deliberately chosen to deliver certain types of information.

Prompts from the system components are distinctive, allowing the responses made by the microcalcifications and ill-defined lesion detection algorithms to be easily distinguished. Prompts produced by the micro-calcification algorithm delineate the shape of the cluster that has been detected. Similarly, ill-defined lesion prompts consist of an ellipse that circumscribes the detected lesion with an additional margin of ten percent by area. However, subjects suggested that they might learn to recover additional information from prompt characteristics:

> "Multiple prompts on this one. They're all micro-calc. I'm not dismissing them out of hand, but just even looking at the prompt, at the way it's outlined on here, it looks like vascular calcification, and indeed that's what it looks like on first looking at the film." (J-2.11)

Subjects' responses to case 1.61 provide an interesting example of how subjects' interpretation of the meaning of a prompt can differ. The system has produced two ill-defined lesion prompts. Prompt B circles a large region of dense tissue in the upper part of the right breast. Prompt C highlights a smaller region and lies wholly within the region circumscribed by B.

> "Now, there is an asymmetry -- but I wonder if there is a cyst in the middle of all that. So I would recall this patient. Now, what they've done is they've prompted the whole area -- which I don't think has been helpful. That bit ... B, the whole area, not helpful (...) C, again it's, I think it's not the right area -- so that's not helpful." (H-1.61)

> "... so B, an increase in density and it does merit ... it's helpful ... it will be relevant -- because I can't see the margins of it, that's fine, it's suspicious. Now looking at the other bit, that's what caught my eye to start with, it's gone for another area here, sort of (...?) oblique linear. I think that's breast tissue, I would be recalling the woman anyway so I will see what it's like, but if that wasn't there, then I wouldn't be impressed by that ... C is in the middle of this bit. I'm not quite sure whether it thinks there is a second mass within ... I don't quite know what ... whether it thinks there is a separate mass, because if I was drawing a line I would draw it round here ... So I'm not quite sure what it's getting at." (J-1.61)

Subject H identifies a significant feature within the region highlighted by prompt B, but doubts its relevance because it prompts too wide an area. Subject H also dismisses prompt C as being for the wrong area and proceeds to annotate the extent of the feature as she sees it. In contrast, subject J feels that prompt B is relevant, but has difficulty finding an adequate explanation for prompt C. Prompt C poses a problem for subject J because it both misses the focal region of significance, and occurs within the region of prompt B, which is seen as significant. This poses several questions simultaneously: If the system has identified the region annotated by H as significant (although by a wide margin), why has C been prompted? If a small focal area such as C can be prompted, why is B not more focal? If the entire region given by B is significant, then why bother prompting smaller regions within B at all?

Such questions can be settled by an understanding of the ill-defined lesion algorithm. Processing is essentially done in two stages. In the first stage features within the mammogram are extracted and segmented according to four different scale sizes. This can be thought of as a sieving operation which allows features falling into broad categories of size to be treated separately. Features conforming to each of these sizes are then classified according to known properties of malignant lesions. Each of these steps is independent and the prompts generated are completely independent of one another. The regions highlighted by prompts B and C thus belong to different scale sizes and their significances are unrelated.

Summary: The results of this second study confirm readers' tolerance of FP prompts and provide interesting pointers as to how readers may use prompts as an error management tool in the clinical setting. They also show that readers began to develop quite a detailed understanding of PROMAM as evidence of its behaviour accumulated.

This result might be purely an effect of the study, and may not be relevant to PROMAM's clinical use. Notionally, the user of a prompting system does not need to know how it works, merely to attend to those areas in the image that it prompts. However, our data suggests that readers benefit from investing effort in making sense of the PROMAM's behaviour because this facilitates more efficient and effective use.

The PROMAM system is designed as an attention cue -- its role is to ensure that features with possibly malignant characteristics are not overlooked by a human observer. The presence of a prompt should merely imply that attention is required (because the system is sensitive), but should not imply that a recall decision is appropriate (because the system is not very specific). The responsibility for assessing the significance of a prompted feature, and thus for making a recall decision, should rest entirely with the reader. However, the data shows that sometimes readers were influenced in this way, so using the system in an unintended manner.

Small Scale Clinical Evaluation of PROMAM

To investigate whether the issues raised by the first two studies were relevant to understanding PROMAM's use in the clinical setting, they were followed up by a small-scale clinical trial.

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 (ref. 26). 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. In the prompted conditions, subjects were asked to first examine the films, then examine the prompt sheet, and then to record their decision. Subjects were given an overview of how the system worked, including the types of feature it was capable of detecting, and trained in its use. Training included specific instructions regarding how to use PROMAM, i.e., that prompts were simply attention cues with no classification significance. A full account of methodology and results can be found elsewhere (refs. 11, 26).

Outputs from two of PROMAM's feature detection algorithms were used to generate prompts for microcalcification clusters (ref. 12) and ill-defined lesions (ref. 17). 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 modularity.

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 trial, and after each subject had completed all their allocated sessions.

Dealing with FP prompts: Ideally, readers 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 developed strategies to determine the significance of system information based on an *a priori* assessment of the prompt sheet. For example, one subject indicated that the shape of prompts for vascular calcifications, and the location of prompts for ill-defined lesions, could 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."

Another indicated that some prompts could be assessed from her examination of the prompt sheet alone:

> "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."

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 invested in further scrutiny of the mammogram. In such cases, consideration of possible explanations is not deferred until all the evidence has been gathered (ref. 8). Two subjects, for example, indicated that they might not look back as carefully -- or at all -- depending on their initial assessment.

Predicting prompts: Subjects reported that they were able to develop quite quickly a capacity to predict which features in the mammogram would be prompted.

> "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."

One subject volunteered an explanation for why predicting prompts was useful:

> "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."

There is a cognitive cost associated with this strategy as it requires that readers must form a more accurate model of system behaviour. However, checking whether prompts meet with expectations appears to be an intuitive reaction for readers, and is perhaps essential for establishing and maintaining trust in system performance. We argue also that prediction is a valuable strategy because it implies that the reader has actually made an assessment based on the evidence in the mammogram.

The success of prediction is dependent upon prompting system consistency as *perceived* by the reader. 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 readers would classify as being similar.

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 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. 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 making instructions given in pre-trial training, both questionnaire data and responses given in post-session interviews indicate that subjects sometimes used prompts as classification decisions aids. Subjects referred to occasions where they had found the absence of a prompt 'reassuring':

> "Yes, yes, I think that that is reassuring. It might just be falsely reassuring sometimes."

The quote above indicates that the absence of a prompt was viewed as 'reassuring' only, merely confirming a decision that had already been made. However, subjects also reported cases where the presence of a prompt had made them more inclined to recall:

> "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."

Overall, subjects' comments suggested that the presence or absence of a prompt was most likely to influence a decision when the evidence available from the mammogram alone was ambiguous. It is possible that in these situations subjects attempted to use whatever evidence was to hand, including prompts, to resolve 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."

One subject drew an analogy between prompts and heightened suspicion when another radiologist asks her to examine a case:

> "... it's like when someone shows sets of mammograms 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."

Summary: The results of this third study reveal how readers adapted to using PROMAM in the clinical setting. They revealed that subjects spontaneously improvised their use of prompts in a number of ways which helped them economise on the effort required to deal with prompt system errors. First, they began to apply strategies for determining the significance of prompts based on prompt -- rather than image -- features. Second, they began to actively predict where prompts were likely to appear.

The study also confirmed that not only did subjects use prompts as attention cues, but as decision-making aids when other available evidence was ambiguous.

Conclusions

The use of prompting systems in breast screening is intended to reduce observer error by helping readers to avoid errors of attention. While unambiguous, quantitative evidence of performance improvement must necessarily await the outcome of large scale clinical trials, our studies provide some support for the achievement of this goal. They also suggest that readers can learn how to manage the potentially undesirable effects of prompting system errors on their own performance. However, we also find that presumptions of prompting systems being used purely and simply as attention cues may be misplaced.

One problem is that the design rationale for prompting systems assumes generic difficulties -- i.e., that readers sometimes have difficulty ensuring that the entire mammogram is examined. However, the problems readers face when examining a set of mammograms are actually very specific and highly contingent. For example, the reading of dense, or feature rich, breasts poses demands very different from those posed by lucent, or uncomplicated, breasts. Furthermore, although readers have general concerns that they might, for example, overlook a malignancy, they also have a more specific understanding of particular deficiencies in their expertise. For example, they might perceive themselves to be more or less able to detect and correctly classify particular feature types.

It would be a mistake to believe that error-free and effective use of prompting systems in breast screening can be achieved if the user is expected to treat the system as a mere “black box”, with no understanding of how it behaves, even if the user is a highly skilled reader. On the contrary, our studies show that efficient and accurate use of prompting systems depend on the system’s behaviour being accountable to its users. Readers maintain accountability of their own work in the context of an understanding of their performance characteristics (knowledge about their skills, limitations and expected behaviours). This is often a poor model for accounting for prompt system behaviour, especially where erroneous prompts are simply artefacts of the methods used to analyse the image.

One solution may be training. Prior to the small scale clinical trial of PROMAM we sought to provide an account of its behaviour through training. However, the training material (though informed by the earlier studies) has still to mature. Developing and evaluating training materials and protocols may prove to be as complex as the development of the system itself.

To address the issue of how prompting systems *should* be used by readers, we may learn from the preparation and use of evidence in current reading practice. For example, readers often organise the ordering of attending to evidence to minimise bias. Though we enforced a protocol in the small scale clinical trial whereby subjects examined each mammogram before examining the prompt sheet, our evidence suggests PROMAM still influenced classification decisions. Further innovations in reading protocol might be appropriate, such as requiring readers to reach a decision before examining prompts, and only allowing ‘routine recall’ decisions to be amended in light of evidence from the system.

Finally, the goal of replacing double reading with a single reader aided by a prompting system may raise wider problems concerning reading practices. Our studies show that readers may exploit of double reading as a way to monitor their performance. There is an informal, collaborative dimension to double reading, and to readers’ management of their performance, which has so far been largely ignored. We argue that the implications of this for single, prompted reading require careful consideration.

References

1. E D C Anderson, B B Muir and A E Kirkpatrick, The efficacy of double reading mammograms in breast screening, Clinical Radiology, 49, 248-251, 1994.
2. J Chamberlain, S M Moss, A E Kirkpatrick, M Michell and L Johns, National Health Service breast screening programme results for 1991-2, British Medical Journal, 307, 353-356, 1993.
3. S Ciatto, M R D Turco, D Morrone, S Catarzi, D Ambrogetti, A Cariddi and M Zappa, Independent double reading of screening mammograms, Journal of Medical Screening, 2, 99-101, 1995.
4. H E Deans, D Everington, A E Kirkpatrick and E Lindsay, Scottish experience of double reading in the National Breast

Screening Programme, The Breast, 7, 75-79, 1998.

5. E R E Denton and S Field, Just how valuable is double reporting in screening mammography?, Clinical Radiology, 52, 466-468, 1997.

6. J Fox, Decision-support systems as safety-critical components: Towards a safety culture for medical informatics. Methods of Information in Medicine, 32, 345-348, 1993.

7. M Giger, Computer-aided diagnosis. In A Haus and M Yaffe, (Eds.) A categorical course in physics: Technical aspects of breast imaging, 283-298. RSNA, 1993.

8. A Gale, Human response to visual stimuli. In W Hendee and P Wells, (Eds.) The perception of visual information, 115-130. Springer-Verlag, 1995.

9. M Hartswood, R Procter and L Williams, Subjective reaction to prompting in screening mammography. In C Taylor, (Ed.) Proceedings of Medical Image Analysis and Understanding ‘97, Oxford, 1997.

10. M Hartswood, R Procter, L Williams, R Prescott and P Dixon, Drawing the line between perception and interpretation in computer-aided mammography. In E Fallon, L Bannon and J McCarthy, (Eds.) Proceedings of the 1st International Conference on Allocation of Functions, 275-290. Dublin: IEA Press, 1997.

11. M Hartswood, R Procter and L Williams, Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography?. In N Karssemejer, (Ed.) Proceedings of the 4th International Workshop on Digital Mammography, 1998.

12. A Hume, P Thanisch, M Hartswood and R Procter, On the evaluation of microcalcification detection algorithms. In Proceedings of the 3rd International Workshop on Digital Mammography, Chicago, 1996.

13. I Hutt, The Computer-Aided Detection of Abnormalities in Digital Mammograms. Unpublished Ph.D. Thesis. Manchester University, 1996.

14. H Kundel, C Nodine and D Carmody, Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection. Investigative Radiology, vol 13, 175-181, 1978.

15. A Lesgold, R Glaser, H Rubinson, D Knopfer, P Feltovich and Y Wang, Expertise in a complex skill: Diagnosing X-Ray pictures. In M Chi, R Glaser and M .J Farr, (Eds.) The Nature of Expertise. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.

16. C E Metz and J Shen, Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis, Medical Decision Making, 12, 60-75, 1992.

17. L Miller and N Ramsay, The detection of malignant masses by non-linear multiscale analysis. In Proceedings of the 3rd International Workshop on Digital Mammography. Chicago, 1996.

18. C F Nodine and H L Kundel, Eye Movements: From Physiology to Cognition. In J K O’Regan and A Levi-Shoen (Eds.) The cognitive side of visual search in radiology. Elsevier Science Publishers, 1987.

19. C F Nodine, H L Kundel, S C Lauver and L C Toto, Nature of expertise in searching mammograms for breast masses, Academic Radiology, 3, 12, 1000-1006, 1996.

20. R Procter, P Thanisch, A Hume, S Astley and A Kirkpatrick, User interface design and data management for digital mass mammography. In A G Gale, S M Astley, D R Dance and A Y Cairns, (Eds.) Proceedings of the 2nd International Journal of Digital Mammography, 405-414. Elsevier, 1994.

21. C Savage, A Gale, E Pawley and A R Wilson,.To err is human to compute divine? In A G Gale, S M Astley, D R Dance and A Y Cairns, (Eds.) Proceedings of the 2nd International Journal of Digital Mammography, pp. 405-414. Elsevier, 1994.

22. E L Thurfjell, K A Lerneval and A S Taube, Benefit of independent double reading in a population-based mammography screening programme, Radiology, 191, 241-244, 1994.

23. N J Wald, P Murphy, P Major, C Parkes, J Townsend and C Frost, UKCCCR multicentre randomised controlled trial of one and two view mammography in breast cancer screening, British Medical Journal, 311, November, 1189-1192, 1995.

24. R M Warren and S W Duffy, Comparison of single reading with double reading of mammograms, and change of effectiveness with experience, The British Journal of Radiology, 68, 958-962, 1995.

25. J C Wells and J Cooke, Film reading practice of the UK breast screening units, The Breast, 5, 404-409, 1996.

26. L Williams, R Prescott and M Hartswood, Computer-aided cancer detection and the UK National breast screening programme. In N Karssemejer, (Ed.) Proceedings of the 4th International Workshop on Digital Mammography, 1998.

27. L Williams, M Hartswood and R Prescott, Methodological issues in mammography double reading studies, Journal of Medical Screening, 5, 202-206, 1998.

Biography

Mark Hartswood, Institute for Communicating and Collaborative Systems, Division of Informatics, University of Edinburgh, Edinburgh EH9 3JZ, Scotland. Telephone: +44 (131) 650 5168. Facs: +44 (131) 667 7209. Email: mjh@dcs.ed.ac.uk

Mark Hartswood has spent five years working on the PROMAM project researching human factors issues in computer-aided mammography. Currently he is completing his Ph.D. and working as a Research Associate on a project applying participatory design methods in the development of an EMR system at Edinburgh Royal Infirmary.

Rob Procter, Institute for Communicating and Collaborative Systems, Division of Informatics, University of Edinburgh, Edinburgh EH9 3JZ, Scotland. Telephone: +44 (131) 650 5177. Facs: +44 (131) 667 7209. Email: rnp@dcs.ed.ac.uk

Rob Procter is a senior lecturer in the Division of Informatics, University of Edinburgh. His research interests lie principally in the field of human factors, particularly methodologies for the design and implementation of IT systems. His current research includes computer-aided mammography, participatory design in the clinical setting, and social learning processes in IT systems implementation.