Human Factors in Computer-aided Mammography

Mark Hartswood

Doctor of Philosophy
University of Edinburgh
1999

Abstract

Breast screening requires film readers to exercise considerable expertise when examining breast X-rays (or 'mammograms') for signs of malignancy. Understandably, errors are sometimes made, and the screening programme is continually investigating ways to improve detection performance. In recent years, interest has grown in using computer based prompting systems to assist with reading. Prompting systems use image analysis techniques to identify possible cancers within a digitised mammogram and cue film readers to their location with the aim of preventing cancers from being overlooked.

Conventionally, quantitative methods, such as ROC methodology, have been used to test whether prompting can improve film readers' performance in a laboratory setting. Recent studies suggest, however, that a purely quantitative approach to the evaluation of medical decision-support systems is inadequate. Satisfactory laboratory benchmarks are not, by themselves, sufficient to guarantee user acceptance — of equal importance are the system's impact in the workplace, and users' subjective appraisal of its utility. This thesis describes the application of qualitative methods to the evaluation PROMAM, a prompting system intended for use in the UK Breast Screening Programme.

A qualitative analysis of clinic work practices show reading to be a situated activity with important collaborative dimensions. Tensions were found to exist between making decision-making visible (hence rendering it accountable and providing a reference by which performance can be monitored) and the possibility of being biased by exposure to the decision processes of others. It is argued that use of PROMAM offers a similar mix of advantages and pitfalls, and that lessons can be learned for prompting from how these tensions are managed for conventional forms of evidence. In subsequent investigations of prompting it was found that readers' interpretation and use of PROMAM were often problematic. Readers often had difficulties understanding prompts, and used them in ways contingent on the particular problem at hand rather than purely to aid detection.

It is argued that effective prompting is not only a problem of achieving sufficient system performance, but also one of ensuring prompts are comprehensible, accountable, and appropriately used. Achieving the latter requires an understanding of how readers make sense of prompts in the context of their conventional reading practice.

Acknowledgements

I wish to express gratitude to my supervisor, Rob Procter, for his guidance and encouragement during the writing of this thesis. I also would like to thank Peter Thanisch and Roger Slack for their valuable comments.

Thanks are also due to members of the PROMAM team, including Linda Williams, Pat Dixon, Lance Millar, Ally Hume and Steve Heddle, with whom it was a pleasure to work.

I would also like make mention of Graham and Alvero, my office mates, who helped to preserve my sanity with their support and camaraderie.

None of this could have been achieved without the support and love of my wife Eve, and I have to mention my son Sam (10 months), without whom this thesis would have been completed a year earlier.

Table of Contents

Chapter 1

Introduction

This thesis explores human factors issues relating to the use of prompting systems to assist with cancer detection in the context of the UK Breast Screening Programme (UKBSP). The work was conducted as part of a requirements capture and evaluation programme for the PROMAM [1] prompting system. Detection aids, like PROMAM, are computer systems that use image analysis techniques to locate cancers within a digitised image of a breast X-Ray. Film readers (typically trained radiologists) are 'prompted' for any suspicious findings made by the system, with the aim of preventing cancers from being overlooked.

Statistical methods, such as ROC methodology, are widely advocated for the evaluation of prompting systems [93]. Their application involves testing whether particular system configurations are capable of enhancing an observer's ability to distinguish between benign and malignant features [92]. The use of statistical methods can be equated with an 'engineering' approach whereby the complex interactions between human observer, system and X-Rays are treated as if concealed within a 'black box'. The processes by which a decision is arrived at are glossed over, and only the results of decision-making are examined.

An engineering approach maintains tacit assumptions relating to both the nature of radiological expertise and the properties of prompting systems themselves. Expertise is treated as if it were a purely mechanical phenomena, residing within the brain of the expert, which can be reproduced on demand in an experimental setting. Different prompting systems with similar numerical performance characteristics are treated as equivalent, even though systems based around different image analysis techniques are likely to produce different patterns of responses. In consequence, expertise is stripped of its social setting, and qualitative differ-

[1] PROMAM is an acronym for PROmpting for MAMmography. The PROMAM system was developed in a joint venture between the Royal Observatory at Edinburgh and the computer science department at Edinburgh University.

ences between prompting systems are ignored.

Quantitative methods, by their nature, preclude certain types of interpretation and promote others. By using quantitative methods it is possible to record success or failure in terms of some chosen performance metric. If a system is found to be unsuccessful, then without the possible recourse of examining how readers actually use and make sense of prompting information, the failure is most readily framed in terms of inadequate system performance. It becomes difficult to explore alternative explanations, for example, that readers are not using the system as intended, or are misinterpreting the system's responses.

The motivation for the work described in this thesis was to address the omissions of a strictly quantitative approach to evaluation by using a combination of quantitative and qualitative methods. Rather than focussing solely on system performance, the aim was to develop an understanding of both the collaborative features of screening work, and readers' interpretation of individual prompts. In this way wider human factors issues might be addressed, including: the format and timing of prompt delivery, training requirements of readers using the system, and how prompting can be effectively integrated into existing reading practices.

The remainder of this chapter is in four parts. The first sets the scene by giving an overview of breast screening in the UK and of the PROMAM system. The second considers known problems with the deployment of decision-support systems into medical practice. The third provides an overview of methods for the evaluation of decision-support systems in medicine, and provides a justification for the qualitative approach adopted in this study. The final section gives an overview of the remainder of this thesis.

1.1 Breast cancer screening

1.1.1 Overview

Breast cancer is the commonest form of cancer in the U.K. 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 [46]. Mammography (radiological imaging of the breast) remains the best method for early detection of breast cancer, and mammography screening programmes operate in many countries. In the UK, women between the ages of 50 and 64 are invited to attend a clinic for screening mammography every three years.

In general, screening tests are not intended to be diagnostic. Instead a test is used that is minimally invasive and that can be rapidly applied to make a provi-

sional assessment of suspicion. For test positive cases, more specific procedures are then employed to establish the truth of the initial findings [95]. Where a suspicious finding is made during screening mammography, the woman is then invited to an assessment clinic where additional procedures are conducted, which may include additional radiographs, ultrasound and biopsy.

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 prevalent round — 6.3 per thousand, compared with 3.4 per thousand in the incident round [22]. The task of reading films is a difficult one, not least because the small number of cancers present is hidden in a background of largely normal cases. A high level of perceptual and interpretive skill is required to ensure that as many cancers as possible are detected, whilst limiting the number of unnecessary recalls for assessment.

1.1.2 Signs of cancer

Screening mammography is an X-ray procedure for visualising breast tissue. A mammogram is a projection of these three dimensional structures resulting in a composite density map of the breast in two dimensions. As a consequence cancers are sometimes partially or completely obscured by overlapping normal tissue, and overlapping normal tissue can present as regions of high density and thus mimic the presence of a cancer. Furthermore, it is possible for benign processes to have an appearance similar to those that typify tumours, and conversely, for tumours to masquerade as benign processes (some examples of these are described below). Thus the degree of certainty about whether a feature within a mammogram is due to a malignant process can vary considerably — some lesions appear as unequivocally malignant, others might only be mildly suspicious, and some tumours are 'occult' — they are not visible on a screening mammogram at all. Mammograms are examined for evidence of an abnormality by one or more experienced readers. The types of feature that are indicators of malignancy include [31, 32, 117]:

Microcalcifications These are small deposits of calcium visible on a mammogram as tiny bright specs, often described as having an appearance similar to grains of sand or salt. Breasts will often contain some calcification due to benign processes, for example, it is common for vessels to calcify — giving a characteristic 'tram line' appearance on the mammogram. An assessment of the number, shape, size, distribution, variability and stability of microcalcifications is used to characterise a presentation as possibly malignant

or benign. Malignant presentations are typified by a small, focal cluster of microcalcifications that have a variable size and shape, and irregular margins. In contrast, benign presentations usually have smooth margins and are diffusely scattered in both breasts. Microcalcification clusters are an important sign of malignancy as they are often associated with an early, pre-invasive stage of tumour development.

Ill-defined and spiculated lesions These are areas of radiographically opaque tissue, often called "masses", appearing as a 'bright patch' on the mammogram that might indicate a developing tumour. Typically, lesions that are circumscribed (having a visible border), with a bright halo and of a low density are often the result of benign processes — for example, cysts have this appearance. However, a small number of cancers will present as well-defined masses. Lesions that do not have a well-defined edge, that are of medium to high density with a heterogeneous texture are characteristic of a malignant process. Often the density is surrounded by fine radiating tendrils (or 'spicules') due to the infiltration of surrounding tissue, giving a stellate appearance. Some types of cancer will infiltrate the surrounding tissue without producing a central mass, and present as an area of spiculation or distortion.

Architectural distortion Infrequently a cancer may present as a disturbance of the alignment of normal tissue to the nipple, or as a disturbance to the edge of the glandular disk. There may also be spiculation present without an associated mass.

Asymmetry Focal areas of asymmetry between the left and right image are common and in the main benign. However, if an asymmetry has a similar shape on two views, is opaque or associated with other signs, then this may be indicative of a cancer. Very few cancers are detected on the strength of asymmetry alone.

Secondary signs A tumour may cause the contraction of ligaments that support breast tissue, leading to focal areas of skin retraction. If the tumour is located centrally nipple retraction or inversion may occur. A tumour extending to the skin may cause skin thickening or ulceration.

1.1.3 Reading

When mammograms have been developed, they are checked for technical quality and are loaded onto viewing boxes for reading. Typically each clinic will have a number of automated viewing boxes that can hold between 250 and 800 films, depending on the model. Films are illuminated by a strong back light, and extraneous light is occluded using blank, exposed, films. Each mammogram is examined by at least one trained film reader who typically is also a radiologist. Many clinics operate the practice of 'double reading', where the opinion of two or more readers is combined to give a final decision for each case. Within clinics 'film readers' are often referred to as 'radiologists'. However, members of other medical specialities may also be employed as film readers if given the appropriate training, so in this document the generic term 'film reader' is used. Typically the involvement of non-radiologists is limited to centres that practice double reading [129].

The number of cases read in a single session is highly variable, and will depend on the availability of readers and the workload in the clinic. Films are typically grouped together in 'batches', corresponding to all the women who attended a contiguous screening session. There is typically some paperwork associated with each of these batches, describing, for example, how many cases are in each batch, who loaded the films onto the viewer, the date the viewer was loaded, the dates when the films were read, and who performed the reading. For each batch the record bags are piled in the order that the films are displayed on the viewer. The film reader will work through the cases on the viewer and mark his/her decision on the screening forms. Some readers mark each decision as it is made, 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 usually falls to the second reader to ensure that the cases are removed from the viewer before the normal cases are taken down [2].

Lesions with one or more properties sufficiently different from background tissue (if they are relatively large or bright, for example) may be conspicuous to a film reader 'at a glance' when an initial appraisal of the films is made. Less conspicuous lesions may require systematic searching of the mammogram to loc-

[2]The above description is informed by an investigation into the work practices in 6 UK screening centres. Details of this investigation is reported fully in chapter 3

ate [98]. Typically, film readers first examine the mammogram in overview and make a global comparison between different views and with previous films [53]. This serves to identify any 'pop-out' features (those that immediately capture attention), and also to provide orientation for the subsequent focussed search [20]. Various magnification aids might be used to assist with this examination. Film reading is thought of as comprising two distinct steps: detection — locating potential lesions on the mammogram and classification — determining their clinical significance [87].

1.1.4 Reader error

Film readers are prone to making two categories of error: false negative (FN) decisions, where a cancer is overlooked or misclassified, and false positive (FP) decisions, where a 'normal' case is recalled for further investigation. Both types of error can be consequential. There may be psychological morbidity associated with recalling a women to an assessment clinic, as well as a financial cost [119]. If a cancer is missed then it might present clinically prior to the subsequent screening round as a so-called 'interval cancer'. Delays in treatment lead to a worse prognosis and increased treatment costs [46].

The performance of a screening test is usually expressed in terms of two metrics: sensitivity and specificity. Sensitivity is defined as the ability of the test to detect all of the occurrences of the disease in the screening population, and specificity as the ability to correctly identify those without the disease [95]. Sensitivity and specificity are defined in the following way:

$$sensitivity = \frac{TP}{TP+FN}, \quad specificity = \frac{TN}{TN+FP}$$

Where TP refer to true positive decisions, and TN to true negative decisions. The formulae assume that the test set contains at least one cancer and at least one normal case respectively.

In breast screening, the sensitivity of individual readers, of clinics, and of the programme as a whole is difficult to assess at any given time because information about FN decisions is not immediately available. An assessment of interval cancers provides the only means of establishing FN rates, however, complete data is only available at the end of each screening interval. Care is required to distinguish interval cancers arising from false negative decisions from 'true intervals' (cancers that have developed quickly between screening rounds), those that are 'occult' (not appearing on the original screening films), and those that are due to sub-

optimal technique (for example, if part of the breast had not been imaged). False negative interval cancers are defined as occuring where there was sufficient evidence on the original screening mammogram to warrant further investigation [114]. Methods for identifying FN interval cancers are described in [37, 114]. Simpson reports that out of a total of 167 cancers presenting between screening episodes in three screening centres, 26% were due to FN decisions, 51% were true intervals, 3% were due to inadequate imaging, 4% were occult, with 14% remaining unclassified [114].

1.1.5 ROC methodology

A useful metric of observer performance should be independent of the frequency at which a looked for disease entity naturally occurs, and also of an observer's subjective bias in setting the criteria for making a positive decision [92]. Metrics with these properties are available through ROC analysis, which describes the inherent ability of an observer to distinguish between normality and disease. Unless the observer can perform this discrimination perfectly then there will be cases where evidence of disease will be perceived where disease is absent, and vice versa. In these cases a decision is made according to the observer's confidence that the signs are sufficiently significant to warrant further action. This decision will depend on an appraisal of prior probabilities and value judgements about the costs and benefits of making a particular decision (that is, how many FP and FN decisions to allow). Where the probability distributions for confidence in positive and negative decisions overlap, then the sensitivity and specificity achieved will be dependent on the confidence threshold employed. This is shown schematically in Figure 1.1.

Since confidence thresholds are labile (they can be intentionally changed, and may be influenced by the particular circumstances or setting of the trial), then comparisons of sensitivity and specificity (between different observers, or between the same (human) observer at different times) will not be a true comparison of diagnostic accuracy.

It can be seen from Figure 1.1 that if the confidence threshold is continuously adjusted then an inverse relationship holds between specificity and sensitivity. If sets of values for sensitivity and specificity are determined for differing confidence thresholds over repeated observations and are plotted in a unit square, then an ROC curve is obtained by fitting a line through these points (Figure 1.2). A ROC curve shows all the possible tradeoffs between sensitivity and specificity as an observer's confidence threshold is varied, and thus represents the inherent

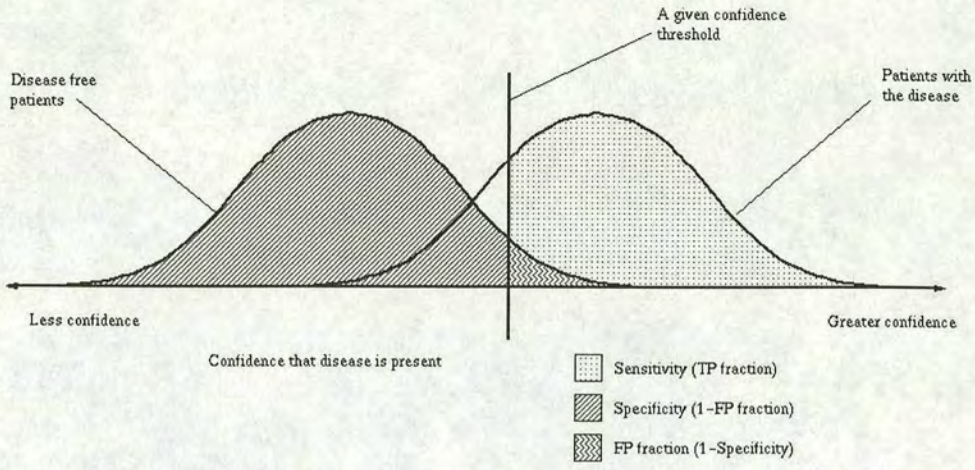


Figure 1.1: Shows the overlapping probability distributions of an observer's confidence that disease is present or absent. This is only schematic, the method does not require that the probability of observer's confidence be normally distributed, only that the distributions overlap. One possible confidence threshold is shown by the solid horizontal bar. All cases to the right of this bar would be called positive, and all to the left would be called negative. Adapted from Metz [92]

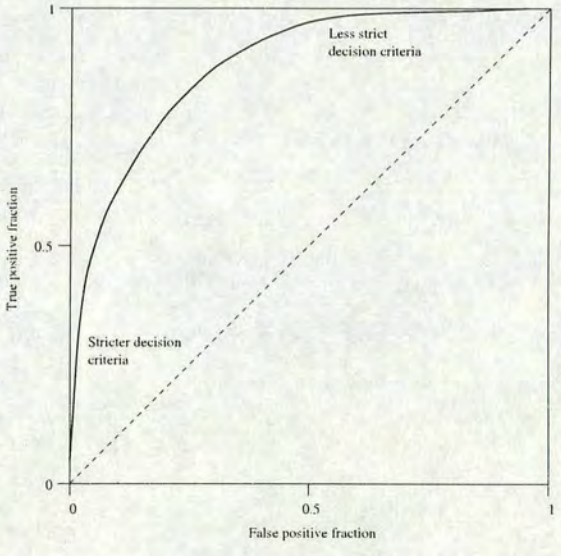


Figure 1.2: The bold line shows an example ROC curve. The dotted line line indicates the ROC curve produced by chance performance. Adapted from Metz [92]

ability of an observer to discriminate between the presence and absence of a disease. The area under the curve is often used as a index of performance and represents the average sensitivity achieved if the specificity is sampled randomly between 0 and 1, or alternatively, the average specificity if sensitivity is similarly sampled [92]. Other indices are available, and are discussed in [91]. As described, ROC methodology is applicable where decisions about abnormality correspond

to a film as a whole. Variations in the technique have been developed to allow for the correct localisation of an abnormality (localised ROC or LROC) and in cases where there might be more than one abnormality per film (free-response ROC or FROC) [93].

In a typical ROC experiment to determine the performance of human observers, subjects are asked to rate their confidence on a rating scale, forcing them to simultaneously maintain a number of different decision criteria. The technique is also applied widely to assessing the performance of computer based detection and classification systems. In this case, a range of sensitivity and specificity values are obtained by varying the operating parameters of the system.

1.1.6 The PROMAM system

The PROMAM system has three key operational stages. First, mammograms are digitised after being exposed and developed in the normal way. Second, the digitised images are analysed by two specially-developed image processing algorithms designed to recognise microcalcification clusters [69] and ill-defined lesions [96]. Third, the results of the analysis are then incorporated into the reading process as prompts. The aim is to draw the reader's attention to specific areas of the mammogram that are judged by the system to merit close visual inspection.

The introduction of digital techniques into screening mammography automatically opens up options for employing some form of soft-copy image presentation interface. The use of soft-copy images for diagnosis purposes is growing in many medical areas, but the requirements for screening mammography are especially stringent. The resolution requirements, for example, exceed the capabilities of the current generation of soft-copy displays [137]. These limitations might be overcome by exploiting the various options that soft-copy imaging provides for image manipulation and enhancement, but their use typically adds to the time required to inspect the image. A feasibility study conducted prior to the inception of the PROMAM project established that the time and effort required to use the prompt interface was a key factor in its acceptability [109]. Film readers involved routinely in breast screening would be unlikely to accept a prompt user interface which increased significantly the time spent on the unambiguously normal breast. For this reason, readers liked the simplicity of paper, form-based prompt user interfaces which have, in addition, the virtue of fitting in easily with current reading practices; paper is handled routinely during the reading session. Accordingly, this was the option that was eventually selected [109].

The aim of the detection aids like PROMAM is to improve the sensitivity of film readers without compromising their specificity. Within the project it was recognised that enhanced performance could be treated as a resource that might be applied in a number of different ways. A prompting system might not only be used to address performance issues, such as reducing interval cancers and detecting cancers at an earlier biological stage (hence improving prognosis and reducing costs), but also to maintain the same level of performance whilst using fewer, or less skilled, human resources. For example, by substituting double reading with computer assisted single reading, or by supporting non-radiologist film readers [28].

1.2 Decision support systems in medicine

Prompting systems are designed both to emulate the expertise of human film readers, and to elevate the performance of those film readers who use them. Prompting systems can therefore be thought of as a type of 'expert' or 'knowledge based' system designed to enhance the quality of film readers' decisions. This section briefly considers the difficulties previously encountered in the design and implementation of medical decision support systems (MDSSs) and how they might apply to prompting systems such as PROMAM.

The early promise that expert systems would master the intellectual aspects of medical practice [113] remain largely unfulfilled. Of the many MDSSs implemented, few have found routine use. A consistent finding is that a demonstration of effectiveness in the laboratory does not guarantee successful deployment in the workplace [3]. Explanations for the failure of MDSSs fall broadly into three categories:

1. Expert system technologies have not met performance expectations [118]: MDSS developers have been unable to deliver systems that meet promised operational specifications.

2. Design and development methodologies have been inadequate [47]: MDSS developers have misunderstood how human and MDSS performance may be best combined.

[3]Intelligent systems have been produced for 20 years for use in medicine. They have not, however, been widely adopted. One senior researcher in the field has estimated only 10 or so are in routine use, including ones used by a single doctor only a few times a week [47].

A survey carried out in 1989 reveals that out of 25 European organisations involved in the development of diagnostic systems only three were in use outside of their site of origin. Also a survey carried out in May 1992 of an AI in medicine mailing list with over 600 subscribers was only able to identify 6 systems that had been or were in routine use. Cited in [63].)

3. There have been broader methodological failings: MDSS developers have been unable to grasp that the culture and values of practitioners may be such that they will be resistant to using MDSSs [75].

These problem categories can be equated with three specific issues: scope, role and work practice.

1.2.1 Scope

The technical difficulties associated with meeting operational specifications are typically more severe for MDSSs that target general application domains. This is because the knowledge base for general domains is often less well defined: knowledge from many sources may be integrated under a variety of different reasoning strategies to reach a decision. In more specific application domains, the knowledge base is often better formalised, and the reasoning process limited to a few well-defined strategies, thus both knowledge and reasoning become more amenable to computer representation [16]. There has been a move away from systems that try to duplicate the general diagnostic capability of a physician towards systems that focus on more specific problem domains [97].

1.2.2 Role

Some MDSSs support decision-making by simply providing information that can assist physicians to reach their own conclusion, for example, performing a literature search. At the other end of the scale there are MDSSs which offer their own interpretation of the facts, i.e., automated diagnosis. In general, the latter are more difficult to design, more difficult to deploy in a working environment, and often are difficult to use. An issue of particular importance is control. For example, the physician may have the power both to decide when to use the MDSS, and to decide how to act on its advice. On the other hand, MDSS use may be compulsory. In general, the latter tends to be resisted by physicians [76], whereas MDSSs that give useful reminders or alerts have been well received [27].

1.2.3 Work practices

Work practice issues in MDSS applications are inevitably multi-faceted, and problematic for designers. Despite their complexity, it is not that issues of work practice and social change are necessarily intractable, but rather that they are often ignored by MDSS developers. Heathfield and Wyatt argue that a pre-occupation with computer artifacts and Artificial Intelligence (AI) techniques prevents the

development of a useful and coherent philosophy for the design and implementation of decision aids, claiming that the systems which might be the most rewarding to develop are not necessarily ones that doctors would find most useful. Heathfield and Wyatt urge a more problem centred, rather than an artifact centred approach [63]. Forsythe is concerned that culture of the (AI) community also serves to restrict the possible approaches to the evaluation of systems. Technical, formal, quantitative modes of thought and problem simplification are useful disciplines for the construction of knowledge bases and inference engines. However, such approaches tend to be habitualised and applied to other areas of system development. Social phenomena, which tend to be complex, qualitative, subjective and resistant to formal analysis, are often ignored. Consequently little consideration is given to how systems are perceived and evaluated by users, if systems are compatible with user needs, or how well a system fits into work patterns and organisational structure [47].

1.2.4 Scope, role, work practices and prompting systems

On the surface, issues of scope and role do not appear to be problematic for prompting systems. Prompting systems focus on a specific problem domain — that of distinguishing certain types of malignant presentation from benign structures within a breast image. Their role is to aid detection, so prompting might be positioned with systems providing alerts and reminders which have found wider acceptance in clinical practice. The use of prompting information is discretionary — in fact, the rationale for prompting states that the decision to recall a lesion for further assessment should rest with the film reader and not with the system. However, a clear definition of a system's intended purpose and capabilities should inform the use of that system as well as its design. Work presented in later chapters examines the accuracy of readers' judgements about the role and scope of PROMAM, and how these judgements are informed by exposure to the system.

The acceptance of MDSSs is often as dependent on social factors as it is on performance issues. Despite this finding, work practice issues have remained largely unexplored with respect to the deployment of prompting systems. One of the aims of this thesis was to ground investigations of prompting in an understanding of how reading is socially organised. In order to anticipate the types of difficulty posed by work practice issues, the following section explores further those barriers to the adoption to medical technologies that are rooted in the social nature of medical practice. Social dimensions to the interpretation of medical evidence

are particularly relevant to the implementation of MDSSs, and are discussed in detail.

1.3 The medical context

1.3.1 Technology and social change

The adoption of a new technology by an organisation often results in changes to work practices, sometimes in ways that are difficult to predict [110]. In practice, a new system often not only affects the work of its primary users, but also lines of communication, the distribution of power and status, and co-operative working practices [80].

These issues are illustrated in a case study of the introduction of a medical record system called PROMIS [76]. PROMIS was designed as an embodiment of PROMR (Problem oriented medical record) – a novel approach to organising the medical record which attempts to address some of the problems inherent in the traditional format. PROMIS was intended to provide corrective feedback on the health care provided by physicians. It enforced compliance with PROMR and actively volunteered information to instruct and guide clinical decisions – requiring justification from users for deviations from the protocol. The philosophy of the developers was that physicians should relinquish control to the computer in order to improve practice. During formal evaluations house staff were found to be hostile to the system. They resisted attempts by PROMIS at direction and tried to circumvent its inflexibility. Furthermore, they claimed that the system adversely altered their relationship with patients, that it changed staff communication patterns and increased the time they spent performing record keeping activities. However, the results of the evaluation revealed that these claims were substantially untrue. PROMIS was more readily accepted by nurses, patients and pharmacists, as it demanded greater participation than was traditional by these groups, thereby enforcing greater use of clinicians' professional skills. House staff saw this as a threat to their professional status and autonomy and felt that their professional judgement was being undermined. In reaction, house staff resented the nurses, pharmacists and PROMIS [76].

Evaluating organisational change as a consequence of the introduction of a new information system can be problematic. For example, it can be difficult to disambiguate transitional from long term effects, between changes due to actual use of the system, and those due to users thinking about the domain in a different way. Furthermore, if the system is a component of some wider strategy

of organisational change, then it may be difficult to distinguish between changes brought about by the system, and other deliberate changes that are part of the same strategy. Finally, because so very few medical decision support systems have been used routinely for a long period of time, little is known about the longer term consequences of their use (which, for example, might include de-skilling) [14].

1.3.2 Values and practices

Success at introducing computer applications into medical contexts is not uniform. Young differentiates between systems that are used routinely to help provide improved clinical services (for example. radiotherapy dosage, laboratory systems, vaccination and immunisation recall) and systems that while having the potential to improve quality of care, are much less widely disseminated (for example, for drug prescription, diagnosis, medical audit and history taking). He argues that the successful systems seek to support a doctor's role — not to replicate it, nor do they interfere with the way the doctors work, or cause them to incur any new burden. Unsuccessful systems, however, often demand doctors change their work practices, increase the perceived burden on doctors and seek to replace some of the decision-making role of the doctor [138].

Where new technologies are applied across boundaries delineated by medical disciplines then acceptance may not be uniform. For example, Kaplan reports a case study of the implementation of a Medical Information System (MIS) called Technicon [76]. A major goal of Technicon was to facilitate nursing activities by reducing nurse's administrative duties and assisting them in health-care planning and management. Typically nurses were enthusiastic about the system, however, physician responses were mixed. The introduction of Technicon was viewed more favourably by surgeons than by internists. Surgeons see themselves as incisive decision makers with an aptitude for technology. Internists are inclined to view themselves as "thinking doctors", and were resistant to a technology which they perceived as challenging this role [76].

At an individual level Kaplan argues that technology is more likely to be accepted if it neither change the nature of a physician's work, nor challenges what a physician considers to be the essence of medical practice [75]. Kaplan summarises some of the values held by medical practitioners which are important in this respect:

- Quality of patient care.
  One of the core values held by physicians is a concern for the quality of patient care. Thus physicians are receptive to new technologies which they

see as improving outcomes, reducing risks and making invasive procedures unnecessary. For acceptance, it is less important whether new technology actually does improve the quality of care, but whether it is perceived to do so.

- Art vs Science.
  Medical applications which make medicine more scientific (eg increased accuracy and precision) are more likely to be accepted. However, it is believed that medicine is ultimately an art, and technologies which challenge a physician's judgement or which dictate practice are likely to be rejected.

- Physician - patient relationship
  Applications are more likely to be rejected if they are seen as interfering with the doctor–patient relationship, or are seen as dehumanising medicine. Technology can move the locus of patient care away from the patient to the conference room thus increasing social distance [17].

- Professional autonomy.
  Physicians see themselves as autonomous, responsible and self-governing and believe that only their peers are competent to judge their actions. Applications which check physicians actions against protocols or which attempt to direct or constrain the physicians action would be seen as violating their professional status and are likely to be rejected.

Ost argues that the introduction of new technologies engenders a tension whereby the innovation works both for and against a physician's autonomy and status. On the one hand, status is enhanced by increased reliance on the physicians' esoteric and specialised knowledge built around the adoption of a particular innovation. On the other, status can be adversely affected by the tendency of major innovations to fragment specialities, compartmentalise knowledge, and to create a greater dependence on the medical authorities supplying the capital for the innovation [107].

1.3.3 Medical evidence

Of particular importance for the deployment of MDSS are the existing mechanisms for establishing the credibility of medical evidence. Information pertinent to diagnosis is compiled by physicians from a number of different sources and is represented on a variety of different media. These might include oral accounts from patients and other physicians, written accounts in the form of medical records and

laboratory reports, and information embedded in visual representations, such as X-rays and CAT scans. The selection and integration of information to achieve a diagnosis is a socially complex task as well as a cognitively demanding one. For example, diagnosis can be a collaborative activity and is typically grounded in hospital routine [13]. Medical evidence is rarely taken at face value, rather its validity is judged according to the perceived credibility of its source [26]. For example, patients may be perceived as good or bad historians and the validity attached to the account they give will vary accordingly. Evidence obtained from colleagues is appraised in a similar way — the opinions of someone recognised as a 'good doctor' may be highly valued, while those of someone less well regarded may be given less weight. Even the validity attributed to lab reports and image data may dependent on the reputation of the laboratory or the technicians associated with its production [77].

The process of establishing credibility is not carried out in isolation, often credibility (as well as meaning) will be negotiated between physicians, and between physicians and technicians. Other criteria used to ascribe validity will include personal experience and personal values, as well as values established at a group or organisational level. Typically the latter are embedded in the hierarchical and social organisation of a hospital, so, for example, the opinion of a student doctor will carry less weight than that of a consultant. Given that status is associated with credibility, participants in the diagnostic process will seek to maximise their reputation in this respect. In an analysis of the discourse between physicians, Cicourel states that encounters involve more than a simple transfer of information, participants also seek to promote their credibility by using tactics designed to demonstrate their competence [26].

Judgements about credibility are one way that evidence from a given source might not be taken at face value. Another is that interpretation depends on a wider understanding of the physical process that the evidence only partially describes. For example, in an ethnographic study of the use of an interpretive facility on an echo-cardiogram (ECG) machine, Hartland recognised that a normal cardiogram is not purely a function of traces on a readout, but an "achieved state" [61]. That is, diagnosis is arrived at by application of further information, including, in the interpretation of ECG traces, the patient's age and physique, and also the considerable experience of the cardiologist. There is no set definition of what constitutes a 'normal' cardiogram and a cardiologist may classify several conspicuously different traces as being normal. The interpretative facility lacked the ability to apply this 'common sense' knowledge to its diagnosis, and was thus

prone to error. Occasionally the machine would interpret a normal cardiogram as being indicative of an abnormality and as a consequence physicians' would often edit patients' records so that an erroneous readout was not included. In hospitals with plentiful cardiology experience the performance of the system was judged sufficiently poor to be largely unused.

In a social milieu where maintaining and establishing credibility is of considerable importance a decision support system is likely to be at a disadvantage. While in essence a MDSS will be performing a task similar to that of a colleague, it will not possess the social skills employed by more regular colleagues to negotiate meaning and assert competence. One of the findings in a survey designed to examine physicians attitudes to medical consultations was that physicians prefered the concept of a system which functioned as much like a human consultant as possible [121].

Where establishing the credentials of a system proves difficult there is the potential for users to make 'errors of trust' or 'errors of mistrust' [100]. An error of trust occurs if the faith someone has in a system is unwarranted by system performance – with the effect that erroneous system results might be accepted. An error of mistrust refers to treating the results from a relatively accurate system with skepticism. Muir suggests that an attempt should be made to 'calibrate' a user's trust – that strategies should be employed to ensure that users have an appropriate level of trust in a system's functions [100]. This may take the form of explaining, as far as possible, how the system works, its operational constraints, how these might be affected by changes in the system's operating environment, to make explicit the system's domain of competence, it's history of competence and what criteria have been set for acceptable performance. Thus users will be in possession of the information required to make an informed judgement as to the appropriateness of system output.

Users of the system will have their own criteria for acceptable system performance – with expectations that may be too high (perhaps because of fear that the machine will usurp the user's role) or too low (if the user is unskilled, or finds a task tedious and thus prefers to abrogate responsibility to the machine). Muir argues that the root of this problem is that people find it difficult to cope with situations where a user's authority may be delegated to a machine, but where the user is still held responsible for the decision. She suggests that people ought to be respected as the ultimate decision makers, and that decision making systems should be designed and presented as tools that users can choose to employ in the process of solving a problem [100].

However, according to Hollnagel it is inevitable that some responsibility is delegated to the computer when decision support systems are used [66]. He suggests that responsibility in decision making requires that one has knowledge of "what the decision is about...the conditions, the alternatives, [and] the consequences", and that "the problem of responsibility, therefore, becomes one of understanding correctly the information and knowledge that describe the situation".

Usually the role of a decision support system is to overcome limitations in human cognition and so promote more effective action within some problem domain. The advice generated (or decision made) is produced by a transformation of the available raw data. The human agent with notional responsibility for the decision may neither have access to the original data available to the system, nor to a full understanding of its decision-making procedure. Even where the raw data is available, the time frame for the decision might preclude a human agent from making a thorough assessment. Hollnagel argues that providing support for information processing in this way necessarily entails surrendering some degree of control over the situation, and hence also some degree of the responsibility [66].

He suggests three possible approaches to this problem. Firstly, to define precisely divisions of responsibility according to a functional decomposition based on task analysis. Secondly, to make the decision-making process employed by the system transparent, so that the quality of its decisions can be assessed. Finally, if decision support systems are to have responsibility, then they should also incorporate rules describing the limitation of its applicability.

The first approach has been adopted in the development of decision support aids for mammography. A distinction is drawn between detection and classification aids [58] which is broadly in line with the (notionally distinct) perceptual and analytical tasks involved in image interpretation [87]. Detection aids give the location of features or regions that the system has identified as potentially suspicious with the aim of ensuring a more complete examination of the mammogram. A judgement about whether any action should be taken remains solely with the reader. In contrast, classification aids give an indication of the degree of suspicion associated with a feature or region with the express aim of assisting a reader decide the most appropriate course of action. This distinction and its rationale is discussed in greater depth in Chapter 2.

In the case of the application of detection aids, the original data is available, the human decision-maker is skilled in its interpretation and the use of the aid is discretionary. However, responsibility issues are still of importance. The film reader must choose whether or not to consult the system, and they must be

aware of the scope and limitations of the system's abilities when interpreting its response.

The work presented in later chapters suggests that the distinction between detection and classification aids is difficult to maintain in practice, that users form a (sometimes erroneous) model of the system's credentials based on observations of its behaviour, and that the system's actions need to be accountable in order for an appropriate interpretation of those actions to be made.

1.4 Evaluation strategies

Both quantitative and qualitative approaches are advocated for the evaluation of MDSSs. The former are concerned with enumerating outcomes, for example, diagnostic efficacy and cost effectiveness [93, 136]. The latter with broader social and contextual issues, such as implications for work practices, and user acceptance, and to enable a critical appraisal of design rationales (for example, by identifying discrepancies between intended and actual usage) [52, 62].

Recently it has been argued that evaluation in Medical Informatics should involve a complementary synthesis of these approaches [52]. The evaluation of the PROMAM system has proceeded according to this model. In the following sections the advantages and limitations of both qualitative and quantitative approaches are discussed. The discussion of qualitative methods focusses on the evaluation of aids for mammography, however, the principles are similar for other types of MDSS. The discussion of quantitative methods has a broader focus. This, in part, is because qualitative methods have not hitherto been applied to the evaluation of decision support systems for mammography.

1.4.1 Quantitative

Quantitative evaluation of detection aids for mammography encompasses a broad scope of activities with a variety of aims, from laboratory benchmarks, small scale experiments with clinicians through to clinical trials.

Laboratory testing involves a comparison of performance of a system against some previously determined standard. In these types of investigation, 'ground truth' is defined according to pathology data obtained at assessment clinics. Where certainty about normality is required, then older films may be used to avoid the possibility of interval cancers being present [70]. Laboratory testing allows performance bench marks to be established. This might be done as part of a training-test cycle, or to establish minimum requirements for clinical test-

ing [93]. A further goal might be to make comparisons between different algorithmic approaches, however, care has to be taken interpreting results obtained from different data sets [19].

Clinical testing involves comparing film reader performance with and without the assistance of a decision aid. One advantage of laboratory evaluation over clinical evaluation is that many trials can be performed without the involvement of film readers — sparing their time and any associated cost. A major disadvantage is that it is difficult to relate system performance to potential gains in the performance of readers when using the system. Typically, laboratory tests are simply used as a metric for comparing systems when using different parameters or techniques. However, with the prospect of clinical evaluation, it is particularly important to be able to choose an operating point that would be most likely to improve film readers' performance. The option of repeating clinical tests for many operating points is often not a tenable one. The principle method for assessing laboratory and clinical performance involves ROC analysis, however, qualitative methods have also advocated for laboratory testing. For example, Hume suggests an approach whereby the types of system response (for example, FP prompts for, say, vascular calcifications) are enumerated for different operating parameters and used as a heuristic for determining a setting that might be acceptable in clinical trials [69, 68]. A suggested heuristic for determining that a system might effect improved sensitivity in practice is its performance at detecting interval cancers [64].

A further related problem is that it is not possible to know what minimum level of system performance is required to effect an improvement in reader performance. This makes it difficult to establish realistic goals for system development and laboratory benchmarks. One solution is to perform experiments using simulated prompts with the desired performance characteristics [70, 4] [4]. Such 'small-scale' experiments, where the both the system's responses and the test set used is carefully manipulated, also help bridge laboratory testing and full clinical evaluation. A disadvantage of this approach is that some degree of 'realism' is necessarily sacrificed. For example, the incidence of breast cancer in the UK screening population is approximately 0.5% [46], so to achieve a statistically significant result, either heavily biased test sets must be used, or a full clinical trial must be conducted involving at least 90,000 women [131].

[4] These and other studies are discussed in greater detail in chapter 2

1.4.2 Qualitative

Initial approaches to the evaluation of clinical information systems (including MDSSs) involved the adoption of the Controlled Clinical Trials (CCT) paradigm [136]. Heathfield *et al.* contend that the underlying assumption for using CCTs is flawed, that is, that an information system is somehow comparable to a drug, and it should be evaluated in the same way. While CCTs give an indication of relative performance gains due to the use of a system, they do not reveal *why* a system has succeeded or a failed [62]. Understanding 'why' is essential if deficiencies in a system are to be identified and repaired. Forsythe argues that the 'double-blind' nature of CCTs specifically preclude the possibility of qualitative investigations being undertaken [51]. Given that many of the difficulties associated with the successful deployment of MDSSs have a contextual or social basis (as argued in section 1.2), much is precluded by neglecting to take a qualitative approach.

The question remains of identifying appropriate methods for qualitative evaluation. Ethnography has achieved prominence in the field of Computer Supported Collaborative Work (CSCW) and is finding an increasing role within medical informatics [52].

Ethnography is a naturalistic social science technique initially developed by anthropologists for studying native cultures. It involves the prolonged engagement of a field worker who takes on the role of 'participant observer'. Ethnography utilises the strategies people naturally use to make sense of novel social situations, where there is the need to consciously reflect on the meaning of the behaviour of co-participants, build models of roles and an understanding of relationships. Generally, this process results in an acclimatisation, where modes of interaction become habitualised, cultural values accepted, and where conscious reflection becomes largely unnecessary. The ethnographer maintains an objective distance by formalising this reflective process, for example, through note taking and model building [60].

Although ethnography is often concerned with generating a 'thick' description of a culture, it is not a purely descriptive technique. An element of analysis is also involved both to situate findings within existing social theory, and to generate further theoretical descriptions. The balance between description and analysis is dependent upon the goals of the research. Analysis in ethnography is an activity which occurs before the investigation begins, to generate an initial focus for the research, and continues throughout the course of the field work. In practical terms periods of field work are interspersed with periods of reflection, where the

gathered data is organised, inferences made, and explanatory models constructed. Further field work is directed by the salient issues uncovered. This iterative process gives ethnography its characteristic 'funnel' structure as the investigation becomes progressively more focussed with time [60].

Forsythe draws a parallel between the utility of anthropological techniques for systematically uncovering how knowledge is distributed and organised within a culture and the requirements of Knowledge Acquisition (KA) for expert system development [50]. Another strength of ethnography is that it can reveal the tacit and often complex constructs which underpins work and working practices, and so may be used to challenge the a-priori assumptions of system developers [48] and to inform design [44]. For example, Forsythe reports on a requirements elicitation exercise to inform the design of an intelligent information system for migraine sufferers [48]. Ethnographic techniques were used, including 80 hours of documented observations of doctor patient communication in five different clinical settings, as well as semi-formal interview with patients. The results of this work challenged many of the tacit and explicit assumptions made by the designers and, indeed, the physicians participating in the project. For example, that the explanations demanded would be in the form of 'textbook knowledge'. The investigation revealed that patients were far more concerned to understand with the condition would affect their everyday life, and whether it was life threatening, rather than with the physiology of the condition.

Typically there are some problems associated with using ethnography for information system design. Ethnographies can be voluminous, and are generally of a discursive nature. This makes it difficult to integrate information from an ethnography directly onto into the design process [115]. Ethnography as a technique is best suited to small, self contained, confined environments where there is a clear focus of attention for participants and a visible differentiation of tasks. Scaling up to provide information concerning large, diffuse organisations which have many functions is difficult, and would almost certainly lead to a loss of representativeness and depth [67].

Ethnography is based on a non-interruptionist and non-interventionist philosophy. However, if the purpose of an ethnographic study is to inform the design of a system which will replace labour, then the position of a field worker as a 'guest' in the workplace may be difficult to maintain. Furthermore, a good ethnography is likely to depict workplace practices as comprising a balanced web of critically interdependent activities, thus making objectivity when instrumenting change difficult [115].

1.4.3 The evaluation of the PROMAM system

Previously, the evaluation of prompting systems has largely involved the use of quantitative methods in order to demonstrate performance gains for assisted reading. In contrast, a mix of qualitative and quantitative methods are applied to the evaluation of PROMAM in order to explore how prompts are interpreted and used in practice. Initially an ethnographic investigation was conducted into reading practices in six UK breast screening centres. This enabled the interpretation of prompting studies to be informed by an understanding of how unsupported reading is organised. Three separate investigations into prompting were then conducted using prompts generated by the PROMAM system and film readers recruited from the UKBSP. The primary means of data collection were qualitative, and included: audio taped debriefings following prompted sessions, questionnaires designed to capture reader's subjective appraisal of the system, the use of 'think aloud' protocols where subjects were encouraged to verbalise their reasoning, and observations of subjects' use of the system.

1.5 Overview of thesis

Chapter 2: Psychological accounts of radiographic expertise The literature concerning the perceptual and cognitive mechanisms underpinning visual expertise is reviewed. This psychological description is used in later chapters to explain particular sorts of accountability that film readers are shown demonstrate for the content of mammograms they examine. Also considered are how theories of reader error are used justify the development of tools to independently support the notionally distinct (in human observers) processes of detection and classification. It is argued that two widely used empirical approaches to deciding whether and error of detection or classification has occured may be inconsistent, casting doubt on the validity on this distinction. Finally, previous quantitative studies of prompting are discussed, with particular reference to studies designed to show how good prompt generators need to be in order to be effective.

Chapter 3: Work practices in breast screening An ethnographic investigation into work practices in six screening centres was conducted. The aim of this investigation was to ground the interpretation of subsequent prompting experiments in an understanding of how praxis and the working ecology of screening artefacts supports the decision-making process. Although it is often assumed that the primary goal of collaborative practice in screening

(double reading, for example) is to enhance the sensitivity of the screening test, other rationales were also identified, including: training, providing reassurance, performance monitoring and demonstrating accountability within a community of practice. Readers demonstrated a reflexive approach reading by seeking to address perceived weaknesses in their reading skill and by responding mammograms in a manner commensurate with the particular difficulties posed by each individual case. Readers also recognised that some types of practice may bias decision-making, and demonstrated approaches to organising their access to evidence in order to minimise potential biases.

Chapter 4: Subjective responses to prompting Experiment were designed to evaluate three different prompting regimes generated by the PROMAM system. Rather than testing the effect of PROMAM on readers' performance, the aim was to elicit readers' subjective appraisal of the different system configurations. One specific goal was to examine how readers responded to FP burdens which, according to previous quantitative work, would preclude an improvement in performance. The results suggested that readers' interpretation of prompting information was more complex than had previously been thought. Readers appeared to be able to make inferences about the sensitivity of a system from only a limited number of examples of malignant cases. Despite this perspicuity, however, they developed only a naive understanding of the system's function, leading to an inappropriate interpretation of some system responses.

Chapter 5: Detailed analysis of prompted cases The aim of this exercise was to explore in greater detail the issues raised by the previous investigation. In particular, to further understand how subjects use, make sense and learn from the prompting information supplied by the system. This was done by eliciting a verbal commentary from three subjects concerning their interpretation of prompted cases. Although prompting systems aim to address the perceived need for assistance with detection, it became apparent that readers used prompting information to resolve many different types of context dependent difficulties which they routinely encounter. Tolerance to FP prompts was demonstrated when prompts corresponded with features for which readers held themselves accountable, for example, where their attention is warranted so that some apparent suspicion might be discharge. Similarly, subjects demanded an account of the system's responses, and were dissatisfied with the prompts where an account of system behaviour

was not readily apparent. Common misapprehensions concerning the role and scope of the PROMAM system were identified, and used to inform the creation of a training package.

Chapter 6: Pre-clinical trial of the PROMAM system Pre-clinical trials provided an opportunity to examine the validity of the conclusions drawn from the previous work under more realistic reading conditions. Interview, questionnaire and observation data were collected during the course of a prompting experiment involving 2000 archive cases and 5 screening film readers. The data revealed strategies employed by readers to manage the FP burden of the system, and also confirmed that readers used prompting information to support their classification decisions, indicating that the distinction between detection and classification aids may be difficult to sustain in practice.

Chapter 7: Conclusions and further work This chapter examines the contribution made towards an improved conceptual understanding of prompting and discusses value of qualitative methods in achieving this end.

Chapter 2

Psychological accounts of radiological expertise

In this chapter the literature pertaining to the perceptual and cognitive mechanisms underpinning radiological expertise is reviewed. Much of this work has its roots in studies of human vision and attention, and so a brief account of these topics is presented first. Expertise in the interpretation of medical images is considered with respect to the influence of prior information, the organisation of visual search and the criteria for its termination. One of the goals of studying radiological expertise is to better understand the reason for observer error and thereby effect changes in practice that improve performance. A rationale for computer decision-aids is given in respect of psychological accounts of observer error. Finally, investigations concerning the effect of the use of computer decision-aids are considered.

2.1 Human vision

The limits of visual acuity does not necessarily determine the ability of an observer to process information for some region of a visual scene. Mackworth defines the 'useful field of view' as the angle about a fixation point from which visual information can be processed or stored [90]. Depending on the visual task the useful field of view may be much smaller than the limit implied by the physiology of the eye. By comparing the performance of human observers in detection tasks involving the identification of a target from a background either sparsely or densely populated with irrelevant detail, Mackworth found a relationship between the useful field of view and the density of the items composing the irrelevant background. He suggests that the area over which visual data can be processed is dependent on how crowded the visual scene becomes, and hence on the 'difficulty' of the visual

task at hand. He also found that the extent of the useful field of view differs for both task and observer, and speculated that for a complex visual task, an observer's useful field of view would vary continuously depending on the component tasks necessary to achieve the overall goal.

Peripheral vision is viewed as playing a role in discriminating features of a visual scene that are worthy of detailed investigation. A distributed and partial interpretation is performed automatically across the whole of the visual field yielding sufficient 'at a glance' information to direct attention to regions of interest [7, 125]. Typically a shift of attention is accompanied by eye movements called saccades that bring the high resolution fovea to bear [53]. Saccades are under voluntary control, and take in the order of 250ms to plan and 50ms to execute — during execution visual processing is suppressed. Because of their ballistic nature, saccades will often overshoot or undershoot their target resulting in small corrective saccades to centre the fovea.

An important paradigm used to investigate radiological expertise is the tracking and recording of an observer's pattern of saccades and fixations. This is taken to indicate how the search of a radiograph is organised, and where attention is directed. The duration of visual dwell (for some empirically determined useful field of view) is often taken as a measure of the degree of processing afforded to a feature or region [87].

2.2 Visual search

Human visual search has been studied using experiments whereby subjects are asked to identify simple targets embedded in a background of distractors. Typically both targets and distractors are discrete shapes often presented in a regular array on a featureless background. The response time (RT) between onset of the image and a subject's decision is measured as the size of the distractor set is varied over a large number of trials.

An important finding of this work is that some combinations of targets and distractors produce no change in response time as the set of distractors is increased. Discrimination of the target is perceived to be effortless, and subjects often report that the targets appear to 'pop-out' of the display. Other combinations show a linear relationship between response time and set size. The former case typically occurs where the target's visual character is very different from that of the distractor, the latter where targets and distractors have features in common. A further finding concerns search for targets that are uniquely identified

by a conjunction of properties (e.g. a red square) in a set of distractors that contain these properties singely (e.g. either square or red objects). Response times increase with distractor set size, but not as severely as for a search for feature singletons amongst similar distractors (see [133, 135] for reviews of findings using the visual search paradigm).

To account for these results it has been suggested that the visual system comprises of a series of 'channels' or 'feature maps' each dedicated to detecting some basic property of a visual scene (colour, size, orientation and so on) that operate in parallel [123]. For feature singletons (i.e. targets that can be identified by some single unique property) detection is simply a matter of checking the appropriate channel for its contents. Where the properties of the targets and distractors are similar, then a single channel may have representations of a number of features that have to be attended to serially in order to make a discrimination. Thus a distinction is made between automatic and effortless 'pre-attentive' processing, and a serial process that requires attentional resources [123]. The greater efficiency demonstrated for identifying conjunction targets is explained if the information made available by pre-attentive processing can also be used in combination to guide attention. This would enable early consideration of the most likely candidates [133].

An important question is one of 'ecological validity'. How well do the phenomena discovered within the paradigm of the visual search apply to real world search tasks, such as the examination of a mammogram for signs of cancer by a trained observer? In contrast to the stimuli used in visual search experiments, natural 'targets' can be partially obscured and are often blended into an irregularly arranged heterogeneous background. Wolfe reports on a series of experiments that try to replicate the results of visual search using more naturalistic stimuli [134]. The stimuli consisted of computer generated images tiled to form continuous scenes that could be interpreted as stylised aerial views of a terrain with 'lakes', connecting 'rivers', 'mountains', 'cities' and obscuring 'clouds'. The basic results of visual search were confirmed. Search for feature singletons (a blue lake amongst polluted yellow lakes) give response times independent of the size of the distractor set. Similarly, conjunction searches, and searches for features with weak defining characteristics were also consistent with the visual search paradigm.

However, one unexpected result was obtained for a search task that mimicked the detection of an "S" in a distractor set of mirror image "S"s. Where the targets and distractors formed part of a continuous river system performance was significantly worse than when the search was for discrete "S" shaped lakes

amongst mirror "S" lakes. In both cases the theoretical model of visual search would seemingly imply the same serial search task. However, Wolfe suggests that although in a typical serial search task all items are identified as candidates, at least their location is available pre-attentively, thus giving a definite set of features to be searched. He argues that in the continuous condition no discrete candidates are identified and so an attentive process is required both to locate and segment candidates, as well as to test for targets.

2.3 Attention

In the visual search literature a distinction is drawn between pre-attentive and attentive processes. The former are passive and automatic, the latter capacity limited and effortful. In this section the literature on attention is briefly reviewed. The study of attention is concerned with the constraints upon human ability to process information, and in consequence, how perception is selectively organised. The difficulty of dividing attention (consider holding two simultaneous conversations), and the differential effort required by disparate tasks (compare the ease of reading with the difficulty of mental arithmetic) are familiar to all. Kahneman accounts for processing restrictions by postulating a central processing mechanism that has a fixed capacity [74]. This idea is rejected by Neisser, who notes that effortful tasks can become automatic with practice. He suggests instead that the effort associated with particular tasks, and with divided attention, are proportional to how well practised the performer is [101]. However, one would presume that there are limits to skilled performance, and that the idea of limited capacity has value for the performance of a particular task at a given time by a given person.

The idea that attention involves selective processing of sensory information is less controversial. Two important questions that are addressed by the literature concern the scope of selection (what can be attended to) and its organisation (the governing of what is attended to). These aspects are considered in turn below.

2.3.1 The scope of attention

Theories of spatial attention draw an analogy with a 'spotlight' that has a variable beam and shape which can be moved across space independently of eye-movements [108]. Thus what is selected by attention is the region in a scene 'illuminated' by the spotlight. Of interest are the limits of spatially deployed attention. For example, experiments reveal a spatial gradient in attentional ef-

ficiency. If an event is cued (sometimes incorrectly) then a relationship can be demonstrated between the spatial separation of the cue and the target event [36]. A spatial limitation can also be demonstrated in dividing attention, for example, when observers are asked to attend to two stimuli simultaneously, efficiency is reduced as a function of spatial separation [65].

Within the attentional spotlight a further degree of selectivity is possible, based on the perceptual organisation of components of the stimuli. Where a distinction can be made between objects based on gestalt properties, continuous features and perceptual similarities, then these objects may be selectively attended to even though they occupy the same location. For example, Rock and Gutman presented subjects with a series of stimuli consisting of two overlapping shapes, each drawn in a different colour. Subjects were asked to give an aesthetic judgement of shapes drawn in one colour, thus directing their attention to only one of the shapes in each pair. Afterwards subjects were given a surprise recognition test, and it was found that they scored significantly better for the attended to objects [111].

In another experiment, Duncan presented subjects with a number of images consisting of a rectangle containing a sloping line [38]. The rectangle and line each had two distinctive properties. The rectangle was either tall or short, and had a gap in either its right or left side. The line was sloped either to the left or the right, and was either solid or dashed. Subjects were asked to report two attributes from each display. Their accuracy was improved if the attributes pertained to the same object (for example the slope and texture of the line) rather than if they pertained to different objects (for example, the slope of the line, and the size of the rectangle).

Thus attention can be selectively allocated to specific objects with certain spatial limitations. Further issues concern the control and the guidance of attention. That is, the governing of what it is that will be attended to at a given moment, and of what will be attended to next.

2.3.2 The control of attention

A distinction is often drawn between endogenous and exogenous control of attention. The former refers to the 'top down', or 'goal directed' allocation of attentional resources, the latter to the capture of attention by some salient property of a stimulus in a way that is at odds with the goals of the observer [39]. For example, Jonides and Yantis demonstrated that the abrupt onset of a distractor could interfere with the primary attentional task of the experiment [73].

However, Folk *et al.* challenged the notion that attentional capture in this way is not entirely exogenous [45]. Their criticism of the paradigm used to demonstrate exogenous control is that typically both the distractor and the target are both signalled by an abrupt onset. The nature of the task (detection of an abrupt onset target) biases perceptual readiness in favour of abrupt onset events and is thus prone to interference by an abrupt onset distractor. Folk *et al.* showed that the detection of a target signalled by colour rather than by abrupt onset was not interfered with significantly by an abrupt onset distractor, concluding that involuntary capture of attention is not 'hard wired' and inflexible, but dependent on the 'control setting' established by the nature of the task. They suggest that, in the absence of a pre-established control setting, defaults may be used that are based on "long term biases". The mechanisms of attentional control is viewed as demonstrating "a delicate and efficient balance between the rigidity necessary to ensure that potentially important environmental events do not go unprocessed and the flexibility to adapt to changing behavioural goals and circumstance".

Wolfe assimilates these ideas in his 'guided search' model of attention [133]. Sensory information is encoded by a series of feature channels corresponding to basic properties of an image (eg colour, orientation etc.) which are used to provide a measure of how unusual an item is in its current context (the difference in basic visual properties between an item and its neighbour). This gives a degree of 'bottom up' activation. Goal derived information is used to select which of the feature channels are most likely to contain an item of interest and thus contributing 'top down' activation. These two sources of activation are combined in an "attention map" visualised as a series of peaks and troughs that indicated the likelihood that attention will be satisfied by examining a particular aspect of a scene. Attention is viewed as a serial process that searches each region of activation in turn, the magnitude of which governs the order of the search. In this formulation, the control of attention is directed by the goals of the observer, and by the properties of the image.

So far, the control of attentional selectivity has been discussed with respect to momentary appraisal of briefly presented stimuli. The question remains of how the deployment of attentional resources is organised over longer time periods where there is more persistent involvement with a number successive stimuli.

Guidance of attention can be thought of in terms of perceptual readiness, that is, a susceptibility to certain types of information based upon prior perceptual episodes. Priming is defined as occurring when one stimulus (the prime) affects the processing of another (the target). Johnston and Dark identify four types of

priming phenomena reported in the literature: modality, identity, semantic and schematic [72].

Modality Priming can heighten responses due to a particular mode of sensory input. For example, if attention is pre-engaged in a particular modality then responses to startle stimuli are greater for that modality than for others.

Identity For example, prior exposure to a word improves detection in a subsequent recognition test. The effect is reduced if the prime is presented via a different modality, if a different type face is used, or if the shape of the word is different.

Semantic For example, processing of a target word can be improved if the subject is first primed by a semantically related word. (eg: Bread primes Butter).

Schematic Refers to the representations of temporal or spatial relationships between objects or events. Active schemata can have biasing effects on the processing of test stimuli.

As well as attention being directed moment by moment as an immediate consequence of interaction with a stimulus, priming may be effected by chronic perceptual biases. Such biases also operate at different levels of processing. An example of a 'low level' bias is the enhanced performance found in recognition tests found when frequently-occurring English words are used as targets [18]. At a higher level, chronically persistent schemata can reduce the attention given to stimuli in an experimental task, and direct it to stimuli relevant to the subject. For example, Bargh reports the results of a dichotic listening task for subjects who were categorised in a prior test to be aschematic or schematic for the attribute of independence. The latter group's performance was better when words relating to the attribute of independence were presented in the attended channel, and worse when independence related words were presented in the disattended channel [6]. Chronic perceptual biases explain of some seemingly paradoxical aspects of divided attention, for example, a person can attend exclusively to a single conversation in noisy and crowded room, and yet have their attention captured by mention of their name in a different conversation within earshot[72].

Neisser phrases this model of the guidance of attention with respect to visual stimuli in the following way:

> "In my view, the cognitive structures crucial for vision are the anticipatory schema that prepare the perceiver to accept certain kinds of

information rather than others and thus control the activity of looking" [101].

He describes a perceptual cycle whereby an observer's exploration of sensory information is directed by an anticipatory schema, which is in turn modified by the resultant perceptions. According to Neisser, the 'selectivity' of attention is governed by a observer's active schema determining what information they are able to accept at a given moment. Information pick-up is directed by expectations, but not entirely controlled by them. Thus it is not the case that observers literally see only what they expect to see — it is possible to be surprised by the unexpected. However, if no active schema is available for a particular type of information, then that information will not be perceived, surprising or not.

2.4 Conceptual models of radiographic interpretation

Gale *et al.* recognise that the perceptual cycle of Neisser can provide a theoretical basis for understanding the nature of skill in the interpretation of medical images. According to this view, anticipatory schema are selected on the basis of prior information (experience, case history and so forth) which in turn serve to direct how an image is initially scanned, and critically, what information is subsequently picked up [54]. Thus it is possible to explain reports of trained observers overlooking gross abnormalities (for example, [124]) if the observers' active anticipatory schema did not allow for the possibility of their detection. In such cases, Gale *et al.* suggest "...it is the initial schema of the observer which overcomes the actual real data" [54]. Similarly, Lesgold et al distinguish between the performance of expert and novice radiologists on the quality and appropriateness of the schema triggered by examining a given radiograph [89].

Gale suggests a conceptual model for describing radiologic expertise. A pool of possible hypotheses (or schema) conditioned by experience, prior knowledge and expectations are available to the observer. An initial hypothesis is selected, and is confirmed or rejected by an initial global impression. Here 'global impression' refers to the role information obtained at the periphery of vision plays in directing attention. Further hypothesis may be selected, or a foveal search initiated to gather additional information. Continuing rounds of global impression, foveal search, hypothesis selection and rejection continue until the observer is satisfied that sufficient evidence is available to confirm one of the considered hypotheses [53].

Similar models are proposed by Kundel [88] and Nodine [104]. Kundel describes global impression in terms as a "pre-attentive filter" that "limits the number of image features or potential target sites receiving attention" and Nodine suggests that global analysis of the image leads to the detection of conspicuous abnormalities and to the identification of anatomical landmarks. Thus all agree that a global impression phase serves to orient attention to significant features within the image. Kundel, Nodine and others, through eye-movement studies, go on to investigate in detail the subsequent stages of visual search and are able to make a qualitative distinction between different phases of search, and between different types of purposeful glances.

In the following sections the literatures relating to three important aspects of Gale's conceptual model are considered in turn. The first examines how expectations and prior knowledge influence the interpretation of medical images. The second explores how the visual search of images is organised. The third is concerned with criteria for the termination of search.

2.4.1 Expectations

2.4.1.1 The effect of clinical history

Where supplied, a clinical history forms part of the information available to radiologists to orientate their subsequent examination and interpretation of radiographs. A history will typically consist of clinical facts derived from sources other than the image itself; these may include details of previously identified conditions, a family history, the results of a physical examination and of previous of medical tests (which may may also have involved imaging). Of interest is the way this information influences decision-making, and in particular, whether decision-making tends to be enhanced or biased.

In a study of the effects of clinical history on the interpretation of chest radiograph Doubilet and Herman covertly introduced seven test cases into the routine workload of a hospital radiology department on a number of separate occasions [35]. Six of the cases contained a single abnormality, and one contained two abnormalities. Cases were chosen carefully such that the abnormality was subtle, but easily interpreted if detected. Each of the six cases containing a single abnormality was read four times with a clinical history suggestive of the abnormality present, and four times with an unrelated history. The case containing two abnormalities was read four times with a history suggestive of one of the abnormalities, and four times with a history suggestive of the other.

A significantly greater sensitivity was achieved when cases were read with a

suggestive history (72%) compared with when they were read with an unrelated history (16%). Although FP decisions were made, the design of the trial precluded the calculation of a FP rate. However, all of the FP decisions were consistent with the unrelated histories.

Doubilet and Herman pose the question:

> "Does the suggestive history make the reader carry out a more careful visual search for abnormalities related to the history, or does it make the reader more likely to interpret a questionable finding as abnormal?".

Since the abnormalities were specifically chosen to be perceptually challenging but diagnostically unambiguous, the observed improvement in sensitivity is attributed to improved visual search. However, the authors recognise that using diagnostically unambiguous cases precludes attributing performance changes to an influence on subjects' criteria setting, and so the possibility of this effect cannot be ruled out. In fact, the influence of unrelated histories on FP decisions might be explained in this way.

An important question concerns what can be legitimately inferred about the effect of history on interpretation under normal clinical circumstances. In Doublet and Herman's study, a comparison was made between the effects of suggestive and unrelated clinical histories. However, the unrelated histories were carefully chosen to be suggestive of conditions that were not indicated clinically, or by interpretation of the images, and therefore might be better described as 'misleading' rather than 'unrelated'. It is plausible that radiologists may encounter misleading histories as part of their unmanipulated caseload, or that additional abnormalities may be revealed by imaging which are neither clinically manifest nor related to the initial reasons for performing the investigation. If such occurrences are relatively infrequent then radiologists may neither have the opportunity nor feel the necessity to develop skill in this respect. The effect observed may be due to misleading histories degrading performance, rather than suggestive histories enhancing performance.

Babcock *et al.* also address the question of whether the effects of clinical history are due to improved visual search, or changes in subjects' confidence thresholds for abnormal findings [5]. They examined experimentally how manipulating clinical histories affected the interpretation of pediatric chest radiographs for childhood bronchiolitis.

The results showed that significantly more features of bronchiolitis (in the order of 25% to 50%) were reported present on equivocal normal radiographs

when they were read with a positive history, compared with when they were read with a negative history. Confidence ratings for the presence of bronchiolitis also showed a statistically significant increase for equivocal normal cases with a positive history. However, no significant differences between conditions were found for the interpretation of equivocal bronchiolitis cases. It appears that a positive history has the effect of increasing suspicion where the evidence for bronchiolitis is minimal, but that a negative history does not have the effect of lowering confidence where evidence of bronchiolitis is more pronounced. The authors suggest that overall, the effect of misleading histories was to decrease the overall performance by increasing the FP rate.

These findings are discussed with reference to the distinction made by Doubilet and Herman between histories improving detection and altering confidence thresholds. Babcock *et al.* state:

> "Because our normal radiographs did not have any features of bronchiolitis, the reporting of these features in the presence of a positive clinical history could not be due to a change in detection ability, but must be due to an altered threshold for calling a questionable finding abnormal."

It is worth considering the differing circumstances of the two studies. In the study conducted by Doubilet and Herman, the diagnostic task was an open-ended one — subjects examined the radiographs in a context where they might expect to make many different types of finding. In contrast, Babcock *et al.* directed subjects to examine radiographs for evidence of one particular clinical entity. This was done to the extent that subjects were asked to examine and rate specific features known to be associated with this condition. In consequence, subjects were not only told what to look for, but also where to look. Thus the results obtained do not rule out the possibility of clinical history improving radiologists' detection ability since the design of the experiment largely precludes observing any such effect. Furthermore, suggesting that the effects were demonstrated on normal radiographs which "did not have any features of bronchiolitis" might be a little misleading. Histories were actually manipulated for 'equivocal normal' cases, and presumably some features of bronchiolitis must have been present for the initial 'equivocal' classification to have been made. The results might be more accurately stated as suggesting that in the presence of minimal signs of bronchiolitis a positive history would make a judgement of the presence of bronchiolitis more likely.

A study by Elmore *et al.* reveals a similar effect where histories of mammography cases are manipulated [41]. This study differs from the two previously reviewed in that instead of testing 'suggestive' or 'positive' against 'unrelated' or 'negative' histories, the effect of presence of history was compared with the effect of absence of history. Clinical history probably plays a different role in screening as opposed to diagnostic radiology. In diagnostic radiology the purpose of imaging is to confirm a suspicion based on clinical findings. Because clinical history is actually part of the rationale for performing the test it also provides a basis for subsequent interpretation. In contrast, where imaging is performed as part of a screening test the idea is to discover disease entities that do not yet have a clinical manifestation. In screening, interpretation may be conditioned by the range of disease entities that it is possible to detect, where as diagnostic radiology, it may be conditioned by the particular focus of the test. Although occasionally a screening case may be accompanied by a suspicious history which might influence interpretation, it would be undesirable for radiologists to be influenced by the absence of a suggestive history.

Elmore *et al.* found that an alerting clinical history led to significantly more changes towards an abnormal rating, and to an increased number of recommendations for work-up (i.e. for performing additional investigations such as ultrasound or biopsy). In contrast, a non-alerting history lead to more changes towards a normal interpretation, but no significant differences in the type of management decisions was observed.

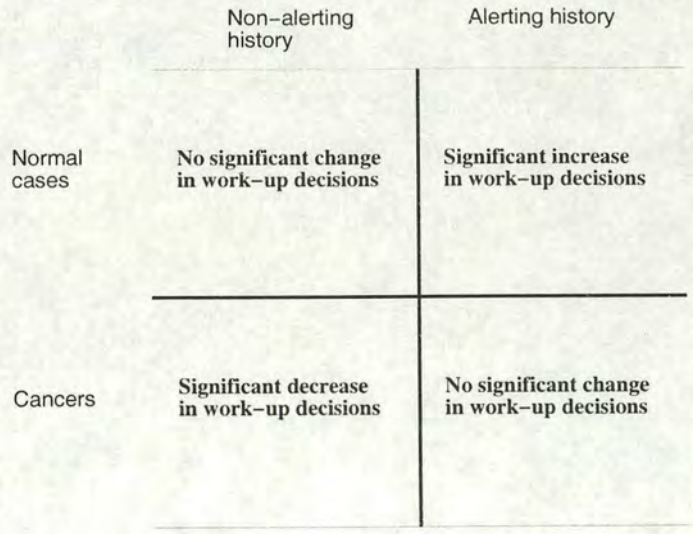


Figure 2.1: Summarises the comparisons made between work-up decisions for subcatagories of cases presented with and without a clinical history. (Elmore et al [42]

Although Elmore *et al.* do not attempt to draw conclusions about the causal mechanisms of the observed biasing effects, it is instructive to examine their results in this way. Figure 2.1 shows some interesting relationships between subset categories. For normal cases there was a only a significant change in work-up decisions where there was an alerting history. Conversely, for cancers, there was only a significant change in work-up decisions where there was a non-alerting history. This suggests that the nature of the biasing effect is dependent on the initial suspicions of the observer. Where it is likely that suspicion is high (as one might expect it generally would be for cases showing a cancer) the moderating effect of a history appears to be towards making a normal decision. Conversely, where suspicion is low, the moderating effect appears to be towards abnormality.

In conclusion, Elmore *et al.* explore the implications of these results for clinical practice. A reading protocol is recommended whereby the radiologist examines the film and reaches a provisional conclusion before attending to the history. It is suggested that by allowing an initial objective appraisal the biasing effects of history may be mitigated.

2.4.1.2 Disease prevalence

Clinical history is a type of prior information that is contingent and specific. No two clinical histories will be identical in detail (although there may be histories that are categorically identical), and a given history is only informative about the case to which it applies. There are other types of prior information that effect a more general orientation to the task of interpretation, one such is the frequency at which a looked for disease entity naturally occurs.

Egglin *et al.* performed an experiment involving the identification of pulmonary emboli (PE) from pulmonary arteriograms in which the prevalence of PE was manipulated in two conditions [40]. Statistical analysis revealed that subjects tended to be more suspicious of cases in the high prevalence condition, and also demonstrated a higher sensitivity. No significant difference in specificity was observed between the two conditions.

Three of the eight PE cases were apparently so obviously abnormal that a majority of subjects in both conditions were confident that an abnormality was present. In the remaining PE cases, large changes in interpretation (for example, two or more points on the rating scale) were made between conditions. Where such large changes in opinion were evident, they were usually made by three or more readers. The authors conclude that the results are case dependent, suggesting that "context bias was therefore most likely to affect the interpretation

of equivocal or difficult cases".

An emerging theme from the work so far reviewed is that the measured effect of manipulating prior information depends critically on the characteristics of the stimuli and the nature of the task. Improved search can be demonstrated where abnormalities are subtle but unambiguous and changes in observers' confidence are shown where search has been eliminated and the signs of abnormality are ambiguous. Babcock *et al.* showed that alerting histories affected normal cases and non-alerting histories abnormal cases. Finally Elmore *et al.* demonstrated that a manipulation of disease prevalence has its greatest impact on equivocal cases.

2.4.1.3 Influence on search strategy

The work reviewed so far demonstrate the effect of observers' prior knowledge or expectations on the interpretation of medical images. Kundel *et al.* demonstrated that prior information can also effect how visual search is organised [86].

Three radiology residents examined two sets of five chest X-Rays, one containing examples of lung nodules and the other examples of clearly visible, but diverse, abnormalities. Both sets shared two normal films in common. When searching the first set subjects were told to look for the presence of nodules, and when searching the second, that the films were either normal, or contained unambiguous evidence of disease.

Eye movement patterns were characterised as either circumferential, localised or complex. The circumferential pattern involves a broad sweep around the edge of the film returning to the starting point in about 15-25 fixations. The localised pattern involves closely grouped fixations 'sampling' a small area of the film, followed by a jump to another area. The complex category describe patterns of fixation that can be described neither as circumferential or localised.

A predominance of circumferential and localised patterns were discovered for the nodule search task, whereas for general search the complex pattern dominated. This was true also for the normal films included in both conditions.

The authors argue that the scan patterns are determined by both the informational requirements (obtaining additional detail, and the resolution of ambiguities) current during any given point of the search [1] and also by the properties of the image itself. Where the task is constrained (searching for nodules) a consistent pattern emerges since an observer can adopt a strategy that is efficient

[1]For example, occasionally subjects made saccades to an equivalent region on the opposing lung, indicating that a comparison with normal structures was being made.

with respect to that task. The lack of any identifiable or consistent strategy for the general search task indicates the complexity inherent in searching for, being attracted by and resolving possible areas of significance when potentially many interpretations have to be considered.

It is clear that prior information not only provides an orientation to likely findings and influences interpretation, but also has has a role to play in the organisation of visual search. The following section discusses in detail a literature that aims to identify and characterise the strategic approaches in observers' visual exploration of medical images.

2.4.2 Guidance of search

2.4.2.1 Peripheral information

Conceptual models of visual search in radiology stress the importance of peripheral vision for organising subsequent search of an image. Investigations have been conducted to determine the depth of information made available from the peripheral visual field, and how this information is used to guide visual search.

One important paradigm involves tachistoscopic presentation to estimate how much information can be gleaned by a single glimpse of a radiograph. Images are displayed for a duration of around 200ms which falls short of the time typically required for a saccade to generate a second fixation. Studies by Kundel and Nodine [85], and by Mugglestone *et al.* [98] report the effects on performance of limiting search time in this way for chest X-Rays and mammograms respectively.

Nodine and Kundel showed that surprisingly good performance was possible from briefly presented chest X-Rays [85]. A sensitivity of 70% was achieved for brief presentation, compared with a sensitivity of 97% for unlimited viewing. ROC analysis revealed better than chance performance for the interpretation of briefly presented images. The authors comment that as well locating potential abnormalities, occasionally enough information was available 'at a glance' for subjects to make accurate classification decisions. It was also found that subjects could selectively attend to and report abnormalities that were located as much as 30° away from the initial fixation point. Performance was reported to be better for larger, high contrast, circumscribed abnormalities, than for those that were smaller and less distinct, or where several disparate signs had to be combined for detection. The authors suggest that observers' attention was directed to regions of the image that show the greatest deviation from normal.

In a similar experiment, Mugglestone *et al.* compared 'flash' presentation with unlimited viewing of a test set of normal and abnormal mammograms [98]. Sub-

jects' interpretation was compared to a previously established radiological standard for each of the cases. The results showed that there was a high agreement with the radiological standard both with flash presentation and with unlimited viewing where there was a greater degree of contrast between target and background (i.e. where the breast tissue was fatty, and where the target abnormalities were masses). Where there was low agreement with the radiological standard in the flash condition, but high agreement in the unlimited viewing, the breast tissue tended to be dense, and the abnormalities tended to be subtle parenchymal distortions. Low agreement in both conditions tended to occur for microcalcifications. The method for tachistoscopic presentation involved transferring the films onto projector slides, resulting in a loss of definition that made microcalcifications difficult to detect in either condition. The authors suggest that although observers' performance is greater than chance for flash presentation, it falls short of the performance observed with tachistoscopic presentation of chest radiographs. This is attributed to the differing natures of chest radiographs and mammograms. Background breast tissue is relatively homogeneous whereas chest X-Rays contain more readily identifiable structure. It is suggested the detecting deviations from normal structure plays a lesser role in the search and interpretation of mammograms than for chest X-Rays.

In another study Carmody examined the limit on detection of lung nodules presented in the peripheral visual field. This involved tachistoscopic presentation of films with nodules at increasing distances from a fixation point [20]. The results showed that observers' ability to detect nodules falls with increasing distance from fixation. With a displacement of 5° the probability of detection fell by a half. An experienced radiologist and a less experienced film reader were used as subjects. Interestingly, the 'limit of detection' for the experienced reader was 15° and 10° for the less experienced subject.

Kundel *et al.* devised an experiment to establish the role periphery vision plays in directing visual search [88]. A method was devised of interactively adding or subtracting simulated lung nodules to a soft copy display of a chest radiograph depending on the direction of an observer's gaze. In this way it was possible to manipulate the presentation of nodules either to only occur in subjects' central or peripheral visual field. In the former cases the nodule would not be displayed until the centre of fixation subtends some minimum angle with the nodule location thus effectively reducing the subject's peripheral field of view for the nodule detection task. In effect, the nodule would appear when the fixation was close enough. In the latter case, they removed a nodule from the display if the centre of fixation

approached the nodule — the nodule would disappear if a subject's gaze were directed closely enough towards it. Using these paradigms two experiments were conducted. The first involved a series of trials where nodules were presented to subjects' central field of view with a series of decreasing apertures. If the central field of view was reduced below 5°, subjects' ability to detect a target, and the time to detection were significantly reduced. The second experiment compared periphery only, central field only and full field viewing. The mean time to hit the nodule was significantly lower where periphery viewing was eliminated. The authors concluded that periphery vision assists in the guidance of search for nodules.

2.4.2.2 Organisation of search

From an analysis of a series of eye-movement studies, Nodine and Kundel suggest a model for observers' search strategy in a nodule detection task consisting of a global impression phase, 'discovery search', and 'reflective search' [104].

If a nodule is detected within 1000ms of the onset of search, then this is attributed to detection made during the global impression phase. A minority of nodules are likely to be detected in this way. Discovery search begins shortly after, and is characterised by an interplay between 'survey' and 'examination' fixations. The former are interpreted as playing a role in re-orientating attention during search, and typically have a duration of 100-200ms. The role of the latter is believed to be analytical, and these typically have a duration of greater than 600ms.

Where a potential target is fixated in the discovery phase, and where there is initially little immediate accumulated dwell on the target area, but a larger accumulation of fixations at a later time, this is termed reflective search.

In a comparison of the detection of lung nodules with phase of search Kundel and Nodine demonstrated that low contrast nodules are generally picked up by reflective search, whilst high contrast nodules are generally picked up in the global phase and by discovery search. The authors suggest that moving from discovery to reflective search indicates a change in strategy from dependence on perceptual mechanisms for detection, to dependence on cognitive mechanisms to disambiguate potential nodules from normal structures.

Carmody *et al.* sought to test the hypothesis that the identification of lung nodules in chest radiographs involved comparisons between candidate nodule sites and normal structures in order to resolve ambiguities [21]. They drew two conclusions: firstly, because in tachistoscopic presentation nodule detection did not

increase with presentation times of 180ms or greater, all the information required to resolve a lesion is not present in the area of the image in which the lesion is embedded. Secondly, that less visible nodules received a greater number of comparative scans. They speculate that this is because radiologists' compare candidate targets with normal structures that might mimic nodules, citing rib endings and end on blood vessels as examples.

2.4.2.3 Novice and expert performance

One method of understanding how skilled search is organised is to make comparisons between the search patterns demonstrated by expert and novice observers.

Kundel *et al.* sought to examine the differences in scan paths between novices and clinicians with differing degrees of experience for a radiology search task [84]. There were some interesting differences. The initial (first sixteen) fixations of the radiologists were characterised as circumferential. In contrast, untrained subjects at first fixated some central portion of the image, and then made short jumps to other regions of the film in what is described by the authors as a strategy of 'local inspection'. Overall, fixations by the radiologists demonstrated a broad coverage, whereas those of novice observers showed a greater degree of clumping around the central region of the film. However, despite the broad coverage, even the experts did not sample all of the film with the fovea.

Scan paths used to examine abnormal images reveal how image properties can influence radiologists' search strategies. The experienced radiologist was able to rapidly fixate the abnormality, whereas the pattern exhibited by the untrained observer was similar to that for a normal film. Radiologists were able to locate lesions in fewer fixations on average than the inexperienced subjects.

In a lung nodule search task, Nodine and Kundel compared a random walk, visual search by a novice observer and visual search by an experienced radiologist [104]. The random walk is modelled on eye-movement data from experienced radiologists, and involves generating a random sequence of fixations based on empirically derived probability distributions for dwell time and saccade length. Comparisons with the random walk model indicate the extent to which visual and cognitive guidance contribute to search. It was shown that the experts' search is more systematic and covers more of the relevant target area, but even the experts' search was not exhaustive and covered only 60% of the lungs. The authors conclude that the experienced radiologist draws on prior knowledge of likely nodule locations to direct (and limit) his/her search.

In another experiment the random walk algorithm was compared with ex-

perienced radiologists searching for nodules in normal and abnormal films [104]. Time to hit 6 nodule sites (the nodules themselves in the abnormal films, and the the location of these nodules in the normal films) to within 3° was found to be 30% slower for the random walk model. The percentage of nodule sites fixated by the random walk model drops off sharply after 3 had been hit. In contrast, the number of nodule sites fixated by experienced radiologists for normal films remains high (>80%) for all six nodule sites. The authors suggest that the ability of experienced readers to consistently hit potential nodule sites even in normal films is driven partly by training and experience. Because these films are normal, there is nothing visually that would attract attention to produce this effect. Data from the abnormal films shows that the percentage of nodule sites hit by experienced radiologists falls off earlier than the random walk model, but less steeply. This is attributed to a termination of search due to nodule discovery.

Nodine *et al.* [105] report on an experiment comparing eye movement data of novices (T-E-), radiologists with training (T+E-), radiologists with training and experience (T+E+), and a randomly generated search path for a mammography task. The target lesions were masses, and two views (craniocaudal and mediolateral oblique) were shown for each case.

Observers with training and experience hit the mass fastest in both the first and the second view. Unsurprisingly, the T+E+ group were faster than the T+E- group, who in turn were faster than the T-E- group. The simulated scanner was faster at hitting targets the than the T-E- group, and it is suggested that this is because the T-E- group may be distracted by plausible false positive targets.

Those with training and experience took 1.4 seconds to hit the mass on the second view, whereas other groups took much longer than this (2.5 seconds on average). The authors suggest that experienced observers demonstrate a greater understanding of the locational relationships that hold between different views of a three dimensional breast.

The authors conclude that the aim of search is to find something "odd or perturbed" about an image, but not to examine every intuitive target, or to exhaustively examine the entire image. They further suggest that thousands of trials of both normal and abnormal images are required to enhance mammographic performance. Experienced subjects were between 13 and 200 times more experienced than subjects who had training only.

Krupinski performed a similar eye-movement study for mammography using three subjects who had experience of reading mammograms regularly, and three who were radiology residents with some experience of mammography [82, 82].

Experienced readers tended to fixate the lesion almost immediately the image was displayed — on average in 0.6 seconds for masses and 0.9 seconds for microcalcifications. If more than one lesion was present, then experienced readers would proceed almost directly to the second lesion. Little systematic scanning was observed. In contrast, inexperienced readers fixated the lesion in 1.8 seconds for calcifications and 1.5 sec for masses, and tended not to proceed directly to the second lesions. Inexperienced readers demonstrated longer overall viewing times, and used systematic viewing patterns (for example, scanning the entire breast area and examining the skin line and nipple region). However differences in time to hit were not significant.

Krupinski states that scanning patterns for the lesion-free films is similar for both experienced and inexperienced readers. She suggests that if nothing is located early on in search, then readers will search the image systematically. Experienced readers covered more of the breast area in lesion-free, compared with lesion positive, images, but still spent less time examining the image compared to inexperienced readers. It was noted that subjects gave bright patches of tissue prolonged attention. Krupinski speculates that this is because it is more difficult to disambiguate calcification particles where their contrast may be reduced. Search is viewed as adaptive, that is, dependant on the specific difficulties posed by the individual features present in the image.

2.4.2.4 Time course of search

Christenson examined how TP and FP decisions are distributed over time for a chest X-Ray search task [24]. Time course data for decisions made by expert and novice radiologists was recorded for a set of chest X-Rays containing diverse abnormalities categorised by region (lungs, mediastinum, heart and great vessels, chest wall, pleura, and upper abdomen). When decisions per unit time are plotted against time, then 'fast' and 'slow' components of search are evident, represented by two linear components of the resulting graph with differing gradients. Responses in the fast phase are associated with 'at a glance' detections, and responses in the slow phase with more detailed examination (guided search). Observations made in the lungs, heart and pleura demonstrated this bi-phasic quality, but observations made in other chest areas showed only the slow component of search. The more experienced observers were able to detect a greater proportion of targets in the fast phase of search. Christenson speculates that in the absence of clinical information that search is initially directed to areas that give a high return per unit time.

2.4.3 Termination of search

This section is concerned with the question of how radiologists decide when to terminate their search of an image. Work in this area has largely focussed on the phenomena (known as 'Satisfaction of Search' or SOS) that occurs where there are multiple abnormalities in an image, but where search is terminated prematurely, or 'satisfied', when only one has been discovered [124]. Satisfaction of search has been demonstrated experimentally by Berbaum *et al.* in an exercise where artificial lung nodules were added to films containing subtle native abnormalities [12].

2.4.3.1 The nature of the SOS effect

Berbaum *et al.*'s initial experiment employed two test sets. The first consisted of normal and abnormal chest X-Rays. The second was a duplicated of the first but with lung nodules artificially added. The presence of artificial nodules was shown to interfere significantly with subjects' ability to detect native abnormalities.

The authors offer two possible explanations for the observed effect. In the first they suggest that termination of search might be a strategic decision. For example, Christensen had previously reported that the probability of an observer making a FP decision increases with search time, whereas the probability of making a TP decreases [24]. If a reader's confidence threshold for making positive decisions is lowered as search progresses, then an early termination of search may serve to improve a reader's overall performance. Secondly, SOS may be due to perceptual set effects. 'Perceptual set' refers to the phenomena whereby an observer primed to detect instances belonging to a particular category has degraded performance for instances falling in alternative categories [59]. The discovery of a nodule may prime readers for associated findings, but impair their ability to make disparate findings.

In a follow-up exercise Berbaum *et al.* [11] examined the time course of SOS effects using the same film set and a similar experimental design as in [12]. In order to establish the time course of decisions an interruption technique was used. The authors were able to demonstrate the SOS effect under these new conditions — performance for the detection of native abnormalities was significantly lower in the presence of simulated nodules. The total search time was independent of the number of abnormalities (both native and artificial) present in the images, and so the authors conclude that SOS effects are not due simply to early termination of search. The results suggested that on average nodules were detected earlier in the search than native abnormalities, which in turn were detected before false positive decisions were made.

Time course graphs showing the cumulative frequency of TP and FP decisions against time indicate that termination of search tended to occur when slopes for detections of FP, native abnormalities and nodules were equal. This is interpreted as suggesting that search is broken off before it is more probable that FP responses are made compared with TP responses, thus adding weight to the earlier conjecture that halting of search has a strategic component.

Nodine *et al.* commented on the studies of Berbaum *et al.*, relating experimental data of their own from a nodule search task involving cases containing one, two or three nodules [103]. Eye movements were recorded in this investigation. An analysis of gaze duration indicated that two thirds of missed nodules were fixated, even in cases where additional nodules were missed. Nodine *et al.* concur that the SOS effect is unlikely to be due to an early termination of search. However, an SOS effect was not evident for films containing multiple nodules. Nodine *et al.* contended that the SOS effect may be a probabilistic phenomenon, and compared the actual probability of nodule detection in cases where there was one, two and three nodules present with the probability of detection expected if nodule detections are independent events. The results showed that the actual probability of detecting a single nodule if two or three nodules was present matched the theoretical probability for these events.

Berbaum and Franken Jr suggested that it is unsurprising that a SOS effect is not shown for a multiple nodule detection task if the SOS effect is due to perceptual capture [8]. This is because perceptual capture implies that fewer attentional resources are available for recognising alternative disease processes. If fact, detection of a first nodule might lead radiologists to search for confirmatory evidence of metastatic disease, thus enhancing the detection of multiple nodules. This may explain a further result reported by Nodine *et al.*— the detection two of three nodules occurred more frequently than was expected by their theoretical model.

2.4.3.2 Clinical history and SOS

In a further investigation, Berbaum *et al.* sought to examine the effect of clinical history on SOS [9]. The film set and reporting procedure were the same as those used for determining the time course of SOS and direct comparisons were made with the results of this previous investigation [11].

It was shown by ROC analysis that native lesions were detected with equal accuracy in conditions with and without simulated nodules when a history suggestive of the native abnormality was given. Under these conditions it appears

that the SOS effect is alleviated. In contrast, subjects' performance for native abnormalities was worse in nodule containing films when a history of metastatic disease was supplied. Further analysis revealed this difference was most likely due to an appropriate history having a beneficial effect, rather than the presence of nodules and an inappropriate history being detrimental. Time course data indicated that native abnormalities were detected earlier than nodules in the presence of appropriate history, but after nodules in the presence of a history suggestive of metastatic disease.

The authors suggest that search must consist of two stages. In the first stage, specific 'feature analysers' are employed (i.e. those relating to the signs exhibited by a specific disease process) that have been selected and primed by the history. In the second stage, search is essentially the same as search without history, where many, presumably unbiased, feature analysers are available. They suggest that the effect of clinical history is confined to the initial, directed stage of search.

Samuel *et al.* suggest an alternative explanation for lack of SOS effect in the presence of a native abnormality history and why a history of metastatic disease does not enhance the SOS effect [112]. In the absence of a history, SOS is mediated by the detectability of individual abnormalities. Thus obvious nodules caused a satisfaction of search where the native abnormalities were subtle. When an observer detects an obvious abnormality, he or she "self prompts" for related features in a manner that is similar to the effect of a received clinical history. SOS is alleviated because the subject is primed to examine the film for the native abnormality. Conversely, history relating to metastatic disease does not contribute to the priming over and above the effect due to rapid detection of the nodules themselves.

In support of this hypothesis, Samuel *et al.* were able to demonstrate a reversal of the phenomena reported by Berbaum *et al.*, whereby conspicuous native abnormalities interfered with observers' ability to detect subtle artificial nodules [112].

2.5 The nature of radiologists' errors

2.5.1 A taxonomy of reader errors

The development of systems to assist with screening mammography depends on an understanding of the nature of the errors they are designed to alleviate. A taxonomy of FN decisions, proposed by Kundel *et al.* for radiologists performing a lung nodule search task, has been widely used as a theoretical basis for understanding reader error [87].

Kundel *et al.* characterise the discovery of an abnormality as a three stage process. First it is necessary for a potential lesion to enter a reader's useful field of view in order for it to be detected. Second, that recognition of a potential lesion also requires that it is disambiguated from background tissue. Finally, criteria have to be applied to decide whether the detected lesion is a true abnormality or a normal variation. Accordingly, they propose errors leading to FN decisions can be categorised either:

Scanning errors where the lesion does not enter the useful field of view,

Recognition errors where the lesion enters the useful field of view, but where the dwell time is less than some critical value for detection to occur, or

Decision-making errors where the lesion is not reported though the dwell time exceeds the critical threshold for detection.

Kundel *et al.* report an experiment designed to establish the prevalence of each of these error categories [87]. Eye-movements were recorded for four radiologists examining chest X-Rays for lung nodules. Test films were constructed by adding simulated nodules to two chest X-Rays (one of a man, the other of a women) so that anatomical background could be held constant and only the nodule position varied.

It was found that TP decisions were associated with an initial average dwell time of 560ms and an average cumulative dwell time of 960ms. The authors assume that a nodule could be recognised if the initial dwell time is greater than 48ms, or if the cumulative dwell time is greater than two standard deviations below the mean dwell time values for TP decisions, i.e. 800ms. When lesion enters an observers useful field of view but is not reported then an error of detection is signalled if cumulative dwell time is less than 800ms, otherwise an error of interpretation is said to have occured. During this exercise, 30% of the FN decisions were attributed to search errors, 25% to detection errors and 45% to faulty decision-making.

Krupinski [83, 82] reports on a similar study performed for mammographic images, comparing the errors made by experienced and inexperienced observers. She reports that the median cumulative gaze duration for both groups of observers is greatest for TP and FP decisions, and shortest for TN decisions, with FN decisions falling somewhere in between.

Experienced observers made 8 FN decisions, 24% percent were search errors, 24% recognition errors and 52% decision making errors. Inexperienced observers

made 21 FN decisions, of which 29% percent were search errors 42% recognition errors, and 29% percent decision-making errors.

Following Kundel, Gale also draws a distinction between categories of FN decisions [55]. Instead of using eye-movement studies, the empirical basis for Gale's classification depends upon whether an observer reports the abnormality, but fails to recall it (a classification error), or whether they fail to report an abnormality (either a detection or a search error). This is done in the context of the PERFORMS test, an assessment scheme completed periodically by the majority of the radiologists involved in the UKBSP.

2.5.2 A critique of the FN error taxonomy

The classification of radiologists' errors in this way requires a number of important clarifications. Some types of observer behaviour reported in the literature cannot be easily accounted for by the proposed taxonomy, for example, the ability of experienced readers to perform detection and classification 'at a glance'. Studies have shown that experienced readers can interpret tachistoscopically presented chest images [85] and mammograms [132] with better than chance performance. Furthermore, comparisons of novices' and experts' eye-movements reveal that novices are more likely to be distracted by intuitive targets. Features of an image are not randomly selected for further consideration, with increasing experience attention is directed to the most likely candidates. This implies that the mechanism by which search is directed by periphery vision involves an implicit classification decision. When 'at a glance' classification occurs, this does not mean that radiologists have the opportunity to apply the full range of strategies available to them for assessing the significance. Rather, the recognition process itself provides enough information for a classification decision to be accurate in some circumstances.

Berbaum *et al.* suggest that classification of error types by gaze duration may critically depend on the type of abnormality and the circumstances in which it occurs [10]. Thus different abnormality types (and even different instances of a similar abnormality) might demand different degrees of effort for their detection and classification. Dwell times for recognition may depend on the mechanisms associated with different recognition tasks. They argue that some types of task may rely on the operation of a number of 'feature detectors' that may or may not work in parallel. Using dwell time to indicate the type of error made could be further confounded if the correct feature detector were not initially employed (since dwell time would accrue while the proper selection is made).

The distinction between classification and detection as suggested does not allow the possibility that an observer may spend time pondering whether a region of the image contains something, as opposed to pondering whether what it contains is significant. For example, there may be a difference between deciding whether a micro-calcification cluster actually exists, and a situation where an observer is certain of its existence, but unsure as to its significance [2]. It is entirely plausible that radiologists employ a confidence threshold for reporting features, as well as for deciding if a reported feature is benign or malignant. This might blur the distinction between classification and detection errors in the methodology used by Gale and Kundel. Where detection is effortful, but where the observer has decided has decided not to report a feature, this would be counted as a detection error by Gale. However, the mode of failure has more in common with a classification error, and would be recorded as such in methodology employed by Kundel because dwell time would accrue.

Another possibility is that recognition may occur, but on this first appraisal the reader may decide that the feature does not warrant additional investigation. In this case, the feature would not receive a prolonged gaze, however, it does not seem proper to define the event as a detection error. If the feature is reported, then by Gale's methodology this event would be categorised as a classification error, but Kundel would call it a detection error because little dwell time would have accrued.

Some classification is implicit in detection; the distinction really being drawn is between the use of different classification strategies. To illustrate, a parallel might be drawn with common strategies employed in the design of computer detection systems. Such systems can employ multiple and distinct processing stages with differing specificities to make the image analysis problem more tractable. For example, the mass detection algorithm employed by the PROMAM system uses two such stages. In the first low specificity step a number of candidate 'bright blobs' are identified. The second stage has a greater specificity. A series of parameters are determined for each candidate (such as texture, shape, density etc.) and are used to make a benign malignant judgement based on the known para-

[2]It might be argued that this type of situation is of little significance. If a sign of malignancy is so slight as to make its recognition as a coherent feature difficult, then it would be unlikely that a positive decision would be made. However, Cowley reports occasions where trained radiologists failed to recognise (screen detected) architectural distortions even after their location had been pointed out [29]. In this situation it would appear that detection is at least as effortful as classification. Furthermore, results obtained by Wolfe (discussed earlier) suggest that when pre-attentive mechanisms cannot be relied upon to locate potential targets, then detection can be particularly demanding.

meters of actual malignancies. Both stages involve classification (bright blob/not bright blob, and benign/malignant), and a loose analogy might be drawn with the detection/classification distinction in human observers. However, drawing this distinction for humans presupposes the existence of independent mechanisms within the brain that have a functional correspondence to low and high specificity processing steps. It may be that these 'steps' are are in fact tightly coupled, thus preventing a distinction from being clearly drawn. For example, Vecera and Farah describe feature segmentation in humans as an 'interactive' process, depending both on discrimination of visual primitives and on higher level object knowledge. They argue that partial representations are subject to interpretation, and if a correspondence is found with a high level representation of an object, then this in turn will guide how segmentation proceeds [126].

In summary, it appears that the detection/classification distinction represents the end points of a continuum (represented by cases where Gale's and Kundel's empirical measurements would agree), rather than hard and fast categories for defining error types.

2.5.3 Reasons for errors

It is not enough to know that certain types of error might be committed, it is also important to establish the reason for error. In addition to search, detection and classification errors, two further types of error have been suggested. The first may occur when there is a failure to recognise the significance of a lesion from an initial overview of the image. The second if an observer fails to recognise that a lesion is actually present after prolonged scrutiny of the lesion containing region. Both these error types have a basis in empirical descriptions of radiological expertise. The former can be understood in terms of how search is prioritised according to processing of information from the peripheral visual field. The latter according to Woolfe's description of 'effortful detection', where, under certain circumstances, the location of targets is not available pre-attentively.

The case of classification errors is somewhat intuitive. If the observer is engaging in a comparison of the feature properties with an internal representation of those properties which typically denote malignancy (which could be prototypical or instance based), or is making comparisons between other regions within the image, then this activity may not be carried out effectively or consistently.

The case of detection errors is more problematic. What is it that prevents an observer from recognising a particular feature within an image for what it is? Kundel *et al.* suggest an additional category of error — "orientation errors",

where an observer is familiar with neither the object of search nor the properties of the background in which it is embedded, as would be the case for a novice [87]. However, there is the possibility of another type of 'orientation' error, where the object of search does not form part of the observer's 'anticipatory schema'. Studies cited in this review indicate that perceptions can be altered by history and prior experience. The SOS effect provides a convincing example of where the ecological properties of an image can influence the detection of lesions. Of course, it is possible that such effects might influence both detection and classification. In the former case, they might influence what is perceived, and in the latter, they may alter an observer's confidence that what is perceived is significant.

An inappropriate orientation to an image (this time, in the sense meant by Kundel *et al.*) can also be used to account for search errors. Studies of radiologists' eye movements reveal complex interrelationships between image properties, prior knowledge (such as locations where lesions are more likely to occur) and where a radiologist's gaze is directed during the course of search. This process of prioritising regions of an image for scrutiny may itself be error prone where a lesion appears in an improbable location, or in an improbable relationship with other components of the image. Alternatively, the process might be adversely affected by constraints placed on the observer, such as fatigue.

2.6 Decision aids for mammography

Design rationales for decision-aids in mammography correspond to Kundel *et al.* 's taxonomy of observer errors, in that detection and classification are perceived as distinct points of failure for a human observer and as processes that can be supported independently. For example, Giger draws a distinction between aids for detection — designed to reduce the number of FN decisions made by radiologists, and aids for classification — designed to reduce the number of FP decisions made [58]. Another way of phrasing this distinction is to say that the purpose of detection aids is to reduce the number of interval cancers, and so improve outcomes by enabling an earlier detection, whereas the purpose of classification aids is to reduce 'unnecessary' invasive procedures (such as biopsy). In the former Giger emphasises that the classification of a computer detected lesion is left entirely to the radiologist. In the latter she stresses the importance of ensuring that the system does not bias the reader towards making a normal decision when this is inappropriate — that specificity should not be bought at the expense of sensitivity.

Possible undesirable effects of using detection aids have also been suggested. Nishikawa *et al.* express the concern that detection aids should not unduly increase the overall FP rate [102]. Here the purpose of detection aids is viewed as one of improving sensitivity without having an undue effect on specificity. Chan *et al.* recognise a further possibility, that the absence of a response from a computer detection aid might bias the radiologist into passing a film as normal [23].

Thus the conception of computer aids for mammography involves consideration of a number of possible effects on the human observer, some desirable, and some to be avoided. It also depends on a cooperative synthesis of abilities between human and computer agents. Whilst detection aids can achieve a sensitivity comparable with that of a trained radiologist, they typically have a relatively poor specificity. Benefits may accrue to the radiologist if the system is able to draw cancers to their attention that they may otherwise have overlooked, but only if FP prompts can be effectively dismissed at a relatively low cost [70].

2.6.1 Detection aids

Prompting studies are concerned with how the use of prompting aids affects observer performance and behaviour. A rough distinction can be drawn between studies that primarily seek to understand how observer performance is affected, and those that are concerned with determining the effectiveness of different prompting regimes.

2.6.1.1 Effects of prompting

Nishikawa *et al.* conducted an experiment to determine if radiologists can easily distinguish computer generated FP and TP prompts [102]. Two subjects were asked to circumscribe three regions that might contain a microcalcification cluster on each of a test set of fifty mammograms. For each annotation, they rated their confidence on a 0-100 point that a cluster is actually present. In this way the subjects were forced to generate a large number of FPs, and ROC analysis revealed how well they could distinguish these from their own TP decisions.

In the second part of this exercise, the subjects similarly rated both TP and FP regions identified by a microcalcification detection algorithm which had analysed the the same set of cases. Performance by ROC analysis was similar to that for the first exercise, and it was concluded that the subjects were able to discriminate between system TP and FP decisions as well as they were able to distinguish between their own.

Chan *et al.* report on an experiment to test if observers' detection performance could, in principle, be improved by use of a microcalcification detection system [23]. Two prompted conditions were used. The first consisted of the prompts generated by the detection system when operating at a sensitivity of 87%, with four FPs generated per image. The second consisted of a simulated level of performance. The TP prompts were generated as for the first condition, but the FPs were produced by running the algorithm with a stricter confidence threshold, resulting in a FP rate of one prompt in every two images. The time available for examining each case was limited to 5 seconds.

A statistically significant increase in performance was found for both prompted conditions when compared to reading unaided. Performance using the enhanced prompting condition was greater than that for the native condition, but not significantly so. In the prompted conditions, subjects' confidence that microcalcification were present typically increased for abnormal cases, but remained the same for normal cases. This suggests that subject were able to recognise FP prompts. In 9 out of the 10 abnormal cases where there was a decrease in confidence ratings compared with the unprompted condition, the system had also made a FN decision. This suggests that subjects were falsely reassured by the absence of a prompt and terminated their search early. The authors suggest that training prior to the introduction of a prompting system should make users aware of the system's deficiencies.

Mugglestone *et al.* examined the effect of prompting on the eye-movement patterns of observers [99]. They used 46 films, half normal, and half containing interval cancers. The cases were presented in prompted and unprompted conditions. The prompts were simulated, and the FP prompts were chosen by a radiologist as areas that could be construed as being suspicious.

The results indicate that prompting increased the number of FP responses on both normal and abnormal cases. However, an examination of subjects' TP decisions in both the prompted and unprompted conditions showed that there was no overall change in confidence ratings, indicating that the presence of the prompts was not making readers overall more suspicious of prompted cases.

FP prompts generated by another radiologist might have a greater likelihood of influencing FP decisions than prompts produced by a detection aid. For example, Nishikawa *et al.* showed that human observers had little difficulty distinguishing computer generated TP and FP prompts, but also that there was little correlation between human and computer FPs for the system they were using [102].

One of the unprompted lesions was detected by five of the six subjects in the

unprompted condition, but only by one subject in the prompted condition. In the prompted condition there were correspondingly more FP decisions made for a falsely prompted region. The authors conclude that under certain circumstances it may be possible for FP prompts to distract attention from an actual lesion.

An analysis of eye-movement data suggested that subjects' search patterns were disturbed by the presence of the prompts. For example, subjects made fewer comparisons between left and right mammograms in the prompted condition. Furthermore, subjects spent a greater amount of time searching the prompted area at the expense of considering other regions of the film.

2.6.2 Effectiveness of prompting regimes

A question of particular importance concerns how good prompting systems have to be to effect an improvement in the performance of radiologists using them. To address this issue, Hutt performed an experiment to determine the effect on performance of prompting regimes with differing FP rates [70]. In each of three prompted conditions the action of a prompting system was simulated to give a sensitivity of 90%, and FPs were randomly placed to give differing FP rates.

The results showed a statistically significant increase in performance as measured by ROC analysis for all conditions except that with the highest FP rate. For abnormal cases no significant difference in work-up decisions was found. However, there were significantly fewer work-up decisions made in the two prompted conditions where ROC analysis showed an increase in performance. This suggests that the improvement in performance was due to subjects making better classification decisions for normal cases, rather than to improved detection performance (i.e. by picking up additional cancers). It is possible that, because FP prompts were randomly generated, it was easy for subjects to distinguish between TP and FP prompts. The simulated prompts might implicitly be giving information that enabled subjects to make better classification decisions.

Because FP prompts were randomly placed (and so are perhaps less likely to correspond with regions that may be misinterpreted as significant) this may negate the effect observed by Mugglestone *et al.* whereby some prompted locations chosen by a radiologist increased subject's tendency to recall [99].

Hutt argues that the effectiveness of a prompting system might in part be related to the TP:FP ratio, that is, the likelihood of a given prompt being correct [3]. He suggests that the results indicate that a prompting system is most likely to

[3] Although Hutt states that the likelihood of a given prompt being correct is the TP:FP ratio, it is actually: $\frac{TP}{TP+FP}$

be useful if the number of falsely prompted cases does not exceed by fifty percent the number of TP decisions. He further contends that an acceptable prompt rate in a clinical environment (where there are fewer abnormalities) might need to be lower than this — that the number of FP prompts should not exceed the number of TP prompts. This is a particularly pessimistic result, as it suggests that the system would have to perform better than a radiologist to be of use.

In an experiment using simulated mammography images and untrained observers, Astley *et al.* sought to test a number of hypotheses in a simulated mammography test [4], again to address issues of appropriate levels of system performance. Computer generated images and targets mimicked a microcalcification detection task in a background with a visual appearance similar to that of breast tissue. Novice observers were used as subjects, and prompt delivery was by means of a pre-cue shown immediately prior to display of the image.

Two experiments examined the effect of observer performance on test sets where the number of targets was systematically adjusted. The aim was to show that results from prompting experiments containing sets biased with target cases can be generalised to screening conditions where the target prevalence is much lower. The first experiment examined whether performance is affected by the prevalence of a target containing images without prompting. No significant difference between conditions was found. The second experiment duplicated the conditions in the first with the simulated action of a prompting system. The sensitivity of the system was maintained across conditions, and for each condition an equal number of FP prompts as TP prompts was produced, thus the specificity improved with decreasing target prevalence. Again, no significant difference was observed between conditions. A comparison between the conditions in the first and second experiment revealed that observer performance improved in the prompted condition. This appears to demonstrate that studies involving biased test sets can be generalised to situations where there is a low target prevalence. However, the condition with the least number of targets had a target prevalence of 10%, which is still far greater than the prevalence of breast cancer in the screening population (0.5%).

Two further experiments sought to test whether the degrading effect of FP prompts is related to the ratio of true to false positive prompts (or the likelihood of a given prompt being correct), or whether it is due to the overall prompt rate. In the first experiment a number of test conditions were generated where the prevalence of target images was kept the same, whilst the sensitivity of the prompting was varied (prompt rates of 70, 80 and 90% were employed). The same

number of TP prompts as FP prompts was produced for each of the conditions, and so the specificity of the system decreased with increasing sensitivity. A trend towards increased observer performance was observed with increasing prompt sensitivity, but this was not found to be statistically significant. In the second experiment, conditions with decreasing numbers of target images were used. The same prompting sensitivity was maintained (thus the number of prompted targets fell in line with the number of available targets), and the number of FP prompts was kept constant across conditions. A trend was observed towards a reduction in performance as the proportion of images prompted increased. Again, this trend was not statistically significant.

These results were counter to expectations. It was thought that performance would be greatest where the ratio of TP to FP prompts was greatest, in effect where the likelihood of a given prompt being a TP is greatest. Instead, it appears that observer performance is determined by the total number of prompts in a given condition. However, subjects' use of prompts may not remain constant across conditions. Astley *et al.* suggest that one factor that determines whether subjects respond correctly to a TP prompt is the their confidence in the system, which is turn related to the observed TP to FP ratio. That is, a subject's confidence in the system is dependent on the likelihood of a given prompt being correct. It is possible that subjects' confidence in the system will influence their search strategy. Where confidence is high, subjects may spend a greater proportion of their limited time examining prompted regions. Where confidence is low, they may adopt a more balanced search strategy, and may be less likely to be influenced by FP prompts.

2.6.2.1 Classification aids

Getty *et al.* tested a classification system's ability to improve the performance of general radiologists, and to see if their performance could match that of mammography specialists when aided [57]. A test set of 118 images (58 malignant, 60 benign) was read by 6 general radiologists without decision aids, then after a short interval and after training, with use of the classification system. In both conditions the location of each abnormality was indicated to the subjects. A statistically significant improvement in performance was observed.

This study raises interesting methodological questions. The system relied on the subjects themselves extracting and parameterising information about the lesion according to a checklist. The classification system then returned a likelihood that the lesion was indeed malignant. The authors suggest that classification (as

performed by human observers) involves two steps — the extraction of appropriate features from the image, and the merging of assessments about those features to reach a decision. It is possible that use of the checklist, and the discipline imposed by demanding the observers supply parameters themselves, has a beneficial effect on performance. A second problem with this particular study is that the training session held immediately prior to the assisted condition involved giving subjects detailed feedback about their performance (compared with that of the system and of specialist mammographers) in classifying malignancies in a separate set of 44 films. This in itself may have improved subjects' ability to make diagnostic decisions.

An interesting question is whether delivery of classification information is best achieved by reduction of assessed features into a probability of malignancy, or by indicating in a rule based or qualitative way the nature of the decision. Both approaches have been shown to be successful [120] [71].

2.7 Discussion

2.7.1 Prompting studies

There are particular methodological problems associated with determining the performance effects of detection aids for screening mammography. For example, the actual prevalence of breast cancer in the UK screening population (approximately 0.5% [46]) makes it difficult to perform experiments that can yield statistically significant results without using test sets heavily biased in favour of malignancy. Detection aids are a relatively new technology, and few systems currently are capable of a level of performance that might be acceptable in clinical practice. In prompting experiments, it is sometimes necessary to simulate system responses where the level of performance required for the condition is not currently achievable with existing algorithms. Alternatively, the task of reading can be made more difficult by applying time constraints. Finally, experienced film readers are a limited resource, and it may be difficult to recruit sufficient subjects to perform the large number of trials usually required for visual search tasks. An analogous task using untrained subjects may be employed to overcome this difficulty. For these reasons, care should be taken when considering the implications of prompting studies for actual clinical practice.

One way of conceptualising the use of detection aids is to view prompting information in the same light as other sources of prior information, with a similar capacity to affect search and to influence decisions. As has been suggested for

more conventional forms of prior information (such as clinical history), reading practices may have a bearing on how decision-making is affected. Similarly, the method of prompt delivery used for investigating prompting, and in actual clinical practice, may be important. For example, presenting prompting information actually on the image [99], or as pre-cues [4], may subtlety affect reader's appraisal of the visual terrain, by both capturing attention and priming for certain types of finding. Another commonality between prompting studies, and studies of the effects of prior information, is that the results obtained appear to be determined (in part) by the nature and setting of the experimental task. However, neither the literature on prompting, nor that on radiological expertise, addresses the nature or setting of actual clinical practice. The *task* of reading is often examined in a way that is insulated from the social context in which reading *work* is usually performed. In an experimental setting, decisions are not consequential, collaborative practices are not accounted for, and consequences of shared values within a community of practice ignored.

The point of departure for the work comprising this thesis was an ethnographic investigation of work practices in UK breast screening clinics. The aim of which was to provide a contextual grounding for interpreting the prompting studies subsequently conducted as part of the evaluation programme for the PROMAM system.

2.7.2 Rationales for decision aids in mammography

A distinction is often drawn between perceptual and cognitive skill in image interpretation. It is argued that detection of features within an image is a distinct and prior step to reasoning about whether those features are significant. Detection is regarded as a watershed event that allows the possibility of a report being made. Furthermore, this event also enables a qualitative change in strategy from a dependence on automatic processing to a consciously guided appraisal of the properties of a detected feature.

Detection aids are designed to play a role in overcoming search errors (where the radiologist has failed to fixate a lesion) and detection errors (where a lesion is fixated, but not recognised). If the latter are viewed as a failure of an observer's anticipatory schema, then, by giving the location and the type of lesion prompted, an observer's perceptual orientation may be altered sufficiently for him or her to perceive a lesion that might otherwise have gone unrecognised. However, this rationale for detection aids explicitly excludes any assistance for readers making classification decisions — they are not designed to assist in rectifying classification

errors (in as much as the classification error is not itself attributable to a failure to apply an appropriate anticipatory schema). In fact, there is concern that prompting aids could degrade performance if readers' classification decisions were influenced. For example, if prompts increase readers' suspicion for benign lesions, or if they decrease suspicion for unprompted malignant lesions. On the other hand, the rationale for classification aids does allow for some influence in the radiologist's own decision-making. It is plausible for classification aids to assist with reducing both FN decisions due to errors of interpretation as well as reducing FP errors, although an emphasis is usually given to the latter role.

The position that detection and classification can be supported independently requires a tacit assumption that these processes are functionally distinct for a human observer. However, in section 2.5.2 it was argued that this assumption may be unwarranted. If this is so, observer errors cannot be characterised unambiguously as failures of either detection or of interpretation, casting doubt on prompting systems' role as merely detection aids. Indeed, prompting studies reviewed in this chapter reveal a number of departures from the model usage for detection aids. While FP prompts do not appear to uniformly increase suspicion [102], they can result in a recall decision in some specific cases [99]. The absence of a prompt may falsely reassure [23], and the co-presence of FN and FP prompts may distract attention from the abnormality [99].

One motivation for the experimental work described in this thesis was to examine how prompting information is actually used in practice, and in doing so, contribute to a clearer understanding of the conceptual basis for decision aids in mammography.

2.7.3 Visual search

Investigations into the visual search of medical images contribute in a piecemeal fashion to an understanding of what is a highly complex activity. Patterns of eye-movements appear to be less amenable to characterisation with increasing experience of the observer and decreasing specificity of the search task. Expert search appears to be defined by its efficiency, rather than by completeness or systematicity. Search is driven by a complex series of interactions between different types of prior knowledge, momentary hypotheses and the properties of the image itself. Experience itself may be thought of as prior information in the form of heuristics that encode likelihoods about the possible location of abnormalities, and the likely significance of possible image features. Furthermore, with increasing experience there appears to be an increasing degree of automaticity;

experts can hit targets faster and make greater use of information available in their peripheral visual field.

Another way of interpreting the complexity of visual search is to view it as being an 'adaptive' response to image variations. Although images produced using a similar modality and technique will demonstrate a broadly consistent visual terrain (they can be can be readily identified as being a chest X-Ray or a mammogram etc.), the details of the features that define the terrain can vary considerably. Specific features may be more or less difficult to interpret and may require differing analytical approaches. Indeed, a qualitative distinction can be drawn between glances that gather information for different purposes (for example, for comparison, analysis and orientation) and the situations in which they are applied. Experience enables observers to bring to bear and organise efficiently strategies that are commensurate with the nature and difficulty of the tasks at hand. The frequency and location of different types of purposeful glances depends on the demands made by individual image features and their relationship to the image as a whole. The ordering of depends on the priority assigned by pre-attentive processing, the results of prior focussed examinations and on knowledge of likely target locations.

In a comparison of visual search between random 2D scenes and naturalistic 3D scenes by recording subjects' eye-movements, examination of the 2D scenes was found to be exhaustive and systematic. In contrast, inspection of the 3D scenes was less exhaustive, but more efficient, suggesting that spatial cues were used to organise the search [116]. Thus visual scenes often not only contain information that describes targets (sought after components that have some relevance to the observer), but also information about how to explore the scene efficiently. Where use can be made of this orientating information then, apparently, this is done quite naturally, that is, without a conscious decision to do so.

In summary, for expert observers, inspection of medical images can be described as effortlessly adaptive — regions of an image are scrutinised to differing degrees in a way that is dependent on the properties of the image, the experience of the observer, and the orientating information available from the image itself. In this thesis it is argued that this adaptive approach implies that readers are accountable to the task of reading in certain specific ways, and that this in turn has implications for the way information supplied by detection aids is interpreted and utilised in practice.

Chapter 3

Work practices in breast screening

3.1 Introduction

The basic principle behind mammography as a screening test is that the appearance of breast tissue on a mammogram reveals in an objective way the underlying causal processes. More specifically, signs of malignant processes should be sufficiently distinct so that a reasonable sensitivity and specificity can be achieved by visual inspection. Psychological accounts of visual skill and radiological expertise are reviewed in Chapter 2. However, a psychological approach necessarily attempts to identify the mechanism by which a single human observer locates and interpret features within an image. A more complete description of expertise should also include an account of how the mechanisms underpinning skillful behaviour are socially situated [43].

This chapter describes the results of an ethnographic style investigation of work practices in two Scottish and four English screening centres (referred to with the letters A through to F to preserve anonymity). The study was not intended to precisely catalogue the significance of each component of the evidence used by readers in decision-making (knowledge acquisition). Rather, the aim of examining how the task of reading films is organised was to explore the context in which radiological expertise is brought to bear.

3.2 Methods

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. Significant qualifying criteria for selection for clinical trials include having a predominantly double reading practice, and sufficient capacity to screen the numbers required over the period of the trial.

Ethnography involves a systematic and structured approach both to the gathering and interpretation of qualitative data that is often grounded by the use of various types of triangulation. A researcher will attempt to identify discrepancies between what people say they do and how they actually behave, thus uncovering tacit assumptions held, for example, about the nature of a task or relationship. Comparisons can be made between the activities of an individual over time, and in different settings, and between different individuals in similar settings, both to challenge the researcher's own working assumptions, and to guide further observations. Ethnography typically depends on a protracted period of involvement with the subject matter (often in the order of months or years) so that detailed and representative data my be collected. In the context of supporting design, some have found shorter field studies — so called 'quick and dirty' ethnography — to be useful [67]. Others argue that to contribute more to design than could be achieved by traditional approaches to requirements capture, the full force of ethnographic methodology should be employed, including the lengthy involvement of trained practitioners [2, 49].

One weakness of the investigation described in this chapter is the relatively short duration of the fieldwork. This has implications for both the representativeness and detail of the data collected. For example, only twenty reading sessions were observed in the six clinics studied. It would have been desirable to have made a greater number of observations that included all the possible combinations of practice (for example, observing each reader reading first and reading second). However, one strength of this investigation lies in the 'natural' triangulation made available by contrasting similar practices carried out in different centres.

Both screening and assessment clinics are usually held within breast screening centres. Additionally, in the English centres studied (clinics B to E), symptomatic work is also undertaken. When a diagnosis of cancer is made at an assessment clinic, the woman is referred to other hospital services for treatment, although typically there is continued involvement of screening radiologists and clinicians. The investigation focusses on clinic activities related to screening work, in particu-

lar, on the preparation and interpretation of evidence used for making a screening decision.

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 film 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 film 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 film reader in centre C.

3.3 Variations in clinic practice

Although all clinics within UKBSP 'do screening', there is often considerable variation in how this is practically achieved. Some of the observed variations in practice that can be consequential in decision-making are summarised in Table 3.1 and Figure 3.3. This section discusses how differences in practice are related to both the broader pattern of practice innovation within the breast screening service and to the constraints imposed by local circumstances. In the remainder of the chapter it is argued that these differences also illustrate how working arrangements of artefacts and procedures are used to support decision-making.

Screening practice in the UK is based on the recommendations of the Forrest Report [46], in particular, that women between the ages of 50 and 64 should be screened every three years using single view mammography. However, screening practice has not remained static since the programme's inception in 1988. The autonomy given to individual screening centres has enabled innovation to proceed by the independent adoption of new practice. Innovative practice 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 false negative and false positive errors [127]. This publication precipitated the adoption of two view mammography for incident round screens as standard practice by the UKBSP.

Clinic	HRT	Blinding	Recall decision	Double reading	Previous films interval	Radiographer opinion
A	Yes	No	3rd reader Arbitration	All cases	3	No
B	No	No	Worst opinion	Most cases[2]	3	No
C	No	No	Discussion of disagreements	Most cases[2]	3	No
D	Yes	No	Worst opinion	All cases	6	No
E	No	Yes	Worst opinion	50% of cases	6	Yes
F	No[1]	No	Worst opinion	All cases	3	No

Table 3.1: Variations in practice between the clinics studied. Differences in the organisation of films on viewers are shown in Figure 3.3. [1] This clinic does not ask about HRT status, but the radiographer will make a note if this information is volunteered. [2] Some cases may be single read due to leave or sickness.

This is a noteworthy example because it also demonstrates how the distinction between screening and assessment tests can change. Typically, screening involves a single test that has a high sensitivity and relatively low specificity. In contrast, assessment of screen positive cases involves the progressive application of a number of different tests that have an increasing specificity (for example, additional mammography, ultrasound, fine needle aspiration, core biopsy, surgical biopsy). Craniocaudal view mammography is often the first test performed as part of an assessment clinic. Extending the incident round screening test to include craniocaudal view mammography represents a tradeoff between increasing its complexity and improving its specificity.

At the time of this investigation, the clinics studied were involved in more or less formal studies of practice innovation. 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 has 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 [33], and by a number of English centres [129]. This is partly because there are methodological problems in ascertaining performance gains due to double reading, resulting in widely varying estimates for its effectiveness [130], and also because of local shortages of trained film readers [129].

In addition to innovation, differences in practice may be a response to variations in local circumstances. For example, clinic A changed from a system of 'worst opinion recalls' [1] to a system of third reader arbitration because their recall rate became unmanageably high. In contrast, centre D discontinued a policy of discussion of disagreements because individual readers held out for their own recall decisions. In centre E, the practice of double reading only 50% of cases is in part a practical consideration. In common with centre C, centre E operates a satellite screening centre where, in contrast with centre C, both mammography and reading are performed. Due to the location of this satellite centre, it is convenient for one film reader to attend to the workload of this clinic alone. Furthermore, two of the film readers in centre E 'got used' to single reading films during a period when they were the only two readers available. Centre E was the only centre to solicit the diagnostic opinion of radiographers, and this is seen as supportive of their single reading practice. Centre E was also the only centre

[1] Where a decision to recall a case for assessment by either reader in a double reading pair guarantees an invitation to attend an assessment clinic.

to operate a system of blinded double reading (where the opinion of the first reader is not easily available to the second reader). Although blinding is often acknowledged as being desirable, it is difficult to implement easily and safely as it requires a duplication of paperwork. In centre E, an electronic system of recording screening decisions is used, with the advantage of making blinding easier to implement.

In contrast to the variations in practice observed between centres, practice within individual centres is surprisingly homogeneous — for example, how mammagrams are arranged on the viewer (Figure 3.3). This can be explained in part by how reading is organised. Many of the activities involving film readers require coordination with other participants (for example, 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 done 'in a quiet moment' or at the beginning or end of the working day. The availability of a given film reader at a given time often cannot be guaranteed, nor can the time they have available for reading be predicted. Thus artefacts 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. This also mitigates against collaboration in screening decisions as this would increase the level of commitment inherent in the reading task. Although reading is a relatively low commitment activity, it 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 double reading would result in a delay of more than a week.

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

3.4 Preparation of evidence

The record of the prevalent and incident round screens, consisting of the mammograms and associated paperwork, undergo a series of preparation steps before being made available for reading. Additions are made to the record's content,

and the components of the record are organised to facilitate both handling in later preparation steps, and ultimately their examination by a film reader. Taken together, these artefacts contain the evidence available to a film reader for making a decision. This section discusses in turn, the production of mammograms, the subsequent preparation of mammograms and other screening artefacts, and finally, the arrangement of these artefacts for reading.

3.4.1 Mammography

It is possible to image the breast using combinations of a variety of angles and positions to give different views. The mediolateral oblique view, taken by compressing the breast diagonally in a line from the shoulder to the stomach, visualises most breast tissue and for this reason is the primary view used in screening mammography. A cancer may be mimicked, disguised or absent in a single view, thus a second view taken from a different angle may help to resolve any ambiguity. The recommended second view is the Craniocaudal (CC) view, taken by compressing the breast horizontally. If a woman has large breasts or is difficult to position, then additional views may be taken at the radiographer's discretion to ensure full coverage.

Mammographic imaging is performed by a trained radiographer. Good technique is important to obtain maximum coverage of the breast and to ensure the image is free from obscuring or confusing artefacts. The radiographer pulls the breast away from the chest wall and applies firm compression, avoiding skin folds and overlapping extra-mammary structures. The breast is spread as compression is applied to separate glandular tissue, thereby reducing composite shadowing. A large component of a radiographer's skill lies in adapting the basic techniques to cater for physical variations in the women screened.

Film readers may recall a case for a 'technical repeat' if they feel that imperfections interfere significantly with their ability to interpret a mammogram. Considerable emphasis is placed on the quality of the mammography. In all of the centres studied radiographers were involved in weekly self and peer assessment of their work to ensure that national guidelines are met — technical repeats should comprise less than 2% of screening cases. When screening in static units, except in centre F, usually a radiographer will ask the woman to wait until the films are developed in case a repeat is necessary. In centre F cases are pre-read for technical quality by a radiographer and mammograms that may warrant a technical repeat are displayed on a dedicated viewer for final judgement by a film reader. However, even in centre F, if it is felt that there may have been a problem at the

time the mammogram was taken the woman may be asked to wait. Similarly, if it is known that the woman has made a long journey (for example, if she has missed screening on a mobile unit in her locality, and in consequence has travelled to the screening centre itself) the mammograms are developed and technical quality ascertained before she is discharged.

There can be a discrepancy between national guidelines that describe the characteristics of technically acceptable mammograms, and what a film reader will find acceptable in practice:

> A radiographer is talking to a trainee about a case where skin folds are visible on the mammogram. She states that the importance of skin folds depends on the radiologist: "our radiologists do not mind skin folds as long as they don't obscure...".
>
> *Observation in processing room (field notes C)*

Although technical repeats in centre F will always involve an additional appointment and therefore inconvenience for the women involved, the overall number of repeated cases might be reduced because here technical quality is assessed by film readers, rather than by radiographers. The former are concerned with 'diagnostic' quality, and the latter often with more abstract definitions of 'technical' quality embodied by collegiate guidelines.

Where repeats are taken at the time of the initial screening there are often limits to the number of retakes performed, and after the first repeat the woman is often told that she can go. Again there can be exceptions to this rule, for example, if the films have been taken by a trainee radiographer further repeats may be required.

Radiographers strike a delicate balance between comfort, convenience, radiation dose, and achieving quality mammograms:

> A radiographer tells the trainee that she doesn't have to put so much compression on as she did for the last case. She states that it is important that the ladies do not find the process too painful as they want them to come back. She adds that it is different with symptomatic ladies as they have something wrong with them and so greater compression can be given.
>
> *Observation in processing room (field notes B)*

Technical quality is not solely dependent on the skill of the radiographer, but also on the physiology of the women screened. For this reason it is sometimes only possible to produce an imperfect mammogram. In these cases the radiographer will make a note for the film reader explaining the difficulties so that a

technical repeat is not inappropriately requested. The woman is also advised that the radiographer has not been able to image all of her breast, and consequently screening may be less thorough in her case.

Screening in outlying areas is often performed on 'mobile units'. These are usually vans equipped with changing facilities and X-Ray sets, but not with processing facilities — films are transferred to the screening centre to be developed. Radiographers therefore do not have an opportunity to examine the mammograms as they take them. In the static units continual monitoring of performance is available as a matter of routine, whereas the lack of immediacy in mobile units can be disconcerting:

> A radiographer stated that when first working 'blind' (i.e. not being able to see immediately the films that had just been taken) on the mobile units it felt very uncomfortable, and that she was very keen to get back to the screening centre to see what the films were like.
>
> *Observation in processing room, (field notes C)*

Radiographers claimed that screening on mobile units typically generates greater numbers of multiple films because radiographers may be uncertain that they have achieved sufficient coverage.

3.4.2 Preparation of artefacts

Paperwork and films for each case are kept together in a 'film bag'. A screening form is produced for each woman for each screening round attended. In the current round, this form is usually attached to the outside of the film bag until after data entry for that case has been completed. The screening form not only helps to identify the case, but also acts as a temporally organised record of each screening episode. The layout and information collected on a screening form varies between centres, but commonalities include: screening ID, name, address, date and location of screening, number of films taken and in what view, comments made by the radiographer, and the film reader's decision. Some of this information is redundant, for example, the number of films taken and the view can serve as a failsafe check to ensure that all the relevant mammograms are available to be read. The form provides a link between the temporally and spatially separated screening activities. For example, radiographers use the form to convey information that may be pertinent to a film reader's interpretation, such as difficulties experienced performing imaging.

Radiographers will also take a brief history of each woman screened and record this information on the screening form. This might include: whether the

woman has had previous breast surgery, whether there is any family incidence of breast cancer, and whether she herself has noticed any lumps or changes. The radiographer will also note the location of any scars or blemishes that might masquerade as a lesion on the mammogram.

Before they are developed, mammograms are marked to identify them as belonging to a particular woman. The marking usually consists of the woman's name and her screening number (or 'screening ID'). This is achieved in most of the centres studied by use of a 'light-stamp' — a device that transfers information on a typed or handwritten label to the film photographically. In centre E the name is typed onto a lead strip — this is placed onto the X-Ray set and exposed as the mammogram is taken. Both methods create a permanent and integral means of identification. It is usual also to place lead markers on the X-Ray set that indicate the view (e.g. 'CC' or 'Obl') and the breast ('L' or 'R') so that this information too is transfered to the films as they are taken. Again, centre E is exceptional in that markers identifying a view are seldom used (this is discussed further below). After the films have been exposed and then processed, adhesive labels are used to indicate the screening round. Name, side and view markers are checked and corrected if necessary with hand written labels. It is at this point that films are checked for technical quality, and where repeats are necessary the substandard films are labelled as being technically inferior — in centre C this is done by cutting off one corner of the film.

As with the screening forms, some of this information is redundant as it is often available from other sources. For example, the view can usually be identified by examining the films themselves, and the screening round by examining the screening form. In fact, the view is mostly not identified on the mammogram in centre E because this is believed to be overly redundant. Here only 'fronts' [2] are labelled as these can sometimes be confused for CCs. This 'redundancy' does, however, serve to make some types of information available 'at a glance', for example, technically poor films where a corner has been cut off are immediately obvious if placed on a viewer. This facilitates the selection and arrangement of the appropriate films for reading and also enables a film reader to quickly ascertain that they are using appropriate evidence to make a decision, for example, that previous films and current round obliques have not been transposed, that 'fronts' are not interpreted as obliques, etc.

[2] 'Fronts' are additional views taken of the front of the breast where it is thought that the whole of the breast may not fit on a single film.

3.4.3 Selection and arrangement of artefacts

Films are loaded onto automated viewing boxes for reading and are typically grouped together in 'batches', corresponding to all the women who attended a particular contiguous screening session or 'clinic'. There is often paperwork associated with each of these batches describing, for example, how many case are in each batch, who loaded the films onto the viewer, the date the viewer was loaded, the dates when the films were read, and who performed the reading. For each batch the record bags are piled in the order that the films are displayed on the viewer.

Two types of automated viewer, Planilux and RadX, were observed in use in the clinics studied. The RadX viewer (Figure 3.1) consists of two independent horizontally moving bands on which the films are mounted. Hand or pedal controls can be used to position the films on the viewer — films not in view are wound onto drums in the body of the machine. The Planilux viewer (Figure 3.2) consists of a series of frames for displaying films. Hand controls are used to select a frame from a cassette located in the base of the viewer. Table 3.2 shows the number and types of viewer used in each of the centres studied.

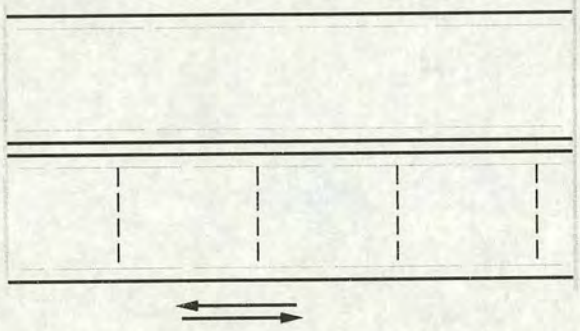


Figure 3.1: A schematic of a RadX viewer.

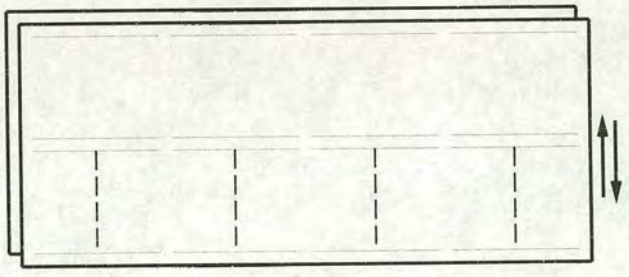


Figure 3.2: A schematic of a Planilux viewer.

All centres except centre C employ an X-Ray helper, whose role appears to be one of 'oiling the wheels' of screening. X-Ray helpers will flexibly assist radio-

Viewer type	Clinic					
	A	B	C	D	E	F
RadX	3	3	2	3	2	1
Planilux	0	0	1	0	1	3

Table 3.2: Types of film viewers available at the clinics studied

graphers by processing films, and they are responsible for placing films on viewers for reading and for taking them down again afterwards (with the exception that film readers assume responsibility for removing any cases recalled for assessment). Additionally, X-Ray helpers play a role in ensuring the integrity of the decision making process by checking both that only cases referred for routine recall remain on the viewer, and that there is a valid decision associated with each case. In centre C all these tasks are carried out by radiographers.

Although readers can access any of the artefacts in the screening record, a selection is made from the available artefacts which are then organised on and around the viewer in a specific fashion. In this way, the evidence considered to be the most relevant to interpretation is made the most easily accessible. X-ray helpers are assisted in this task by the labelling in the preparation stages described above.

Technically poor films are not usually used as evidence for screening decisions, and except in clinic B, are not placed on the viewer for reading. Not displaying technically poor films reduces both the overhead on reading and the possibility that inappropriate evidence is used for making a decision, whereas displaying them may make additional evidence available to the film reader. Technically poor films are seldom discarded, partly because if a women does not re-attend for a technical repeat then the imperfect films may be the only ones available for making a decision.

In centre B, the X-Ray helper checks each case to see if the woman has attended an assessment clinic previously, and will orientate the assessment form so that it protrudes from the film bag, thus drawing the attention of the film reader.

Figure 3.3 shows the arrangement of films employed by each of the centres studied for typical incident and prevalent round screens (when multiple films are taken, then these layout strategies may be modified). Choice of film layout appears to be influenced by a number of factors. It is notable that in clinics A and D, where films are arranged linearly, only RadX style viewers are in use. Use of Planilux viewers often requires a film reader to stretch or change position to examine films located of different parts of the viewing frame, thus the more compact arrangements demonstrated in clinics B, E and F may be easier to attend

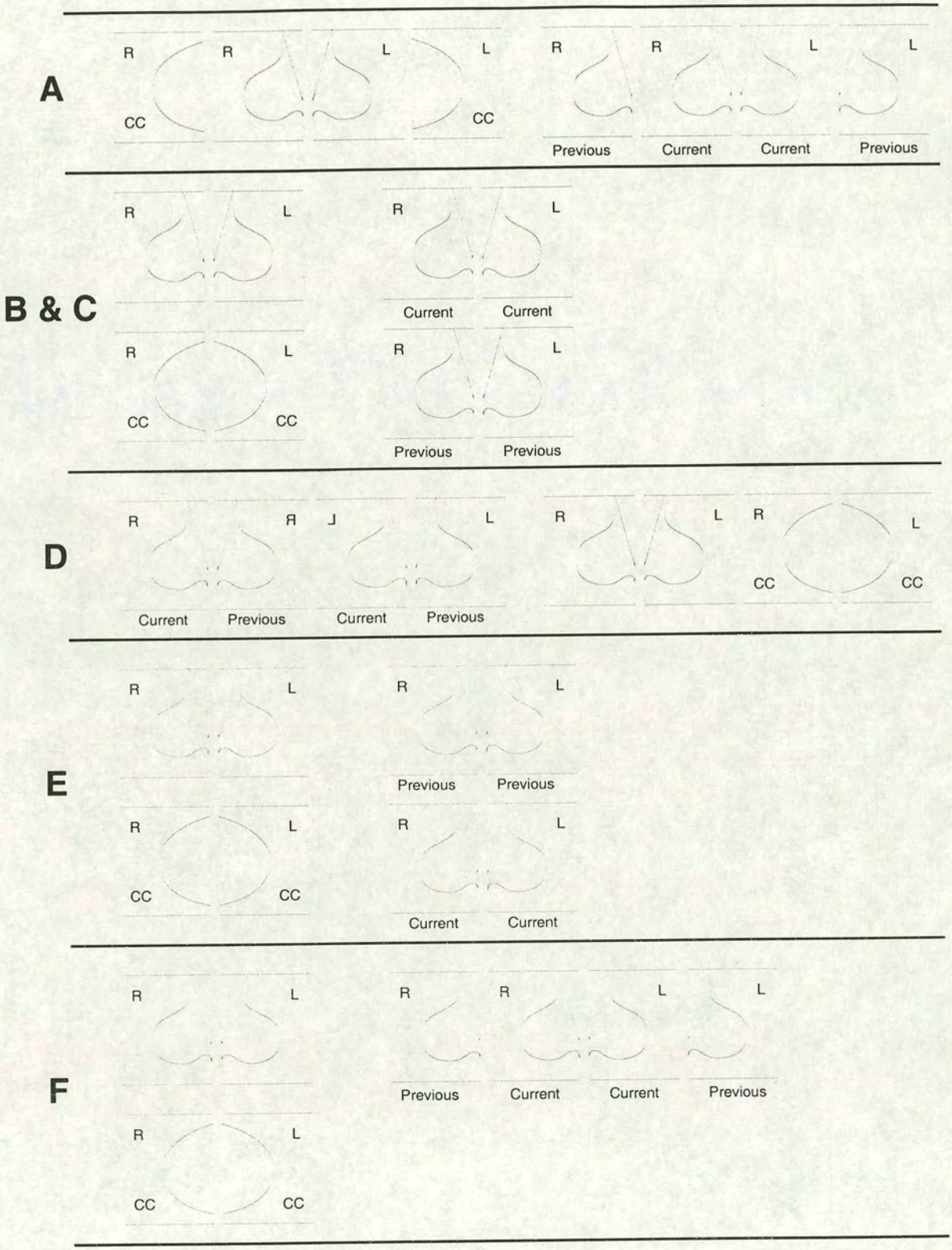


Figure 3.3: Shows how films are presented on viewers in the clinics studied.

to. In addition, readers in clinics other than A and D often make use of electronic devices that synchronise the movement of the viewing bands thus making this compact arrangement manageable on RadX style viewers. So in centres that use a combination of Planilux and RadX viewers, a similar layout style can be easily maintained on each. An exception to this is centre B, where a compact arrangement of films is employed solely on RadX type viewers, with no additional control for synchronising the viewing bands. In centre F, the sole RadX viewer is used for films identified by radiographers as potentially technically imperfect — a purely linear arrangement is employed on this viewer.

Generally artefacts are organised according to how a reader will attend to them, but the specific arrangement of films on the viewer serves also to accentuate significant relationships between different views. Centre D employs a novel arrangement whereby the previous left oblique and the current right oblique are mounted with the face of the film turned towards the viewer. This allows a juxtaposition that emphasises asymmetry over time at the expense of an emphasis on asymmetry between breasts. The arrangement of CC views relative to obliques in centre A facilitates comparison between different views of the same breast. In centre A, incident and prevalent round screens are displayed on different viewers. The reason given for this is that because the position of craniocaudal and previous films relative to the obliques is similar, there is a concern that CC films may be accidentally neglected as previous films are not always attended to. The potential for this type of error may be exacerbated by the greater number of incident compared with prevalent round cases typically generated by screening. Centres C and F also routinely separate prevalent and incident cases, but rather than employing different viewers these are sorted within a batch so that incident cases are examined first, and prevalent round cases later. In centres B and D organisation of prevalent and incident round cases is dependant to a greater extent on how invitations for screening are organised. This sorting is achieved apriori in centre D, as each clinic, and therefore each batch, consists entirely of either prevalent or incident round cases. A similar pattern holds for screening on mobile units in centre B, although a clinic in the static unit would typically consist of both prevalent and incident round cases intermixed. In centre E, prevalent and incident round cases are intermixed for all clinics, and are not sorted when they are put up on the viewer. This may account for the discontinuity in the arrangement of films in centre E, where the oblique view films are placed above CC view films, or below previous obliques thus facilitating a 'context switch' when prevalent round cases are encountered. The layout of films in centre F represents

a hybrid approach. Here a linear arrangement is used for incident round cases, and a compact arrangement for prevalent round cases. It may be that there is a tradeoff between having a compact arrangement on the Planilux viewers, and maintaining a lateral arrangement between previous and current round obliques to aid comparison.

The point here is that the arrangement of evidence for screening is subject to a number of constraints. These are due to both how screening is organised, and the physical properties of the artefacts involved. The final arrangement of artefacts inevitably entails several tradeoffs, including: the emphasis given to different sorts of relationships between mammograms, the ease of establishing the identity of cases (e.g. as either from the incedent or prevalent round) and a regard for the comfort and physical limitations of the reader.

3.5 Reading

Each mammogram is examined by at least one medically qualified film reader who typically is also a trained radiologist. The number of cases read in a single session is highly variable, and will depend on the availability of other readers and the workload in the clinic. A film reader will work through the cases on the viewer and mark his/her decision on the screening form. The decision of a film 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 film reader decides that a diagnosis may be inaccurate because of imperfect mammography then repeat screening films may be requested.

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 cases are removed from the viewer before the normal cases are taken down.

Many English and all Scottish screening centres practice double reading. This involves the separate examination of each case by two film 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 by combining the decisions of the individual readers. In the clinics studied, three different strategies were employed for deciding the final outcome (Table 3.1). 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 decided through discussion between the two initial readers.

Where screening decisions are recorded on a screening form, the second reader has access to the first reader's decisions as a matter of routine. This is because the screening form is attended to as a source of various types of evidence, (for example, 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 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, which 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, for example:

> "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)*

> "[I] believe that blinding will prevent bias when reading. For example, if the first reader indicates a feature is benign then the second reader might be biased into non recall. When blinded [you] have an entirely fresh opinion."
>
> *Interview (written notes fr1-B)*

> "In a double reading team there is a tendency for radiologists to converge in their performance characteristics, maybe due to personality

dominance. Feedback on performance occurs at review clinic and at review of interval cases."

Interview (written notes fr1-E)

One reader suggested that the degree of influence that access to first reader decisions has might be a function of experience:

Reader: "...and the fact that [someone] had already first read it normal, you see that should not make a difference."

Observer: "Do you think it does on occasion?"

Reader: "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 indeed affect decision-making. Data taken from a published double reading study [1] indicates a strong relationship between the reported sensitivity of each of the participating radiologists and the percentage of time they are second reader (Figure 3.4). 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)

Although 'batching' cases when reading may be used as a time-saving measure, it also delays a reader from attending to the evidence on the screening form until a decision has been made. This practice may therefore also serve as an ad-hoc blinding mechanism.

In centre D, readers generally have a preference for reading first, and it was suggested that one reader had a notably strong preference in this respect in order

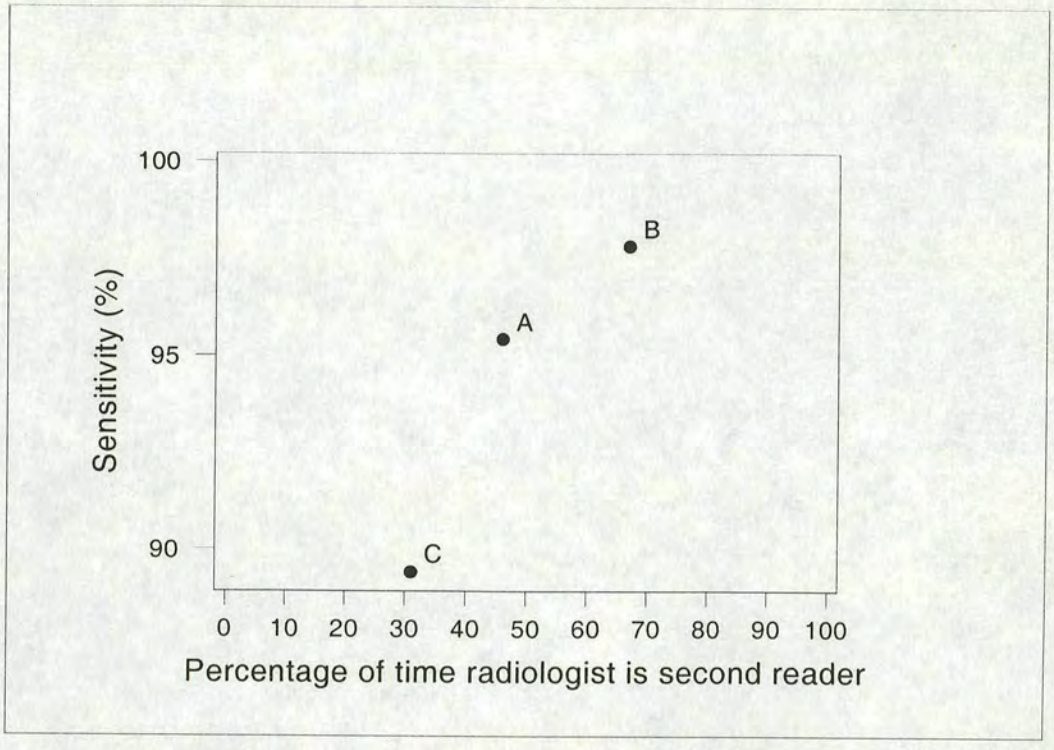


Figure 3.4: Shows the relationship between readers' sensitivity and the percentage of time each reader read second. The letters A, B and C denote the readers participating in the study. The data is from [1], and is discussed in [130]

to get "a clear run at things" (fr1-D). Reading first may be desirable because independence is guaranteed. However, in centre D there are further potential benefits for the first reader. Here assessment clinics are organised so that the radiologist conducting the clinic is usually given cases recalled from batches of cases 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. Thus they seek to establish a balance of experience between reading independently, and being exposed to the first reader's decision. Furthermore, when second reading one reader (fr4-A) reads the batch in 'reverse order'. This

is done with the assumption that the first reader is likely to be more fatigued towards the end of his/her reading session and may therefore be more likely to make mistakes on cases towards the end of the batch. As a second reader, this reader tries to avoid the same pattern of fatigue. Such practices do indicate that a distinction is made between the properties of the first and second read. However, some aspects of clinical practice serve to minimise this distinction. In the clinics studied that use a 'worst opinion recalls' decision pathway it is the responsibility of the second reader to remove cases recalled for assessment from the viewer. The practice is notable because if the first reader recommends recall then this guarantees that the case will actually be recalled — the opinion of the second reader is of little practical consequence. Furthermore, the removal of recalled cases by the first reader is arguably a less error prone handling strategy. Leaving cases on the viewer gives the second reader an opportunity to examine the case and reach his or her own conclusion. This practice also helps maintain a similar context for decision-making for the second reader, who may be working with the expectation that he/she will chance upon significant cases at a particular rate.

A double reading system involving discussion of recall decisions demands a greater degree of commitment from readers. Readers in the clinics studied not only acknowledged the logistical difficulties associated with implementing a system of discussion, but 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: "...if you have a system of discussion to decide recalls then the process might be biased by the dominant personality" (field notes fr1-B).

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)*

Notionally the purpose of double reading is to improve the cancer detection performance of screening. For example, published studies that testify to the benefits (or otherwise) of double reading are typically concerned with its effects

on the sensitivity and specificity of the screening test [1, 3, 25, 34, 33, 122, 128]. Only one makes a passing reference to other potential roles (that it places an emphasis on teamwork [33]). However, in the centres studied the practice of double reading appears to fulfil a number of roles that are arguably as important as any performance effects. These include:

1. a means of training film readers,
2. to give a measure of reassurance, and
3. to provide feedback on performance.

Although there may be acknowledged disadvantages to unblinded double reading, and although readers appear to value their autonomy as decision-makers, the unblinded nature of double reading facilitates the activities listed. These are discussed in detail below.

3.5.1 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 film 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 at this stage, they do not 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. (fr3-B)
- So that any 'unnecessary' recalls can be stopped by the second reader. (fr3-E)
- So that the second reader can act as a check and detect any missed abnormalities. (fr4-B)

Thus training is organised to take advantage of the structure of double-reading to provide a safe and 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 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" (fr1-C).

Moving from reading first to reading first or second may not necessarily be viewed simply as a response to a reader's growing experience. One reader (fr3-D) suggested that it is important for novices to gain experience as an independent first reader, and also by examining the decisions of other film readers made available by reading second. This distinction between the roles of first and second reader is emphasised in centre B where the second reader is seen as being the 'most important', and as having ultimate responsibility for decision-making.

3.5.2 Reassurance

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 felt that she was tired at some points and stated to the clinical director that she wanted 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 (senior) reader (fr1-E) had detected, but since that time, has only 'missed' one that the senior reader had detected. She stated that double reading provides "reassurance" in these circumstances.

Another reader from centre E used 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). [3] He further stated that if he ever discovered that he had missed an obvious cancer then he would go back and read all the cases in that batch again. He recalled an occasion when he discovered (by checking the first reader recalls) that he had missed an "obvious" spiculated lesion. He stated that on this occasion he did read this set again (field notes fr3-E).

3.5.3 Feedback

One role of double reading closely related to that of providing reassurance is to give feedback about a reader's performance. The first reader's opinion on each of a set of cases effectively provides a standard against which the second reader might compare their own decisions. Even in centre C, where double reading is seen primarily as having a training function, one experienced reader suggested:

> "...the two consultants like to read against each other as well as against the inexperienced radiologists."
>
> *Comment made while reading (field notes fr1-C)*

Unblinded double reading provides a framework where this type of comparison may be routinely made. Additional effort is required to access the first reader

[3] Refers to a subjective judgement on a five point scale, where 1 would be benign or normal, and 5 would be definitely malignant.

decision where blinding is enforced. As suggested in the account of one reader's (fr3-A) practice in centre E in the previous section, readers sometimes go to the necessary lengths to obtain the first reader's decision where reading is blinded. So although it was previously suggested that blinding is difficult to implement safely when using a paper based reporting system, the utility of unblinded double reading for providing feedback information may also mitigate in favour of not blinding.

Feedback gained by the second reader in this way may fulfil a number of functions. As suggested above 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 informal monitoring practices had evolved to include notes made on the reporting form, during the course of reading, concerning lesions considered to be benign by the film reader. This practice of annotating benign lesions in effect 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)*

> "It's good to know the second reader has seen the same thing (...) for example, that something hasn't changed (...) the second reader gets conformation that they are thinking along the same lines."
>
> *Interview (written notes fr1-A)*

This suggests that annotation enables both 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

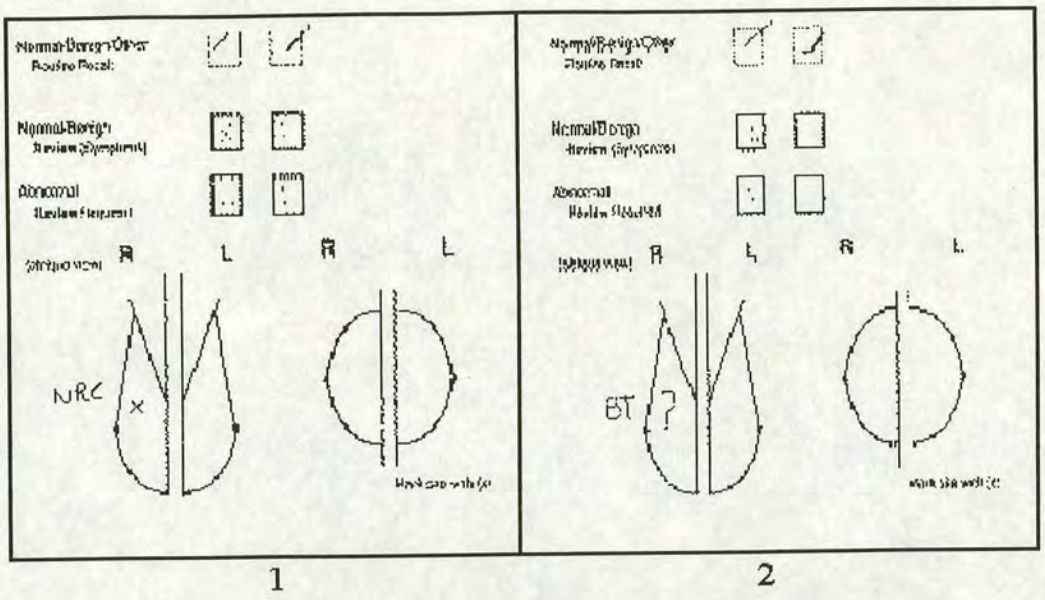


Figure 3.5: Examples of first readers' benign feature annotations.

discharged. Annotation may thus serve to demonstrate a reader's accountability to the decision-making process. An example annotation from centre A is shown in Figure 3.5.

The scope for annotation is exemplified by the comments of one reader, who, on identifying "more new lymph nodes" in a mammogram, stated that this presentation is "quite common" and "not worth commenting on" (field notes fr4-A). The reader in this case did not feel compelled to communicate his observation, implying a tacit understanding that his colleagues possess a similar level of skill and would easily reach a similar interpretation. This does not imply that readers are guaranteed to annotate the same features: individual readers will operate at different confidence thresholds for annotations as they do recall decisions. The same reader later commented that he "annotated benign stuff less than other readers" (field notes fr4-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 their 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.

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, rather than for recalled cases that turn out to be normal. This 'virtuous' difference between reader's interpretations accounts for the performance gains reported for double reading. In fact, double reading serves to compensate for both intra- and inter-observer variations [94]. 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, this is the region where most false negative and false positive 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 between the decision-making and monitoring aspects of a reader's work, where access to a first reader's decision is recognised not only as useful as a metric of performance, but also as potentially harmful if it then serves to bias decision-making.

3.6 Interpretation of evidence

This chapter so far has considered the preparation of screening artefacts, and how they are both organised and attended to. The remainder of the chapter focusses their interpretation.

The principle source of evidence available to a film reader for making a screening decision is obtained by visual inspection of current screening mammograms, which are rendered easily accessable by their arrangement on the viewer. However, readers also have access to a variety of additional sources of evidence, recorded in a variety of ways, and which have differing levels of accessability. Some of the mammograms taken in previous screening episodes may also be made available on the viewer, others are kept in the film bag but may be retrieved and examined during the course of a reading session. Film readers also can access the history taken by the radiographer, and the radiographer's notes about the screening process. Previous screening forms, and therefore previous radiographers' accounts and screening decisions (including the results from any assessment procedures) are also available from the film bag. When reading second, a film reader will typically have access to the first reader's decision and comments.

It is tempting to characterise the skill of a film reader solely in terms of their ability to distinguish benign from malignant presentations in a single pair of mammograms. There is a further temptation to believe that the cognitive mechanisms underpinning this ability comprise the totality of a film reader's skill. In the following sections it is argued that film reading expertise also involves: integrating evidence from multiple sources, assessing the quality of that evidence, examining evidence with an eye for perceived deficiencies in a reader's own abilities and further organising access to evidence to avoid the potential for bias.

Individual mammograms themselves can offer multiple sources of evidence concerning the status of a candidate lesion. One striking feature of readers' commentaries is the ubiquity with which they refer to multiple signs when making a decision. This accumulation of evidence often appeared to weigh both for and against a suspicious outcome, for example:

> "Calc in a cluster, elements are sharp and spiky plus an ill defined lesion. Lesion visible on both views - so recall." [The reader states that she is fairly confident in this decision.]
>
> *Comment made while reading (field notes fr2-C)*

Evidence from different sources is combined: the character of the calcifications, their association with an ill-defined lesion, and the presence of the lesion on both views all contribute to a recall decision, and a statement of confidence.

> "Something like a spiculated lesion on the right hand side, not on the left — so is suspicious — but it is there on the previous films, and able to pick it to bits — not concerned".
>
> *Comment made while reading (field notes fr2-B)*

In the above quote the reader weighs the appearance of the lesion (suspicious), and its asymmetric presentation against evidence from previous films (it is unchanged). The reader also talks about 'picking a lesion to bits', often also referred to as 'undressing a lesion'. Overlapping linear structures can often mimic the appearance of distortions or stellate lesions. If a suspicious lesion has a 'spiky' appearance then readers do not necessarily take this at face value, they will try to determine if the lines can be 'traced through' the lesion, indicating that it is probably composite:

> "Bit denser there - all lines seem to go through it rather than be drawn towards it." [The reader decided the feature was benign.]
>
> *Comment made while reading (field notes fr1-A)*

> "...probable cancer on the right side. Is a density, lines not go through, not compress out, is not on the previous film." [The reader decides to recall the case for assessment.]
>
> *Comment made while reading (field notes fr1-B)*

However, even in combination, signs might conflict or be inconclusive. In one episode a reader was observed to recall for an area of asymmetric density located in a 'danger area' that was not present on films from 6 years ago. He thought that it "might be a cyst as the lady is on HRT, plus it has a smooth upper border where clear". He was also able to see a "hint" that the lesion was present three years ago (field notes fr2-D). In this case, although there was evidence for and against making a recall decision, the evidence for benignity was not sufficiently convincing.

The use of multiple sources of evidence can be viewed as a way of reducing the uncertainty associated with decision-making, but not of mitigating it entirely. Technologies such as PROMAM serve a similar function — they make an additional type of evidence available to film readers with the aim of improving the performance of the screening test. However, taking the view that readers would experience few problems utilising novel types of information would be to ignore the complexity inherent in the task of interpreting conventional sources of evidence.

3.6.1 Hormone Replacement Therapy

The recommendation made in the Forrest report [46] that the UKBSP should target women above the age of 50 was based on results from clinical trials that suggested using mammography to screen younger women is less effective. This is partly because the presence of non-involuted glandular tissue in pre-menopausal women reduces both the sensitivity and specificity of the test. One effect of Hormone Replacement Therapy (HRT) is to partially reverse menopausal changes, leading to a patchy increase in glandular tissue. A film reader from centre E suggested that this may also have an effect on the specificity of mammography:

> "In this area, recall rates do not drop between prevalent round screens and later round screens if [the women] are on HRT. Recalls are less specific because fibroadenomas and cysts just keep growing."
>
> *Comment made while reading (field notes fr2-E)*

A film reader from centre A indicated how expectations based on the natural history of normal breast tissue have to take into account the effects of HRT:

> "Shouldn't get a cyst if didn't have any before, if over fifty and if not on HRT."
>
> *Comment made while reading (field notes fr1-A)*

However, only two of the six centres visited asked attendees at screening if they were taking HRT treatment. A film reader from centre B gave one reason for not asking:

> "Not now ask if lady is on HRT, personally not want to be influenced by apparent increase in density."
>
> *Comment made while reading (field notes fr2-B)*

A reader from centre E echoed this view:

> "Not policy to ask if the lady is on HRT, because can't dismiss new small focal masses even if the lady is on HRT."
>
> *Comment made while reading (field notes fr3-E)*

One reader from centre F offered a different rationale:

> "Not ask attendees at (centre F) whether they are on HRT, because not want to encourage the idea that HRT treatment is linked to breast cancer."
>
> *Comment made while reading (field notes fr1-F)*

In contrast, in centre D where the policy is to ask about HRT status at screening, one reader suggested that:

> "[It is useful to ask if the lady is on HRT, it is] easier to make a decision — [you] can have a reason for something."
>
> *Comment made while reading (field notes fr1-D)*

Thus a tension exists between the utility of knowing whether a woman is taking HRT, and the possibility that attending to this information may serve to bias decision-making. A similar tradeoff was observed between the benefits and the potentially biasing effects of knowing the first reader's decision when double reading. In both cases, access to information may be purposefully managed. In the case of double reading, this may be by done by using formal or ad hoc blinding mechanisms, and in the case of HRT status, by simply not asking women screened whether they are taking HRT.

However, simply because information about HRT status is not available does not mean that readers never use their knowledge about the effects of HRT to account for appearances of a mammogram. For example, a reader working in a clinic that does not ask about HRT made the following comments about a particular case:

> "Diffuse density — not worrying — probably due to HRT."
>
> *Comment made while reading (field notes fr3-E)*

Then later in the same session, for a different case:

> "...previous to new films show an increase in tissue density, but almost certainly this is due to HRT. Various patches of asymmetry because of this are of little interest."
>
> *Comment made while reading (field notes fr3-E)*

A reader in clinic F, was asked if she ever speculated about whether a woman is taking HRT, and how this might relate to her decision-making:

> "The most likely situation where you would speculate on it is when you had previous films — changes in density. For the most part I think you use it to reassure yourself, say what all these extra splodges are — probably due to HRT. If there's more than one extra splodge then that's (usually?) fair enough. But if you have only got one, extra bit, (...?) then you probably are in a situation where you would bring it back to reassure that it was some (...?), and I think you do have to be very careful again not to over rationalise — blame HRT when it is not really HRT to blame."
>
> *Interview (transcript fr2-F)*

The effects of HRT form part of a reader's understanding of how the appearance of the breast may change with time, so it is not surprising that readers may consider HRT as a possible explanation for their observations, even in the absence of definitive information about HRT status. Complete neutrality can be difficult to maintain, even when practical steps have been taken eliminate a potential source of bias. However, in the above quote, the reader from centre F (fr2-F) implies a degree of mental discipline is required, and employed, to limit what might be inferred if speculating about HRT.

3.6.2 Additional views

In addition to the mediolateral oblique view, craniocaudal views are taken for prevalent (first) round screens. In clinic A, prevalent and incident round cases are mounted sorted and mounted on separate viewers for reading. Readers' responses to the separation appear to emphasise the difficulty inherent in reaching decisions for prevalent compared with incident round cases. One stated that decision making is "more difficult", and that "twice as much concentration is required" (field notes fr2-A). Another stated that examining prevalent round cases is "much harder work" (field notes fr1-A). Although the availability of CC views in the prevalent round somewhat compensates for the lack of previous films, it appears that readers still find the task of interpreting prevalent round films more demanding.

Oblique view mammography provides greater coverage of breast tissue than CC views, and so readers tend to "examine the obliques first — then see if can find the interesting feature on the CC" (field notes fr1-A). Some cancers are more obvious on the CC view, so if the reader does not find anything of immediate interest on the Oblique, they will then reverse this process — scrutinising the CC view closely and using the oblique view as a reference (interview, transcript fr2-F).

CC views play a similar role to previous films in that they provide a reference for disambiguating composite shadows:

> "Very large density in top of CC view, nothing equivalent in the oblique view and very 'spread out'." [The reader does not recall the case for assessment.]
>
> *Comment made while reading (field notes fr2-B)*

> "Just checking that [the left of disk on the right hand side] is Ok - not there on CC view."
>
> *Comment made while reading (field notes fr2-B)*

"looking at this density [in the oblique view] but is spread out on the CC so is OK."

Comment made while reading (field notes fr2-B)

Because of the different projection, a lesion apparent in an Oblique view might appear in a number of possible locations on the CC. One reader was observed to use a pen to describe an arc from the nipple to assist searching for the lesion on the CC. (field notes fr1-A).

Some readers suggested that there are specific 'rules of thumb' for attributing degrees of suspicion to the evidence available from a comparison of Oblique and CC views. For example, one reader suggests that "More often than not a suspicious lesion would be apparent in both views" (field notes fr2-C). Another stated that "More likely to get composite shadows on obliques than on CC films, because of the different way the breast is compressed. So composite shadows are more suspicious on CC views" (field notes fr1-A).

Thus it appears that interpretation in two view mammography not only involves an independent assessment of the notable features in each view, but also a comparison between views based on expectations about what relationships might hold for benign compared with malignant lesions.

One reader stated that one of the advantages of two view mammography is that readers can be more sure about their opinion: "they can be more certain in deciding that a film is definitely normal." (interview, written notes fr1-E). This suggests that in addition to improving screening performance, two view mammography may also provide readers with a degree of reassurance concerning the accuracy of their decision-making.

Radiographers at their discretion may take CC views in the incident round if there are difficulties with positioning. For the same reason they may take lateral or mediolateral views. For large breasts it may be the case that the whole of the breast cannot be imaged on one standard size film. This requires that either a number of films are taken to cover the whole of the breast, or, where the facilities are available, a larger film size is used.

Multiple views can be used opportunistically, for example, one reader suggests that:

"Front views are quite useful, because if suspicious of something on one view, then can quite often dismiss it on the other." [For women with larger breasts where there is an overlap in the tissue recorded on films.]

Comment made while reading (field notes fr1-A)

In another episode, where the front of the breast on the oblique was blurred the reader stated "[I] have a CC, so can see area — otherwise it would be a technical recall" (field notes fr1-D).

3.6.3 Previous films

For incident round screens, previous films are loaded onto the viewer for comparison with current round films, thus enabling a reader to assess how a lesion may have changed over time. A lesion may be treated with a greater degree of suspicion if it is new, if it has grown appreciably or has increased in density, or if the number of calcification particles present has increased, for example:

> "Opacity at the bottom of the breast — caught my eye, but was there last time. Microcalc behind the nipple — but was there last time."
>
> *Comment made while reading (field notes fr2-B)*

> "Deciding what to do about this case — it is probably a lymph node — but it has grown..." [The reader recalls the case for assessment.]
>
> *Comment made while reading (field notes fr1-D)*

> "Left breast increase in density more than the right" Checks intermediate films and notes that is on HRT. "Has been like that for some time." [The case is not recalled for assessment]
>
> *Comment made while reading (field notes fr1-D)*

> "More calcification than before. Long coarse calcification - very typical of benign ductal calcification."
>
> *Comment made while reading (field notes fr4-A)*

As the above quotes indicate the use of previous films does not always entail a yes/no indication of suspicion based on an unambiguous indication of change. If a lesion is visible on previous films, then an assessment is made as to whether any changes in size are significant. Readers also have to consider any differences in the context of natural breast changes (for example, the involution of glandular tissue), and how these changes might be affected by external factors, such as HRT. Furthermore, readers have to consider the possibility that an apparent increase in density is artifactual, that is, due to slight changes in projection creating a composite shadow where there was none visible before.

Readers suggest that they frequently refer to previous films to ascertain or confirm the status of a lesion. One reader estimated that he examined previous films for this purpose "every third or fourth case" (field notes fr3-E). Another

suggested that "Maybe half you look back to the previous films" (interview, transcript fr2-F).

Readers were observed to identify tissue or structures in the previous film around the site of the new lesion, sometimes by using other breast features such as blood vessels as landmarks, and then to consider how this tissue might have overlapped to form a composite shadow on the current round films (Comments made by fr1 and fr2 in centre A). This process might not always result in unambiguous conclusions, for example, one reader identified: "Slight distortion, possibly due to existing tissue seen on previous view." which had a "Low level of suspicion" (field notes fr4-A). He marked a no recall decision on the screening form, but changed his mind after examining the film again.

In four of the six clinics, films from the immediately previous screening round are examined. In the other two clinics (D and E), if the woman is being screened for the third or fourth time then films from two screening rounds previous are examined. One reader stated that:

> "...policy decided after looking at cancers that had shown up in the third round, and also at some interval cancers and noticed that there were some changes that were more obvious over 6 years than over three."
>
> *Comment made while reading (field notes fr3-E)*

However, in clinics D and E, readers were observed to occasionally retrieve and examine films from the previous screening round to discount natural changes that also might be emphasised by this longer interval. For example, in one episode the reader noticed an asymmetric increase in density, retrieved and examined the immediately previous films and noted that the women was taking HRT. She then concluded that "...she has been like that for some time" and decided not to recommend recall (field notes fr1-D).

3.6.4 Technical quality

When assessing the evidence in a mammogram readers also consider the quality of that evidence by taking into account the technical quality of the radiography. This can include an assessment of positioning (coverage), exposure (penetration), movement (blurring), film type and processing (brightness and contrast). Insufficient compression can also lead to poor penetration, and because tissue may not be adequately separated, an increased risk of composite shadowing [81].

Assessing variations due to technical quality is particularly important when making a comparison with previous or additional views. Slight changes in pro-

jection between screening rounds can result in apparent changes that are really due to composite shadows on the current round films:

> "Tiny changes in projection can give huge differences in appearance — especially with obliques. Same one taken 5 minutes apart by two different radiographers can look quite different."
>
> *Comment made while reading (field notes fr1-A)*

Some changes in appearance may be due to the approach taken by the radiographer. In some centres, when taking incident round films, the radiographer will examine prevalent round films and adjust the exposure according to the density of the breast to achieve better penetration (field notes fr3-A). Furthermore, coverage (the amount of tissue imaged) may vary between rounds, either because of differences in positioning, or because the presence of skin folds has masked part of the breast in one of the films. For example, in one episode where the previous films for a case had been obtained from a different centre, the reader stated that she believed that the opacities that could be seen on the current films were "present on the previous films", but that the relevant tissue was folded behind the pectoral muscle during compression (field notes fr2-C).

The possibility that the absence of a lesion in a previous film might be explained by imperfect technique can lead to a protracted investigation to determine if this is the case. In one episode, a reader identified what he thought might be a lymph node over the left pectoral muscle that had slightly ill-defined margins and which was not visible on previous films. He said that it may not have been imaged, and that it looked as if more tissue had been imaged on the current films. He retrieved the films from the previous round but still could not locate the lesion. He then placed the current round and previous round films back to back on the viewer to see if the same area has been covered. Finally he decided that the lesion had probably appeared in the interval and recommended a recall for assessment (field notes fr3-E).

Again, readers may also make use of landmarks within the breast to make an assessment of coverage, for example:

> "You look at the vessels, there's that 'Y' shape which is there on that film [current], so anything that is there and down and in from that, which that [the feature] is, is not going to be on that [previous] film."
>
> *Comment made while reading (transcript fr1-F)*

If a woman has moved between screening areas, then an attempt is made to obtain previous films from where she was previously screened. Variations in procedure between centres can be problematic:

> "...different compression and technique can mean that films can look strikingly different."
>
> *Comment made while reading (field notes fr2-C)*

In another example, a reader had discovered asymmetric breast tissue on the right side:

> "This looks like normal breast tissue 'shadowing' but there is nothing resembling it on the previous films. The previous films are not from this screening centre — perhaps that's why they look so different".
>
> *Comment made while reading (field notes fr1-D)*

Processing chemistry or the film type used may also change with time, and this knowledge can be used to explain apparent changes in the appearance of a breast, for example, in centre E, where films taken six years previously are used, one reader was observed to retrieve the most recent previous films for a case where she had discovered a focal increase in density. She could detect no change between the more recent previous films and the current films, and concluded that the apparent increase in density was due to a "whiter film" (field notes fr2-E).

3.6.5 Context

Readers may do not always take the appearance of lesions at 'face value' — they will frequently look to other sources of information to for confirmatory evidence. Further to this, readers also suggest that their approach to interpreting mammograms and their interpretation of specific features within a mammogram can both be dependent on a variety of contextual factors. These are discussed below.

3.6.5.1 Relationships between notable features and breast tissue

Readers may interpret potentially suspicious lesions as being components of more widespread phenomena within the breast:

> "...they're all over the place, they are punctuate, and sort of small and round, and there's nothing irregular about them, and the fact that they are on both breasts — not just on (one side only?)."
>
> *Comment made while reading (transcript fr1-F)*

The same reader later used similar reasoning to decide that an opacity is probably benign:

"If it had been sitting here on its own, I probably would have brought it back, but because there's — yeah maybe it just depends on how I'm feeling I don't know — but on that view it's not significantly different from the others, it looks a bit denser on that view but that could just be (...?). But because there's a lot of them again they're less suspicious, so I think I will let that one go."

Comment made while reading (transcript fr1-F)

Thus an assessment of normality may be based not only on the appearance of the feature itself, but also on the appearance of other, seemingly unrelated, parts of the mammogram. In the second quote the reader notably suggests that she would be more inclined to recall the suspicious feature if it had presented in a different context. In the above examples the reader is making a judgement that the observed presentation is based on some underlying benign causal process which is indicated by multiple occurrences of similar types of feature. Similar types of contextual judgements may also be based on the likelihood that particular circumstances or general appearances can be prone to producing a suspicious looking, but benign, presentations:

"Low density patch. The breast is made of lots of loose bits. [The patch] is similar to lots of other areas [and] so is not a problem"

Comment made while reading (field notes fr1-D

"If you know that they are on HRT for instance you might accept patches in one where you wouldn't accept in another."

Interview (transcript fr2-F)

"Very many lines in the breast can make it look as if there is a feature - [such breast types are] prone to showing composite shadows - play a probability game."

Comment made while reading (field notes fr3-A)

In the first example above the reader suggests that suspicion may be mitigated in the light of known effects of HRT. In the second, it is suggested that the character of the breast itself will make certain types of confusing presentation more likely, and that an analysis of specific features would have to take this into account.

3.6.5.2 Density of breast tissue

Readers often distinguish mammograms according to differences in density of the breast tissue, claiming that it is more difficult to both identify and characterise

lesions in denser breasts. Cancers presenting as ill-defined lesions may be masked by normal tissue, also the contrast between microcalcifications and bright (dense) breast tissue may be reduced. A number of readers draw attention to this difficulty, for example: "Small dense breasts, can't see a lot in there" (field notes fr2-B), "Dense breasts are difficult" (field notes fr2-C). Some readers suggest that they respond to this difficulty by taking "extra care" when examining dense breasts (eg field notes fr2-C). Taking 'extra care' might involve a longer visual search, perhaps with the aid of a magnifying glass. Also, readers may look specifically for features that might be masked, for example: "Dense, patchy ones are quite difficult - try to see through the tissue to any underlying distortion." (field notes fr1-A) "...if they are particularly dense breasts, you do have to look with (...?) care to exclude distortions — that could be hiding within the density." (Interview transcript fr2-F.) This suggests that readers will adjust the effort afforded to examining films depending on their perception of difficulty.

There is a suggestion that variations in tissue types means that "Some breasts are easier to say are OK than others" (field notes fr2-E), for example:

> "Well, if it's a completely lucent breast, and it's been well positioned — a good technique — then you can be almost completely certain. It's very difficult to say that anything is completely normal, and you don't know for instance if the lesion has been left off the mammograms. It's really only in the completely lucent breasts you can be as confident as possible."
>
> *Interview (transcript fr2-F)*

Thus readers work with the knowledge that variations in tissue type can effectively limit the certainty associated with their negative decisions. One reader suggested that it would be useful if a prompting system were able address this problem, if the absence of a prompt could be reliably used to indicate that a breast is completely normal (interview, written notes fr1-E).

3.6.5.3 Location

Studies have indicated that a significant proportion of interval cancers occur behind the glandular disk and in the retro-aureole area [15]. Readers described these areas as either 'danger areas', 'review areas' or as 'the milky way' (because the appearance of the retro-glandular area). Again an adaptive response is indicated — some readers suggest that they 'prompt themselves', or 'were trained' to examine review areas with particular care. Others indicated that readers treat features found in these areas with a greater degree of suspicion, for example, a

reader from centre C stated that a density is "more worrying if found in this area" (field notes fr2-C) and when examining one case, a reader from centre A stated: "so it's an asymmetry in a review area - therefore it's a bit more sinister..." (field notes fr5-A).

Film readers are typically involved in assessment clinics where they may take on the role of clinicians or radiologists. Some comments made by film readers indicate that they sometimes consider the implications of their recall decisions for carrying out assessment procedures:

> "Deciding what to do about this case — it is probably a lymph node — but it has grown — it is not at all dense — very difficult to assess — shouldn't affect judgement — but does — on balance will have to bring back the case".
>
> *Comment made while reading (field notes fr1-D)*

> Radiographer has made a comment 'calcs'. She goes back and identifies the calcs — "so few of them — it would be difficult to biopsy" "Difficult one" — has a look at the previous films. "Calcs are in a line so is more worrying. Would be unhappy if [I] had this case at my assessment clinic...low down in the breast so difficult to biopsy. Could do mags." The reader then recalls the case. She takes another look at the film and states that she thinks that the calcification is probably bening and modifies the recall form.
>
> *Comments made while reading (field notes fr2-E)*

It appears that it may sometimes be difficult for readers to make decisions in a neutral way with respect to their knowledge of assessment procedures. In the first quote the reader demonstrates an awareness of this possibility.

3.6.5.4 Decision-making performance

Another aspect of context is a reader's perception of their own level of skill. Several readers expressed opinions about aspects of their own general performance abilities. For example, one reader stated that she was not so good at detecting distortions, suggesting "either good at them or not good at them" (field notes fr2-A). Another suggested that her particular skill lay in detecting "patchy asymmetry, distortion, microcalcs" (field notes fr1-D). A reader from centre A suggested that he has "a stricter criterion for suspicion" than his colleagues (field notes fr4-R), whereas a reader from centre B suggested that she tends "to bring more cases back than [my] more experienced colleagues" (field notes fr2-B).

Thus it appears that readers have beliefs that they are able to articulate about the level of their skill and abilities relative to those of their colleagues. Readers may also use this knowledge to adapt their reading strategy:

"My approach tends to be to look (positively?) for things that I know I'm not so good at ... there are certain things that you do have to prompt yourself to look at, one of them being the danger areas."

Interview (transcript f2-F)

This is a similar approach to that taken in other circumstances where there are recognised difficulties. For example: identifying distortion of architecture in dense breasts, or as the reader above suggests, detecting lesions in the 'danger areas'.

One way readers might monitor their day-to-day performance is to compare their recall rate with the frequency at which they might expect to recall. Suspicious cases should present randomly, but readers suggest that they are aware that variations in distribution of recalls over short time periods can potentially influence their decision-making:

"Paranoia can set in if have a large number of films that have passed as normal — might think 'what have I missed?'."

Comment made while reading (field notes fr1-B)

"If you get to the end of a session, the end of a pile of reporting and you haven't recalled anything, then you think 'this is (...), maybe I've missed something' then in the next bunch you find that you will recall every other one. So it averages out."

Interview (transcript fr2-F)

3.7 Discussion

3.7.1 Interpretation of evidence

A description was given in Chapter 1 of presentations that might indicate a malignant process. However, categorisation of lesions is less clearly delineated than this taxonomy might suggest. Lesions are often indicated by a combination of traits, each adding to a reader's suspicion by degrees according to their saliency. However, readers do not take the appearance of a lesion at 'face value' — they actively explore how appearances might be accounted for by artefactual or by benign presentations. This might involve a more detailed appraisal of the lesion itself ('undressing'), or reference to other types of available evidence (previous films, additional views, radiographers' notes).

Readers also consider how appearance might be related to contextual factors, such as the general appearance of the breast and expectations about appearance

based upon the age of the women screened. These factors might indicate a greater or lesser likelihood that a suspect lesion is actually benign or an artefact.

Thus experiential knowledge of typical appearances of benign and malignant lesions is not the only type of knowledge that is brought to bear. Interpretation is related to a wider understanding of how the three dimensional structures of the breast are represented on a mammogram, and of the natural history of normal breast tissue. Decisions are also made in the light of knowledge about how the presentation of both benign and malignant lesions might change over time, and how their appearance might be altered under different projections.

In contrast with the initial visual appraisal, this extended approach to interpretation is problem-solving in character. For example, it may not be immediately obvious which region in a CC or a previous film corresponds with that containing the lesion – readers may have to be resourceful in identifying landmarks in order to make a comparison, and they may then have to decide if the lesion is an artefact (composite shadow) by reasoning about the redistribution of normal tissue in the current projection. Readers may also make use of the evidence available to them in creative and opportunistic ways, for example, by using the area of overlap on multiple films to get a 'second look' at an area of interest.

Often these activities involve more than just an appraisal of the visual and documentary evidence at hand — readers may seek out the decision made in the previous screening round and examine films from other screening rounds or previously attended assessment clinics. Additionally, they will re-organise artefacts in order to make particular relationships clear, for example, by putting a previous and a current round film back to back to assess coverage.

Even when all the available evidence has been exhausted, it may still be difficult for a reader to formulate a clear interpretation — some ambiguity may remain. As well as specific presentations that are difficult to interpret, there are circumstances where it is more difficult for a reader to be certain of their decisions, for example, where there is dense breast tissue, or when interpreting prevalent round cases.

Readers will also assess the quality of screening evidence as part of the decision-making process. Thus the interpretation of evidence is tied closely to an understanding of the process and circumstances of its production. Examples of this include a reader's assessment of the technical quality of a film and of the variations that might be due to different processing regimes in other screening centres. If the technical quality of a film is poor, then a reader may request a 'technical recall' — the woman is invited back to the clinic and the screening procedure is

repeated. However, technical quality is not only dependant on the skill of the radiographer, but also on the physiology — there can be circumstances where it is difficult or impossible to achieve technically perfect mammograms. Furthermore, if there are lesions on the surface of the breast (moles, scars etc), these can appear suspicious on the mammogram and may result in an unwarranted recall to an assessment clinic.

3.7.2 Collaborative aspects of screening

Radiographers and dark room technicians collaborate with readers' decision-making in an indirect way by contriving an appropriate and supportive arrangement of the artefacts involved in screening. Radiographers can be thought of as collaborating more directly in the decision-making by making additions to the content of these artefacts. In centre E this includes the diagnostic opinion of the radiographer, but in all centres radiographers supply information necessary to prevent erroneous interpretations. Because of the large numbers of women screened, the production and interpretation of mammograms is rarely carried out contemporaneously, thus a film reader will lack first hand experience of any difficulties encountered or of peculiarities associated with individual cases that might have a bearing on their interpretation. Radiographers will give an account of such occurrences by making notes on the screening form which serve to tie together these temporally separated activities. In doing this radiographers demonstrate an awareness of the nature of a film reader's skill, and in particular of specific limitations of its application.

In the clinics studied, the practice of double reading represents a collaborative approach to decision-making. Although the notional purpose of double reading is to improve detection performance, it also enables the responsibility for decision-making to be shared. In this way readers gain reassurance that the effects of variations in their performance are minimised. Also, double reading is exploited as an informal mechanism for both the coordination and monitoring of performance. The utility of double reading for this purpose has been extended in some clinics by the practice of providing annotations for the benefit of the second reader. This may serve to make readers more accountable by demonstrating their vigilance and skill. It may also serve to reinforce normative interpretations within readers' community of practice. However, readers are wary of explicit collaboration to decide individual cases, and may purposefully manage their access to the first reader's decision to reduce the possibility of bias.

The deployment of prompting technology may have an impact on the informal

collaborative dimensions to screening practice if, for example, double reading was to be replaced by a computer assisted single reading. The standard available for comparison of reader's performance would be provided by the system's, rather than another film reader's, responses. Although it may be possible to still derive a degree of reassurance that a thorough investigation has been made, the opportunities for establishing and maintaining interpretive norms would be diminished.

3.7.3 Reflective application of skill

The results of this study suggest that film readers demonstrate an awareness of the extent and limitations of their expertise. One implication of this is the way readers strategically bring resources to bear to compensate for perceived limitations. For example, they will pay greater attention to 'danger areas', they will examine dense breasts more carefully, and will be on the alert for features that are easily masked, or that they believe themselves to be poor at detecting. Readers are also concerned that although access to certain types of evidence (for example HRT status or the first reader's decision) can be potentially useful, it might also serve to bias their decision-making. When seeking to account for the visual evidence readers express a concern that it is possible to 'over rationalise' where there is uncertainty — that is, to make an inappropriate decision by giving too much weight to a particular piece of evidence.

Readers therefore 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 to how particular conclusions are drawn from particular types of evidence, and suggests pitfalls and biases to which a reader may be subject. Readers apply their skill in a self-conscious or reflective way; they are selective about what evidence they will consider and will actively manage their access to that evidence to avoid preconceptions and to maintain impartiality. Examples of this include not asking about HRT status at screening, or not attending to a first reader's decision until they have formulated their own. In this way readers demonstrate and seek to maintain their independence as decision-makers.

3.7.4 Implications for prompting systems

The treatment of information within an image by a computer based detection system is likely to be less exhaustive than that of a human observer. For example, a system might be able to reliably detect micro-calcification clusters, but

be unable to make an interpretation based on the character of the constituent particles. Furthermore, the scope of the system's analysis is often highly localised, it will not be able to associate the occurrence of individual features with more global phenomena, for example, by considering the significance of a 'mass like' feature in the context of 'busy breasts'. Currently, the PROMAM system cannot make comparisons between left and right breasts, between views or with previous films, does not consider the age of the women, the circumstances of the mammogram's production, or the technical quality of the films. Thus in contrast to a human observer, a computer based detection 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, computer based detection systems are unable 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 on a computer system's ability imply that a reader cannot use their knowledge of the behaviour of film readers (as they do to interpret the decisions and annotations of colleagues) to reliably account for the behaviour the system. Furthermore, the PROMAM system is unable to supply an account of its own decisions.

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 effect.

Chapter 4

Subjective responses to prompting

4.1 Introduction

Prompting studies reported in the literature have largely been concerned with quantitative evaluation, often necessitating manipulation of either the system's response, reporting conditions (for example, by imposing a time limit) or the composition of test sets (such that they are not typical of what a film reader might expect to see during the course of a 'typical' reading session). Their goal has either been to demonstrate the effectiveness of a particular prompting regime, or to establish minimum performance requirements for prompting systems more generally.

However, quantitative methods by themselves do not reveal how film readers use and make sense of prompts, nor how they subjectively evaluate a system's performance and abilities. These issues, although largely unexplored, are potentially significant. For example, without an understanding of how prompts are interpreted it is difficult to determine whether a system is actually being used as intended. A reader's subjective appraisal of a system may inform their perception of its credibility and also its acceptability as a useful tool in the context of other demands on readers' attention and time.

This chapter details an experiment performed to ascertain readers' subjective responses to different prompting regimes generated by the PROMAM system while maintaining, as far as possible, typical screening practice. In particular, realism was established in the following ways:

Representative film sets Test sets of films were selected from four typical days screening at the South East Scotland Breast Screening Centre (SESBSC).

Realistic conditions 'Normal' reading conditions were simulated, including use

of standard reporting forms. A reading protocol was adopted that might be preferred for prompt usage in actual clinical practice.

Realistic system Unmodified output from the two PROMAM feature detection algorithms was generated at different operating points.

The use of representative film sets precludes gathering data with respect to any performance gain that might be achieved by using the system. However, their use does provide a basis for understanding how prompting information might be routinely used and interpreted.

Previously, Hutt had suggested that in order to effect an improvement in observer performance, a system's FP rate should not exceed its TP rate [70]. One aim of this experiment was to examine whether Hutt's conjecture correlates with readers' subjective tolerance of PROMAM's FP burden. Other aims included an assessment of the time penalties due to using the system, the effect of different prompting rates on readers' specificity, and to examine closely how the readers use the prompt sheets in conjunction with their normal reading procedure.

4.2 Material and methods

4.2.1 Algorithms

The PROMAM system comprises two feature detection algorithms designed to detect and prompt for microcalcification clusters and ill-defined lesions respectively.

4.2.1.1 The microcalcification detection algorithm

The core of the microcalcification detection algorithm comes from work by Karssemeijer [78, 79]. The algorithm consists of two important stages. These are:

- noise equalisation (using iso-precision scaling), and
- iterative labelling of possible calcification particles (using a Markov random field model).

Iso-precision scaling is a technique developed by Karssemeijer in which the variable noise level within a mammogram is replaced with a constant, known, noise level. This allows algorithms to estimate background noise locally and hence detect unexpected deviations which may indicate the presence of a particle of calcification. Groups of pixels identified as potential microcalcification particles

are iteratively assigned a probability that they form part of a cluster, that is, they are associated with similarly identified pixels in their neighbourhood. During each iteration particles with a low probability are eliminated, and stricter criteria for defining 'neighbourhood' applied. In this way spurious isolated candidate particles are removed. The final iteration generates a prompt for a cluster of surviving candidate calcification particles that satisfy two clustering rules. The first rule states that two candidate particles belong to the same cluster if they are closer together than some specified critical distance. Thus a prompt might be produced for a line of calcification particles that each satisfy this condition, as well as for a more compact distribution. The second states that a cluster consists of a minimum number of calcification particles — clusters with fewer particles are discarded.

In summary, the microcalcification detection algorithm, in order to generate prompts, uses notions about both the properties of individual calcification particles and the tendency for radiologically significant calcification to occur in clusters. However, this is done in a limited way. The algorithm ignores some of the information used by film readers to distinguish between benign and malignant clusters, for example, as the morphology of individual particles and of the cluster as a whole. The microcalcification detection algorithm will generate FPs for types of calcification that are easily and routinely dismissed by film readers, such as vascular calcification. Furthermore, the algorithm will occasionally mislabel artefacts or noise within the image as calcification particles — generating FP prompts (perhaps in combination with real calcification particles) where the final clustering rules are satisfied. Conversely, FN prompts can be produced if particles belonging to a malignant cluster are mislabelled as breast tissue and where any remaining particles fail to satisfy the clustering rules. One consequence of this highly mechanistic approach is that TP and FP decisions made by the system are likely to have a different character to those made by a film reader.

4.2.1.2 The ill-defined lesion detection algorithm

The ill-defined lesion detection algorithm consists of three important stages:

- multi-resolution analysis,
- feature segmentation and
- classification.

Multi-resolution analysis is a pre-processing step aimed at simplifying the process of identifying possible ill-defined lesions; the effect is to reduce the complexity

of the image by reducing 'clutter'. This involves using a maximum entropy technique to decompose the original image into sub-images of differing scale sizes [96]. Each sub-image represents what would be seen if features with sizes in a given range (say between 2.5m and 15mm) are viewed without the imposition of features of other sizes, but with visual properties such as shape, size, texture and density preserved.

Following multi-resolution analysis, the feature segmentation step seeks to delineate candidate regions for subsequent classification. This is achieved by iteratively marking regions of similar brightness until some boundary criteria are met. The result is a map of the breast, or a series of 'candidate' regions, corresponding to the breast's underlying 'lumpy' structure.

For each candidate region a series of parameters are obtained, including for example, shape, brightness, texture and degree of isolation. Classification of candidate regions as 'normal' (those not prompted) and 'suspicious' (those prompted) is done by using statistical methods to compare extracted parameter values with values previously obtained from sets of known benign and malignant cases. It is worth emphasising that the parameters used correspond to local properties of a candidate region in a single image. The algorithm does not utilise all of the information available to a human film reader to assess suspicion, such as an overall impression of the image, bilateral comparisons, and information from previous films or CC views. Furthermore, the algorithm does not target all of the features that might be classified as belonging to the broad category of ill-defined lesions, such as spiculated lesions that do not have a central mass. The performance of the ill-defined lesion algorithm also depends on the accuracy of the segmentation step, how well the extracted parameters can be used to distinguish between benign and malignant presentations and the number and range of features on which the system has been trained.

4.2.1.3 Operating points

Three operating points were chosen for each algorithm to generate conditions with the prompt rates shown in Table 4.1. The associated sensitivities were derived by running the algorithm on independent sets of pathology proven cancers obtained from the SESBSC.

In the 'high' sensitivity condition, the operating points used correspond to the highest sensitivity obtainable from each algorithm while still producing a reasonable specificity. In the 'medium' and 'low' conditions, sensitivity was sacrificed in favour of obtaining subjects' opinions in respect of systems with improved

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

Table 4.1: Sensitivities and corresponding average prompt rates for the prompted conditions. Prompt rates refer to the fraction of cases with at least one prompt from a given algorithm.

specificity. It was assumed that subjects would not be able to make an accurate assessment of sensitivity because there were only two pathology proven cancers in the test set. It was anticipated that subjects would rate the system's performance according to the FP burden that they would be willing to accept. Consequently, subjects were not informed of the actual sensitivity of the system in each condition, only that they were reading conditions where the sensitivity was 'high', 'medium' or 'low'. Subjects were also told the approximate prompt rate of each of the algorithms, as indicated in Table 4.1.

As can be seen from Table 4.1 the ill-defined lesion algorithm is less mature than the microcalcification detection algorithm, prompting 1 in 2 cases with a sensitivity of 62% compared with the microcalcification algorithm's 94% sensitivity with only 1 in 3 cases prompted. Because the algorithms operate independently, the FP rate will be higher than that produced by the worst performer.

4.2.2 Test sets

Films representing four entire days screening at the SESBSC, totalling 464 cases, were scanned and analysed by the PROMAM system. The only selection criteria applied were that, on each day, morning and afternoon screening clinics would be available from both sessions held in the centre itself and from sessions held in mobile screening units. This was done to ensure a sufficient volume of cases, and also to provide a consistent balance between static and mobile screens. At screening, 25 of the 464 cases were recalled for assessment, two of which were discovered to be cancers. At the time of the experiment interval cancer data was not available for the selected cases, so it is possible that the test set included film reader FNs.

Each of the cases were annotated by the project radiographer with respect to tissue density (lucent, medium or dense) and nodularity [1] (present or absent). Four balanced test sets, each consisting of 116 cases, were randomly selected from

[1] 'Nodularity' refers to the 'lumpiness' of the underlying texture of the breast tissue.

the pooled day's samples. The sets were balanced with respect to the number of cases recalled for assessment, their source (static or mobile), density of the breast tissue and nodularity. The pathology proven cancers were treated as cases recalled for assessment for the purpose of randomisation.

The test sets originated prior to the practice of taking CC views for incident round screens, so CC views were not available for making the original screening decision, nor for decisions made during the course of the experiment. Original screening films were used for the experiment. For logistical reasons (availability of the films, and of viewer space) previous films were not made available to subjects.

4.2.3 Protocol

Four experienced radiologists who had no previous involvement with the development of the PROMAM system nor exposure to the cases used in the test sets, were recruited as subjects from two Scottish screening centres. Subjects are referred to by the letters A to D to preserve anonymity.

The experiment consisted of four conditions, three were prompted at different rates, one was an unprompted control. Each condition consisted of 116 cases. The first five cases of each condition were used to familiarise the subjects with experimental procedure. The remaining cases were read in two sessions consisting of 56 and 55 cases respectively. There was a 15 minute break between these sessions. A Graeco-Latin square design was used to enable effects due to changes in prompt rate to be isolated from subject effects, session effects, and effects due to differences in the test sets. Each subject read each condition, but on different film sets (Table 4.2).

Subject	Sessions			
	I	II	III	IV
A	c2	d3	b1	a0
B	d1	c0	a2	b3
C	a3	b2	d0	c1
D	b0	a1	c3	d2

Table 4.2: The Graeco-Latin square used. a-d refer to film sets. 0-3 refer to the prompt rates (conditions) as None, Low, Medium and High respectively.

Prompt sheets consisted of a hard-copy, low resolution image of the mammogram pair with prompt information superimposed. Prompts for ill-defined lesions consisted of an ellipse surrounding the suspect region, and prompts for microcalcifications consisted of an irregular outline of the potential cluster. An example is shown in Figure 4.1. Prompt sheets were attached to the reporting forms via

a paper clip in such a way that a subject would have to lift the reporting form to examine the prompt sheet. A prompt sheet was produced for each case irrespective of whether that case was actually prompted or not, so that absence of a prompt sheet could not be construed as an erroneous omission.

In the prompted condition, radiologists were asked to use the following reporting protocol:

1. Examine the films,
2. examine the prompt sheet,
3. record a decision on the prompt form,
4. move on to the next case.

The aim was to enable subjects to form an independent assessment of each case before attending to prompting information, thereby reducing the possibility that decisions might be unduely biased by the system's response.

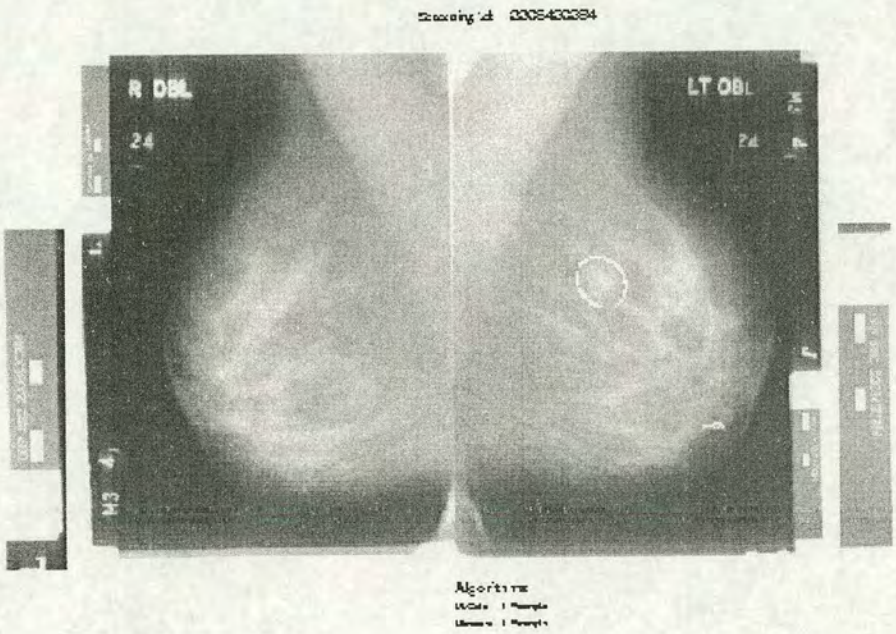


Figure 4.1: An example of a prompt sheet

4.2.4 Data collection

For each case, subjects were asked to mark their decision on the standard reporting form as either recall for assessment, routine recall or technical recall. Where a

recall for assessment was made, subjects were asked to indicate whether the system had correctly prompted for the recalled for feature, to mark its location on a breast schematic and identify the lesion type (for example, mass, calcification cluster etc.).

Questionnaires were administered before and after the experiment and after each condition. A 20 item Likert test was used to assess subjects' attitudes to the system after each condition, where a higher total score indicates a more favourable assessment. [2] In addition, an observation protocol was used to record subjects' actions as they examined each case. This is described fully in section 4.3.5. Copies of questionnaires and instruction sheets for this experiment are included in Appendix C.

4.3 Results

4.3.1 Recall Rate

Figures 4.2 and 4.3 show the number of recall for assessment decisions made by condition. Wald Statistics for type 3 analysis indicate that radiologist differences are the most significant contributor to the observed variation in the recall rate, followed by set difference and session differences. Condition differences failed to reach significance at the 5% confidence level (Table 4.3, a Pr>Chi value of less than 0.05 is significant).

Source	DF	ChiSquare	Pr>Chi
READER	3	29.2525	0.0001
CONDITION	3	7.3865	0.0605
SET	3	14.7659	0.0020
SESSION	3	10.8314	0.0127

Table 4.3: Wald Statistics For Type 3 Analysis

This result indicates the complexity of interactions between film set, and inter and intra-observer variations. It is possible to explore the nature of some of this complexity. Table 4.4 shows the recall decisions made by subjects broken down according to the three categories of tissue density used to classify each case. The visible trend is for greater numbers of technical recalls, and recalls for assessment, to be made for denser breasts. A chi-squared test shows that these results would not be expected by chance if subjects' decisions and tissue type were independent ($\chi^2 = 9.505$, df=2, p=0.050). For comparison, a similar trend is also observed in

[2] A description of the Likert test is included in Appendix A

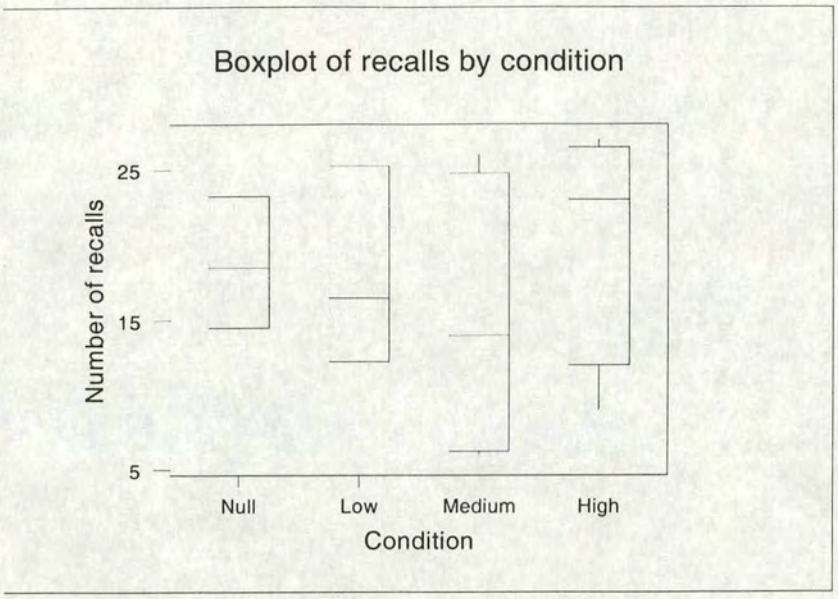


Figure 4.2: Shows recalls made against condition

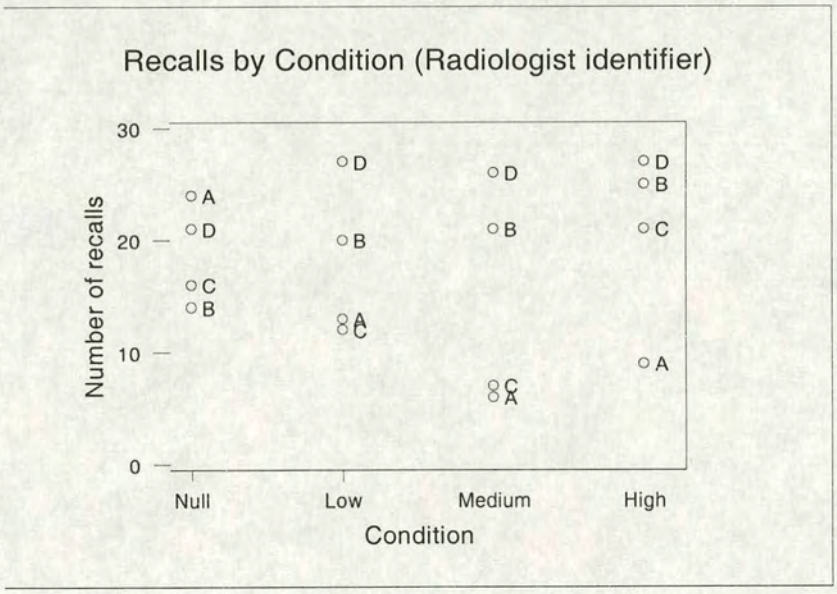


Figure 4.3: Shows recalls made against condition

the original decisions made at screening for these cases (Table 4.5). Again, the differences are significant ($\chi^2 = 23.621$, df=4, p<0.001).

Subjects did not have access to previous films during the course of the experiment. A comparison of recall rate by screening round between original screening

Subjects' decision	Tissue density		
	Lucent	Medium	Dense
Routine recall	267 (84.5%)	1027 (78.3%)	111 (74.5%)
Recall for assessment	40 (12.7%)	222 (16.9%)	27 (18.1%)
Technical recall	9 (2.8%)	62 (4.7%)	11 (7.4%)

Table 4.4: Shows the decision made by subjects for each case according to tissue density. Because each set was read four times there are four decisions per case.

Original decision	Tissue density		
	Lucent	Medium	Dense
Routine recall	304 (96.2%)	1171 (89.3%)	129 (86.6%)
Recall for assessment	0 (0.0%)	84 (6.4%)	12 (8.1%)
Technical recall	12 (3.8%)	56 (4.3%)	8 (5.4%)

Table 4.5: Shows the original decision made at screening for each case by tissue density. Because the cases were double-read, there are two decisions per case.

decisions and decisions made under experimental conditions reveal one implication of this omission. Table 4.6 demonstrates a tendency for fewer decisions recalling for technical reasons, or for assessment to be made where previous films are available (that is, where it is not a woman's first screening visit). The film sets used dated from before CC views were routinely obtained for first time screens. While this difference is significant ($\chi^2 = 14.834$, df=2, p=0.001), no significant difference is observed for a similar analysis of the decisions made under experimental conditions (Table 4.7, $\chi^2 = 2.530$, df=2, p=0.283). It appears that one effect of access to previous films is to improve the specificity of the screening test.

Original decision	Screening round	
	First	Subsequent
Routine recall	460 (86.5%)	1144 (92.0%)
Recall for assessment	36 (6.8%)	60 (4.8%)
Technical recall	36 (6.8%)	40 (3.2%)

Table 4.6: Shows the original decision made at screening for first time and subsequent screening visits. Because the cases were double-read, there are two decisions per case.

Finally, an analysis of system performance with respect to tissue density reveals some interesting trends. Tables 4.8 and 4.9 show the number of cases with at least one ill-defined lesion prompt and one microcalcification prompt respectively. It appears that the ill-defined lesion algorithm is more likely to produce FP prompts with increasing tissue density ($\chi^2 = 17.953$, df=2, p<0.001). Conversely, the microcalcification algorithm is less likely to produce FP prompts with

Subjects' decision	Screening round	
	First	Subsequent
Routine recall	415 (78%)	990 (79.6%)
Recall for assessment	86 (16.2%)	203 (16.3%)
Technical recall	31 (5.8%)	51 (4.1%)

Table 4.7: Shows the decision made by subjects for each case according to screening round. Because each set was read four times there are four decisions per case.

increasing tissue density ($\chi^2 = 6.854$, df=2, p=0.033). An increased frequency of composite structures with increasing tissue density might be a possible explanation for the behaviour of the ill-defined lesion algorithm. This would lead to a greater number of candidate lesions, and consequently to a greater number of FPs. Increasing tissue density may also have the effect of reducing the contrast of calcification particles (and image features and flaws that might be mistaken for calcification). This would have the effect of reducing the probability that a potential calcification particle would be labelled as such by the micro-calcification detection algorithm, thus reducing the overall prompt rate with increasing tissue density.

Mass prompt present	Tissue density		
	Lucent	Medium	Dense
No	56 (70.9%)	157 (48.0%)	13 (34.2%)
Yes	23 (29.1%)	170 (52.0%)	25 (65.8%)

Table 4.8: Shows the number of cases prompted at least once by the ill-defined lesion detection algorithm according to tissue density.

Calc prompt present	Tissue density		
	Lucent	Medium	Dense
No	48 (60.8%)	230 (70.3%)	32 (84.2%)
Yes	31 (39.2%)	97 (29.7%)	6 (15.8%)

Table 4.9: Shows the number of cases prompted at least once by the microcalcification detection algorithm algorithm according to tissue density.

4.3.2 Timing

There is a visual trend for the time to complete conditions to increase with the sensitivity of the detection algorithms (Figure 4.4). However, none of the radiologists felt that use of the prompting system would significantly increase reporting time. All either disagreed or strongly disagreed with the statement "The system

will be time consuming to use" in the attitude test. Also, in the post-session questionnaire, one subject repeatedly volunteered the opinion that the system was easy and quick to use (Table 4.14).

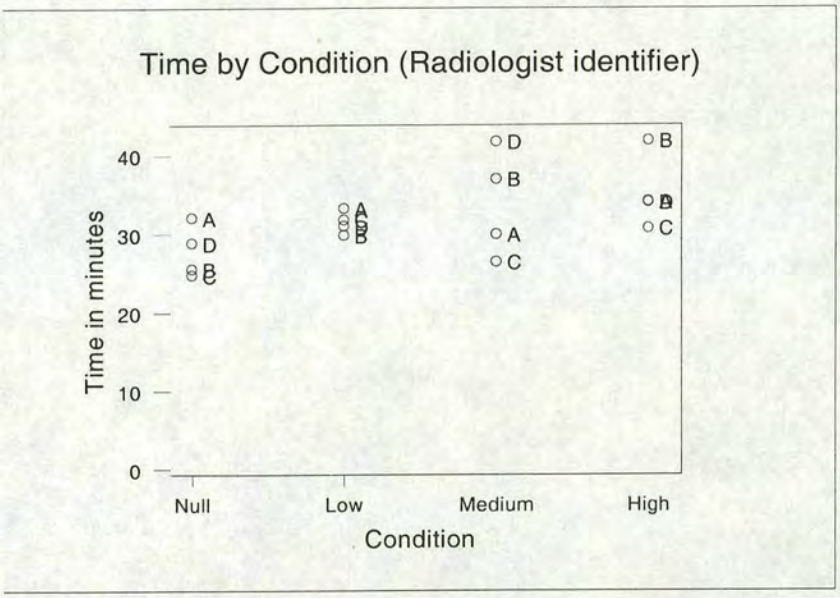


Figure 4.4: Shows time to complete against condition

4.3.3 Opinion

Opinion was measured in a number of different ways, including a 20 item Likert test as a general measure of attitude towards each condition. In this test subjects rated their agreement with a series of statements expressing an opinion concerning the condition they had just read. Assuming that the questions accurately reflect attitude, the higher the total score in a given condition the more favourable the disposition towards the system in that condition. Figure 4.5 shows a box plot of Likert score against condition.

Figure 4.6 shows the same data with the individual scores for each subject (A-D). Except for subject A, scores increase monotonically with increasing sensitivity, suggesting that subjects are better disposed to the system as the prompt rate increases. Wald statistics for type 3 analysis indicate that while effects due to condition and subject contribute strongly to the differences in the Likert score, effects of condition are also significant (Table 4.10).

In the post-session questionnaire subjects were presented with the following statements pertaining to the utility of the system and were asked to indicate

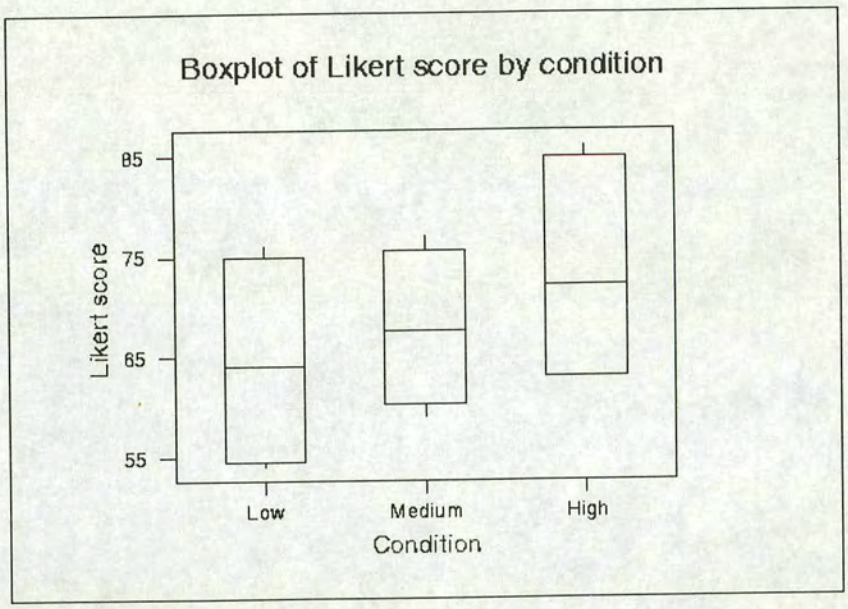


Figure 4.5: Shows Likert score against condition

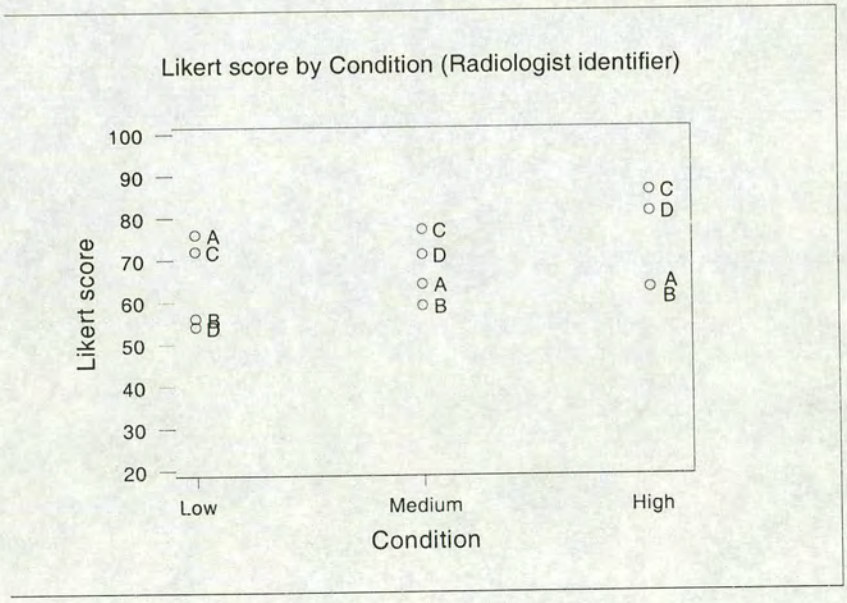


Figure 4.6: Shows Likert score against condition

whether they agreed or disagreed with each:

- Overall, this system would be useful to you in a screening context as it currently stands.

Source	DF	ChiSquare	Pr>Chi
READER	3	40.3323	0.0001
CONDITION	2	9.1863	0.0101
SESSION	3	17.3404	0.0006

Table 4.10: Wald Statistics For Type 3 Analysis

- The mass detection component of this system would be useful to you as it currently stands.
- The micro-calcification detection component of this system would be useful to you as it currently stands.

Figure 4.7 shows the pooled results with respect to condition. The similarity in responses to the 'Overall' and 'Mass component' questions suggests that the performance of mass detection algorithm is the limiting factor when making an overall assessment. The results indicate that the prefered operating point for the micro-calcification algorithm appears to be at the medium sensitivity. However, it might be the 'quality' of the prompts, rather than the overall prompt rate that has influenced this judgement. For example, in response to the question "What do you think the system's weaknesses are?" in the post experiment questionnaire most complaints concerning prompts for vascular calcifications were made about the high sensitivity condition.

After each condition the radiologists were asked to rate the sensitivity of the system. They were asked to state if the mass component, the calcification component and the system overall was too sensitive, not sensitive enough — or just right. The pooled results are shown in Figure 4.8.

Again it appears that the performance of the mass algorithm is the dominant factor in determining subjects' assessment of the system as a whole. It appears that the Medium sensitivity setting represents the prefered configuration for calcifications. However the distinction is less clear cut than for the previous question ("Would this system be useful...") and the picture is further complicated by claims that the system might be 'too sensitive' made by some subjects and 'not sensitive enough' by others at its highest sensitivity. There is ambiguity possible when asking subjects to rate the system with respect to sensitivity; if a system produces too many FPs it might be described as being 'too sensitive'. In this case specificity is really being referred to, but in terms of sensitivity. At the same time, a system might also be described as not sensitive enough if it doesn't prompt for clusters that a film reader might expect it to prompt for.

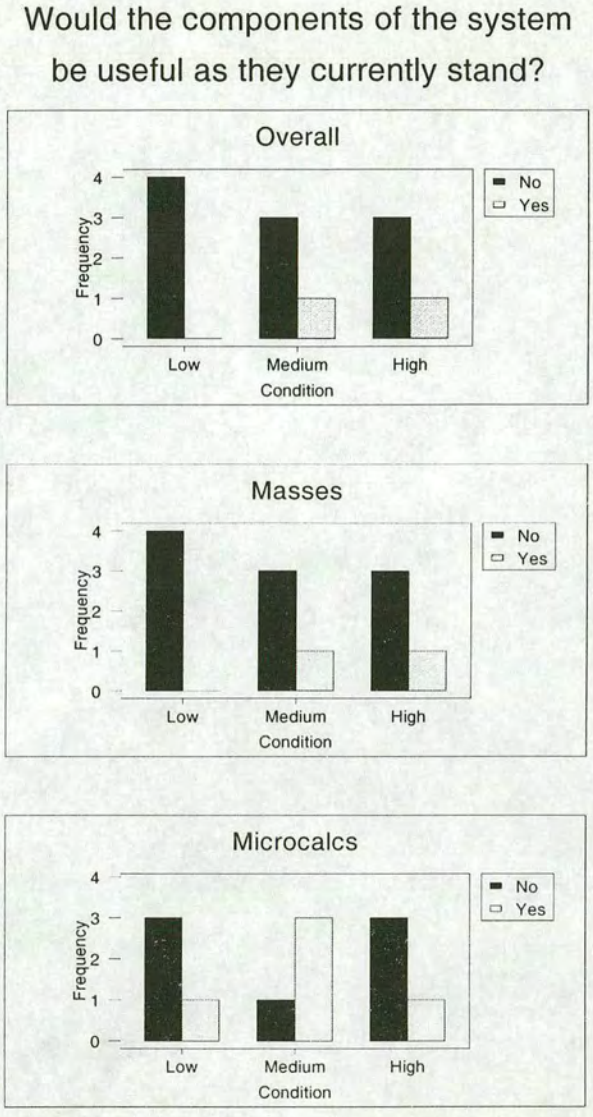


Figure 4.7: Bar chart showing responses to the question: "Do you think that the system (or the mass and calc components of the system) would be useful to you as it currently stands?"

Before and after the experiment, subjects were asked to rate statemens referring to different possible system configurations on a scale of 1 to 5 (where a score of one would indicate that the configuration is most useful, and a score of five, least useful). The configurations suggested were:

1. High prompt rate, where most of the features prompted for are benign, but with a high probability that any malignancies will also be prompted for.

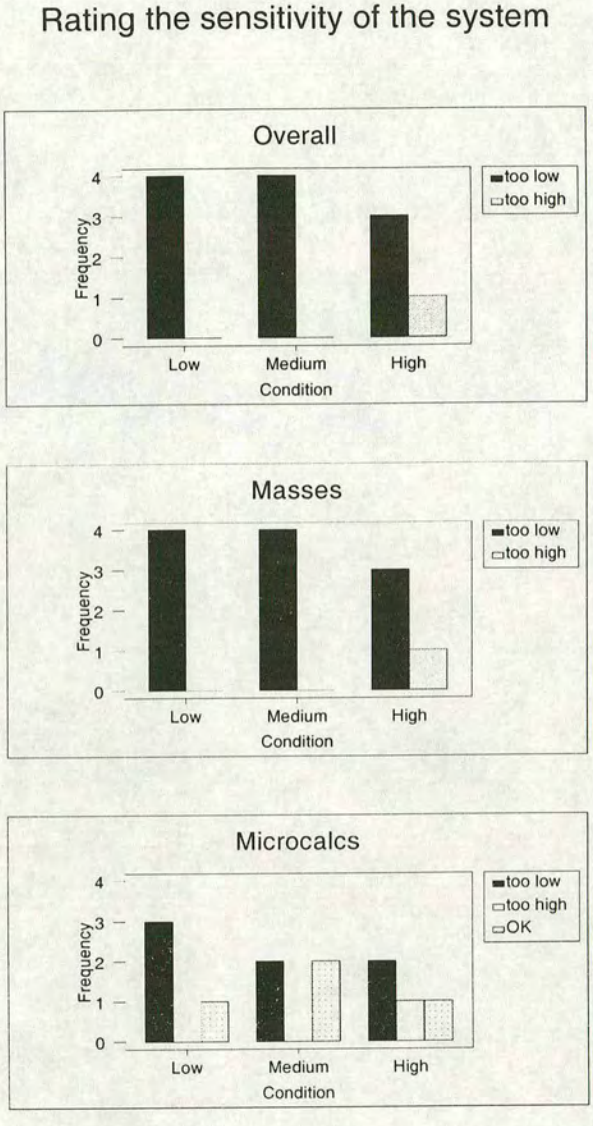


Figure 4.8: Answers to the question: "Do you believe the sensitivity of the system (or the mass and calc components of the system) is Too high, Too low or Just right?"

(Low specificity, high sensitivity.)

2. Low prompt rate, where few of the prompts are for benign features, but with a high probability that some malignancies will be missed by the system. (High specificity, low sensitivity.)

3. A system which is designed to prompt for micro-calcification clusters (whether

malignant or benign) but not other types of calcification (for example, vascular calcification, popcorn calcification). (Calcification clusters only.)

4. A system that will prompt for all types of calcification clusters, rather than one that tries to discard those with benign appearance. (All types of calcification.)

5. A system that will prompt for opacities that can usually be dismissed by radiologists with the aid of previous films or multiple views (for example, composite shadows), as well opacities that are the result of a malignant process. (All types of opacity.)

System configuration	Rating					Average
	1	2	3	4	5	
1. Low specificity, high sensitivity	2	1		1		2
2. High specificity, low sensitivity	1			1	2	3.75
3. Calcification clusters only	3	1				1.25
4. All types of calcification			1	3		3.75
5. All types of opacity	2	2				1.5

Table 4.11: Pre-experiment rating on a five point scale of possible system configurations. A rating of 1 would indicate the configuration is most useful, and a rating of 5 that it is least useful. The greater the average score they less desirable the configuration.

System configuration	Rating					Average
	1	2	3	4	5	
1. Low specificity, high sensitivity	3	1				1.25
2. High specificity, low sensitivity					4	5
3. Calcification clusters only	3	1				1.25
4. All types of calcification		1		2	1	3.75
5. All types of opacity	1	2	1			2

Table 4.12: Post experiment rating on a five point scale of possible system configurations. A rating of 1 would indicate the configuration is most useful, and a rating of 5 that it is least useful. The greater the average score they less desirable the configuration.

Responses to the above questions asked in the pre- and post-experiment questionnaires are summarised in tables 4.11 and 4.12 respectively. Responses to statements 1 and 2 suggest a preference for a system configuration that produces a high sensitivity, and a consolidation of this opinion is evident in the post-session questionnaire. Subjects also rate highly a system that could discard prompts for

non-clustered calcification (statement 3), however, opinion is mixed concerning whether a distinction between benign and malignant clusters should be attempted (statement 4). The responses to statements 3 and 4 bear upon where the division of responsibility between system and users should lie, which is in part embodied by the system's capabilities. Subjects may feel uncomfortable if a system that primarily assists with detection attempts to make a distinction between benignity and malignancy, for example, one possible concern is that malignant clusters could be discarded accidentally. Users of the system might be falsely reassured by the absence of a prompt if they are aware that the system has some level of skill at making this distinction. Furthermore, this mode of action can be viewed as counter to the rationale for detection aids — that a judgement about the significance of a prompted region should lie with the human film reader. In responding this way, a tacit recognition is made that some types of system FPs, if not actually desirable, will be an inevitable consequence of the prefered system configuration. Responses to statement 5 can be interpreted in a similar fashion. Opinion appears to favour a system that does not attempt to rule out candidate lesions by utilising additional information routinely available to film readers, such as previous films.

4.3.4 Subjects' comments

In each post-session questionnaire, subjects were asked five free response questions:

1. What do you think the system's strengths are?
2. What do you think the system's weaknesses are?
3. What irritated you most about the system?
4. What aspects of the system did you find most useful?
5. Can you suggest how the system might be improved?

The responses given are shown in tables 4.14 to 4.18 respectively. When commenting on perceived system weaknesses, three subjects made qualifications about the sorts of things that the system should be able to detect or ignore. In response to the medium sensitivity condition, subject B commented that the ill-defined lesion algorithm was making omissions behind the glandular disk (a 'danger' area). When commenting upon how the system might be improved, subject C also suggests that improved prompting in the 'danger' areas by the

ill-defined lesion algorithm would be desirable. Subject C further comments that the ill-defined lesion algorithm has difficulty with dense breast tissue. These comments suggest a bias in subjects' assessment of the system based on difficulties regularly encountered when film reading. [3] Subject D, when commenting on the high sensitivity condition, draws a distinction between widespread vascular calcification and occasions where only a single site or region is affected, suggesting that only the latter should be prompted. This distinction may refer to a preferred division of responsibility — that obviously benign features should be discarded, but where more detailed consideration is required to dismiss a candidate feature this should be the prerogative of the film reader.

Condition	Number of comments	
	Sensitivity	Specificity
Low	9	1
Medium	6	4
High	5	7
Total	20	13

Table 4.13:

The number of responses made to questions 2, 3 and 5 which can be interpreted as representing a criticism of the system's sensitivity or specificity are shown in Table 4.14. Overall, there appears to be a greater number of criticisms concerning the system's sensitivity, and these apply most frequently to the low sensitivity condition. In contrast, criticisms about specificity appear to be more frequent as the sensitivity of the system improves. Although this result is unsurprising in that it mirrors the system's performance characteristics, it is interesting that subjects feel that they are able to make an accurate judgement about the system's sensitivity. When commenting on what irritated her most about the system, subject B suggested that the system "missed suspicious areas". One strategy for assessing the system's competence, and in particular, its sensitivity, might involve a comparison of the system's response with areas judged to be suspicious.

Comments made by subjects B and D concerning the 'consistency' of the system's responses indicate further how subjects try to make sense of the system's behaviour. Subject B perceives the action of the ill-defined lesion algorithm to contain an element of "apparent randomness", and subject D is concerned that the microcalcification detection algorithm is "inconsistent". [4] Subject D elaborates

[3] These are discussed in detail in chapter 3

[4] Here, and throughout this thesis, the term 'inconsistent' is used to refer to readers' judgements that the system has behaved counter to their expectations of what should and should

by stating that "some benign calcs were prompted, others not", again suggesting that a subjective appraisal of the content of a mammogram is used to inform judgements about system behaviour. Subject D goes as far as to speculate that the perceived inconsistency is "deliberate" — that the system's responses have been intentionally manipulated for experimental purposes. Although subjects appear to be able to make a reasonable judgement about the sensitivity of the system through exposure to prompting information, they occasionally find the system's behaviour confusing and unaccountable.

In response to the question "what aspects of the system did you find most useful", two subjects (A and C) refer to the system's ability to detect asymmetry. However, the system does not make a bilateral comparison, and so is unable to utilise asymmetry information to inform classification judgements about ill-defined lesions. In the briefing given prior to the experiment subjects were not given details about the working of the detection algorithms over and above stating that ill-defined lesions and microcalcification clusters are detected. It is possible that subjects are attempting to understand the system's behaviour in terms of the way they themselves utilise information available from the mammograms, and that sometimes this approach can be misleading.

not be prompted. It is not used in the formal sense to suggest that the system's responses are non-deterministic.

Subject	Condition	Comments
A	Low	Draws attention to certain features. Not too distracting.
	Medium	Draws attention to abnormalities.
	High	Draws attention to asymmetries.
B	Low	Micro-calcification
	Medium	Micro-calcification - tiny clusters.
	High	Small clusters of micro-calcification.
C	Low	Alerting you to areas you may otherwise have overlooked.
	Medium	Prompting review of asymmetries that may be overlooked. Detecting micro-calcification that the reader the reader may miss entirely.
	High	This set of conditions was very useful for masses, only one was not prompted. Distortions are more difficult as one persons distortion is another's normal breast tissue.
D	Low	Quick easy to use. Spots obvious cancers.
	Medium	Simple to understand. Simple to use. Does not make reporting any more time consuming.
	High	As before, easy to use, quick.

Table 4.14: What do you think the system's strengths are?

Subject	Condition	Comments
A	Low	Insensitive (sens)
	Medium	Not specific enough. Too many obviously benign prompts eg vascular calcifications. (spec)
	High	Does not prompt for all suspicious lesions + prompts for many benign lesions eg vascular calcifications. (sens+spec)
B	Low	Soft tissue masses - when not dense were not picked up - especially those behind the breast plate / glandular tissue. (sens)
	Medium	Masses - opacities lying behind the breast plate. (sens)
	High	Vascular calcification distracting. (spec)
C	Low	Not sensitive enough for small clustered micro-calcification and small asymmetries / masses Marked all vascular calcifications unnecessarily. (sens+spec)
	Medium	Some asymmetries and masses not prompted. For me the system should be too sensitive. (sens)
	High	Only a slight increase in sensitivity for masses is required. This set was almost perfect. Masses in dense breast tissue are obviously a problem. (sens)
D	Low	So little help tempted not to bother looking at prompt. Spots only obvious cancers. (sens)
	Medium	Prompting the wrong things currently. (sens+spec)
	High	Picking up vascular calcs. Only appropriate to do so if there are only single area/site affected. (spec)

Table 4.15: What do you think the system's weaknesses are?

Subject	Condition	Comments
A	Low	Did not pick up certain lesions. (sens)
	Medium	—
	High	Too many benign prompts and missed suspicious areas. (sens+spec)
B	Low	Nothing in particular.
	Medium	Apparent randomness of prompting ill defined masses
	High	Too sensitive - masses / vascular calcification. (spec)
C	Low	Nothing.
	Medium	Nothing.
	High	Nothing.
D	Low	No real help at this sensitivity. (sens)
	Medium	Prompting on vascular calcifications. Prompting seems inconsistent to me - some benign calc were prompted others not. (spec)
	High	Inconsistency of prompts (may be deliberate).

Table 4.16: What irritated you most about the system?

Subject	Condition	Comments
A	Low	Draws attention to asymmetries.
	Medium	Asymmetries.
	High	Asymmetry.
B	Low	Micro-calcification.
	Medium	Micro-calcification.
	High	Small clusters of micro-calcification.
C	Low	2nd prompt to areas even if I'd already examined(?) them
	Medium	Micro-calcification detection. Asymmetry review.
	High	Mass and micro-calcification prompts.
D	Low	None in this condition
	Medium	Potentially good double check
	High	Good on spotting calcs

Table 4.17: What aspects of the system did you find most useful?

Subject	Condition	Comments
A	Low	Increase sensitivity. (sens)
	Medium	Improve sensitivity and specificity. (sens+spec)
	High	Needs to improve specificity. (spec)
B	Low	Better detection of soft tissue masses - small. — pectoral muscles. etc. (sens)
	Medium	—
	High	Remove vascular calcification. (spec)
C	Low	Would be useful for ensuring that you missed no small clusters of calcification if sensitive enough for this.
	Medium	Increased sensitivity for opacities and asymmetries, particularly in the areas we miss them ie back of the breast, on pectoral muscle, axilla and inframammary angle. Micro-calcification detection seems about right already. (sens)
	High	Only a small increase in sensitivity required. (sens)
D	Low	Sensitivity+++ (sens)
	Medium	Needs to have the highest possible sensitivity - ie all potential abnormalities prompted but with high "filter" so the benign things I would ignore are not prompted. (sens+spec)
	High	Increase sensitivity to highest possible, most useful for reinforcement. (sens)

Table 4.18: Can you suggest how the system might be improved?

4.3.5 Observation data

The actions of the subjects in the course of reading films were noted for each case according to the following scheme:

e eyeball the film

m use the magnifying glass

p examine the prompt sheet

d mark the decision

The protocol for the experiment instructed subjects to examine the films, examine the prompt sheets, then mark their decision. Table 4.19 shows the frequency of occasions when the subjects either failed completely to examine the prompt sheet (i.e. no '**p**' is recorded in the observation data) and when they marked their decision before examining the prompt sheet (**d** occuring before **p** in the observation data). In the latter case the subjects may have turned the reporting form over after making a decision, then retrieved it realising that they forgot to examine the prompt sheet. Alternatively, if a subject had already decided to recall they may have believed that the prompt sheet would have little further information to offer with respect to their decision, and so only examined it as an afterthought following marking their decision. Taking subject differences into account, there remained a statistically significant variation in the frequency of errors between conditions ($p < 0.0001$ and $p < 0.0111$), with a marked trend for subjects to make an error at the Low, rather than at the High, prompt rate.

This suggests that at lower prompting rates there is insufficient information to hold the radiologists attention either because of the frequency or quality (or both) of the prompts. In response to the low sensitivity condition, subject D stated in the post-experiment questionnaire that the system was of "So little help" that she was "tempted not to bother looking at prompt".

At the lowest prompt rates, on average 19% of cases were prompted for (counting both the mass and microcalcification prompts). Hutt's suggestion that the number of FP prompts should not exceed the number of TP prompts for a prompting system to be effective allows only an overall prompt rate (i.e. including both TP and FP prompts) of 1% [70]. [5] If a more liberal interpretation is made — allowing TP positive prompts to include cases recalled for assessment, then an overall prompt rate of 10%. [6] In both cases the prompt rate would be lower than

[5] Assuming the prevalence of breast cancer is 0.5% and a system that can operate at 100% sensitivity

[6] Assuming a recall rate of 5% and a system that has a 100% sensitivity for recalled features.

that required to sustain subjects' interest during the course of this experiment.

	A		B		C		D		Total	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Low	1	20	6	2	3	1	0	20	10	26
Medium	0	1	0	10	2	0	0	1	2	18
High	0	0	0	5	0	0	1	0	1	8

Table 4.19: Shows the number of occurences in each condition of subjects failing to examine the prompt sheet (1) and subjects marking their decision before examining the prompt sheet (2).

Source	DF	ChiSquare	Pr>Chi
SUBJECT	3	4.9261	0.1773
CONDITION	2	8.9979	0.0111

Table 4.20: Omitting to examine a prompt. Wald statistic for Type 3 analysis

Source	DF	ChiSquare	Pr>Chi
SUBJECT	3	9.8972	0.0195
CONDITION	2	33.7274	0.0001

Table 4.21: Marking decision before examining the prompt. Wald statistic for Type 3 analysis

4.3.6 Further analysis

Figure 4.9 shows the results of the pre/post experiment questionnaire on the perceived value of prompting for particular types of benign feature. Subjects were asked to rate each feature type on a scale of one (useful) to five (distracting). A t-test of the results showed that subjects were significantly more likely to believe that prompting for benign features would be useful after the experiment than they were before it ($p<0.05$).

When making a decision to recall for assessment, subjects were asked to indicate whether the relevant feature had been correctly prompted. Figure 4.10 shows the percentage of correctly prompted recalled cases for each condition against the Likert score for that condition. For the majority of subjects, a monotonically increasing Likert score is apparent as the number of correctly prompted cases in the set increases. These results add weight to the suggestion that subjects' evaluation of the system is informed by a comparison between the system's response and their subjective appraisal of potentially suspicious features.

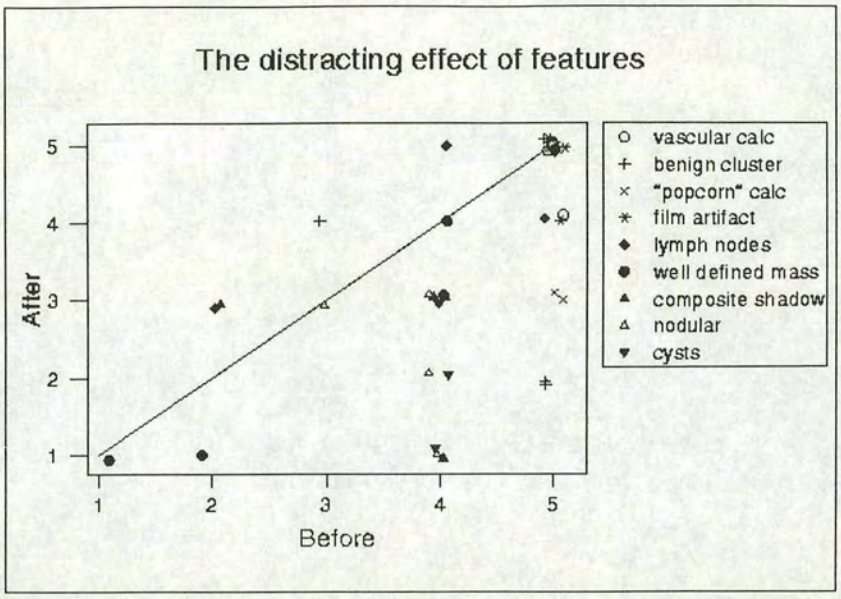


Figure 4.9: Shows the opinion concerning the usefulness of particular types of false prompt (where 1=Useful and 5=distracting) before and after completing the experiment.

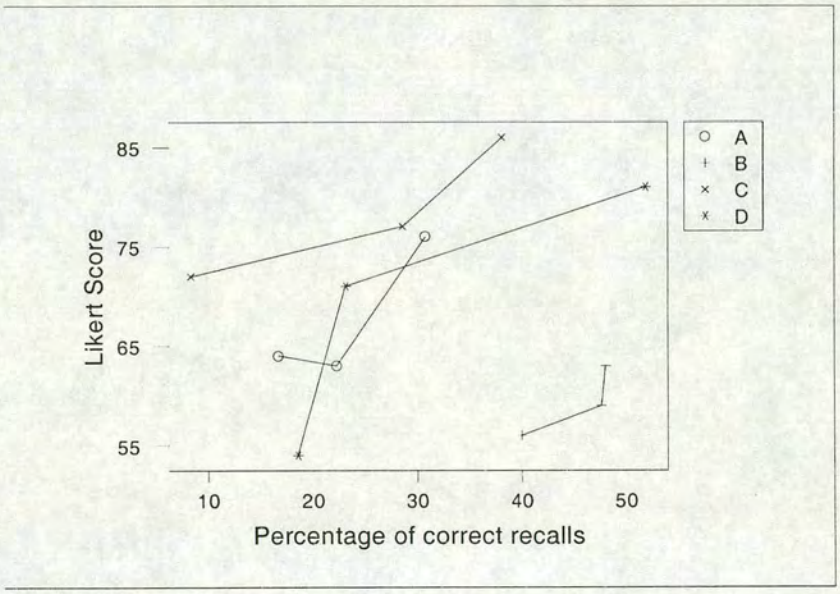


Figure 4.10: Shows percentage of correctly prompted recalls against Likert score for each subject

4.4 Discussion

None of the prompting regimes had a significant effect on the number of cases recalled for further assessment, and only a small effect on the time taken to

complete the conditions. Observation data revealed that below a certain prompt rate subjects were less inclined to examine the prompt sheets. Also, subjects' perceived tolerance for FP prompts was greater at the end of the experiment than at the beginning.

It is difficult to draw conclusions concerning an acceptable prompt rate for the ill-defined lesion detection algorithm as its performance was poor (only 63% sensitivity at its highest setting). Subjects were able to perceive this difficulty and were far more concerned about its sensitivity than its overall prompt rate. By consensus, the prefered prompt rate from the microcalcification algorithm was at its medium sensitivity (86%) where approximately 1 in 6 cases were prompted. However, subjects were unhappy with the number of vascular calcifications prompted for in the high sensitivity condition, and a higher prompt rate might be acceptable if prompts for this type of feature were removed.

Overall, these results suggest that when tested under realistic conditions, film readers' tolerance level for FP prompts is appreciably higher than the upper limit established by Hutt for improved detection performance. Of course, positive subjective assessment may not necessarily coincide with objective performance effects, but it can be argued that earlier work may have underestimated the FP upper limit.

As there were so few true malignancies in the test sets, subjects were not expected to be able to form an accurate picture of the system's detection performance. However, comments made both during and after the experiment showed that their assessment of the system's sensitivity was actually very acute. Figure 4.10 suggests that this judgement was informed by the proportion of recalled cases that were correctly prompted. It can be argued that subjects' tolerance of FP prompts was due to the fact that they were informative of the system's performance characteristics.

It is possible that the effect of FP prompts will depend on their nature. When reading, radiologists consider a number of *candidate* features for recall, but only a proportion of these features result in recall, and only about 10% of recalled cases actually turn out to be cancers. It may be that prompts for *candidate* features would be acceptable to film reader's in a clinical setting, whereas prompts for *other* features would not. The latter would be distracting, and contribute to the degradation in performance found in earlier work. In contrast, the former affords learning about — and positive confirmation of — the system's behaviour. It is probable that this will be important for effective routine clinical use of such a system. In support of this, a comparison can be drawn with the practice of

'annotation' described in Chapter 3, where often benign but notable features are drawn to the attention of a second reader. A prompt might be viewed not only as an explicit directive to examine a region a the film, but also as an implicit source of information about the system's capabilities. However, this information is sometimes either incomplete, or misinterpreted — despite making accurate judgements about sensitivity, subjects mistook PROMAM's operational scope and described its behaviour as 'inconsistent'.

This investigation raises a number of issues concerning the means by which subjects inferred properties of the system from its behaviour. In particular, how subjects were sometimes mislead into adopting an inaccurate model of the system's operation and what additional information would be required to ensure a more accurate interpretation. A follow-up exercise was devised to examine these issues and to explore further subject's apparent tolerance to FP prompts.

Chapter 5

A detailed analysis of prompted cases

5.1 Introduction

Results from the experiment described in the previous chapter suggest that readers are able to make accurate judgements about the performance of the PROMAM system, but are less accurate in their determination of the scope of the system's abilities. In order to examine in greater detail readers' interpretation of prompts a 'think aloud' protocol was used to examine subjects' reasoning for a large number of prompted cases.

5.2 Protocol

Prompted cases from sets 1 and 2 used in the 'subjective responses to prompting' experiment were employed in this exercise. 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 as subjects, none of whom had participated in the prior experiment. 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 perform the following actions for each of the prompted features:

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. State whether they would recommend recall on the basis of the prompted feature (yes/no).

4. Classify each prompted feature (free text response).

5. 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 subsequently transcribed.

The transcripts were initially examined for occasions where subjects had been confused by, or had misinterpreted, the prompts. The opinion of system developers was sought concerning the action of the system in those cases. The transcripts were then examined for other recurrent themes. The data was reorganised accordingly, and then re-examined to similarly structure the uncovered themes. Finally, a conceptual basis was sought that would account for the emerging framework, and relate it to the findings of the previous experiment, the work practice study, and literature concerning radiological expertise.

In the discussion of the verbal protocol, extracts from the transcript are labelled according to the subject (A, B or C), the session (1 or 2) and the case examined (1-74 or 1-70). Thus the label (A-1.23) identifies the extract as belonging to subject A reading case 23 in her first session.

5.3 Setting

The exercise took place in the reading room of the Glasgow Caulder Street screening centre using film viewers regularly used for screening work. The reading room contains three similar RadX type viewers — usually one is dedicated to incident round screens, and the others to prevalent round screens. Space is limited; there is barely enough room for three film readers to be reading at the same time. During the exercise there were several interruptions. These were either questions directed

at subjects by other members of staff, or requests to access artifacts within the reading room. Cases were examined under usual reading conditions — the room was darkened, and care was taken to arrange the films on the viewer in a manner typical for that screening centre.

However, the tasks of reading, and of system use, differed from normal practice in significant ways. In common with the 'subjective responses to prompting experiment' only oblique view mammograms were made available to the subjects (usually either CC views or previous films would also be available during routine screening). The most likely effect of this would be on subjects' specificity. There may also be an impact on their assessment of prompting information — it is possible that a feature that appears suspicious on one view can be seen to be clearly benign on another. In these circumstances, a reader's tolerance to a prompt may be affected.

Subjects were shown only prompted cases, and were therefore exposed to much higher prompt rates than would be the case for a randomly chosen set of films. This may affect their judgement about the usefulness of the system in terms of its specificity. Furthermore, there is no opportunity for subjects to respond to unprompted cases, and thus no data available about how they view system FN decisions where no FP decisions are also made.

This exercise was presented to subjects as one of rating and talking about the system's response, rather than as one of interpreting mammograms. No instructions were given to subjects concerning how they should organise their examination of the mammogram and prompt form. It is significant that early in their first sessions, subjects A and C both adopted a policy of examining the mammogram before referring to the prompt form:

> "Moving on to number three. What I'll do is I'll look at the films first - and then I'll look at the prompts." (A-1.3)
>
> "...look at the films before I get a chance to eye the prompt." (C-1.6)

They maintained this approach throughout the remainder of the exercise.

5.4 Rating data

Table 5.1 suggests considerable inter-observer variation between subjects with respect to the acceptability of prompts. Subject B's rating of responses made by the microcalcification detection algorithm appears anomalous, but can be explained by the positive rating she gave to vascular calcifications. The majority of

the FP produced by the microcalcification detection algorithm were for vascular calcifications — 135 of the 188 calcification prompts were classified as vascular by one or more of the subjects. While subjects B and C were inclined to rate vascular calcifications as 'unacceptable in a screening environment', subject B took an opposing view:

Algorithm	Subject		
	A	B	C
Mass	27.7%	44.8%	70.3%
Calc	8.5%	91.5%	18%

Table 5.1: Shows the percentage of prompts judged to be acceptable in a screening environment for each algorithm.

> "[...] That's arterial calcification. Definitely isn't anything else. As I say I don't know quite if we're trying to get the computer to pick up micro-calc, I don't know how you can get round that. You can't, you'll just have to prompt it and you can dismiss it. So in an ideal situation it would be able to tell this is what it was, but I don't know that that can be helped. It's distracting to be honest." (B-1.27)

Table 5.2 shows that although subject B was inclined to rate vascular calcification prompts as 'acceptable', she shared the view of subjects A and C in suggesting that they are more likely to be distracting than valuable. As subject B's commentary on case 1.27 suggests, she is drawing a distinction between acceptable, and valuable prompts. The system's inappropriate response to vascular calcifications is mitigated because distinguishing vascular calcifications is perceived to be beyond the capabilities of any potential prompting system. This view is not, however, a technically informed one — a system that could distinguish vascular calcifications is plausible. Subject B is stoical — she appears to be suggesting that this is an unfortunate aspect of system behaviour that would 'just have to be lived with'. However, this view might not be sustained with persistent exposure to a system demonstrating this weakness. Subjects' views concerning what are seen as 'justifiable limitations' to the system's performance are discussed in more detail later.

Assuming that the rating scale is being used in a similar way, then subjects B and C appear to demonstrate a high degree of tolerance to what are effectively false positive prompts. Although subject A's tolerance of prompts appears to be lower, she specified a greater number of unprompted regions that, in her view,

Subject	Rating				
	1	2	3	4	5
A	1	11	5	145	26
B	13	8	2	0	165
C	3	15	13	9	148

Table 5.2: Ratings given to prompts produced by the microcalcification detection algorithm. (1=valuable, 5=distracting)

Subject	Rating				
	1	2	3	4	5
A	3	28	11	104	9
B	38	19	2	1	95
C	20	38	39	37	21

Table 5.3: Ratings given to prompts produced by the ill-defined lesion detection algorithm. (1=valuable, 5=distracting)

Subject	Rating				
	1	2	3	4	5
A	0	76	3	1	0
B	12	5	0	0	0
C	6	8	0	0	0

Table 5.4: Ratings given to additional features annotated by subjects. These are features in the mammogram that had not been prompted by the system, but which subjects thought worthy of a prompt. (1=valuable, 5=distracting)

should have been prompted, compared with subjects B and C (Table 5.4). So it seems that subject A is not necessarily intolerant of false positive prompts per se, but she is less tolerant of the sorts of false positives characteristically produced by this system.

A notable feature of the verbal protocol is subjects' commentary on the 'reasonableness' or otherwise of the system's behaviour. While a simple relationship between the perceived significance (in terms of suspicion) of a feature and the appropriateness of prompting that feature might be expected (Figure 5.1), it appears that this is not the sole criteria for judging the reasonableness of the presence or absence of a prompt. In fact, criteria for reasonableness appear to be highly contingent and interdependent — the value of a false positive prompt is seldom judged entirely on the characteristics of the prompted feature alone. Additional criteria include the location of the feature prompted within the breast, tissue type, the perceived effect of the prompt on the film reader and the number of other prompts generated on the same mammogram.

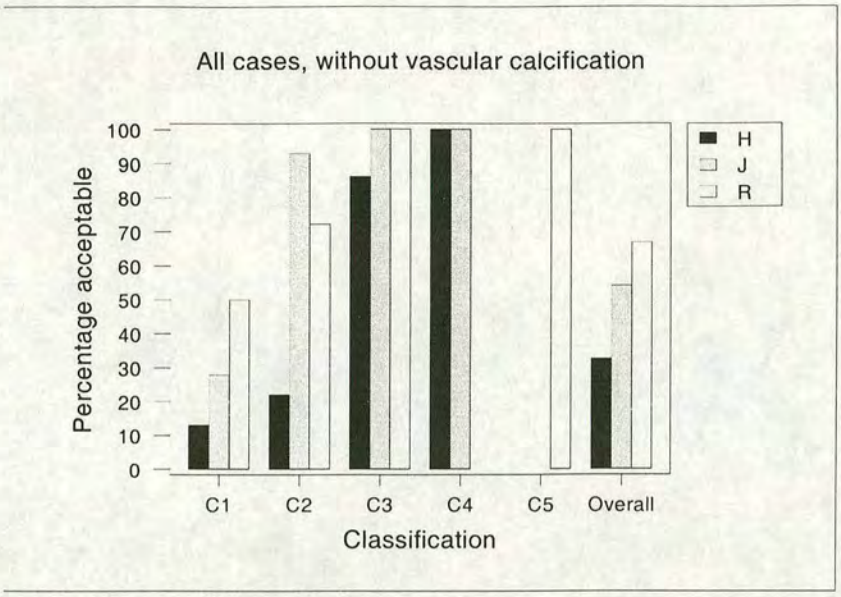


Figure 5.1: Shows the percentage of acceptable prompts against rating of suspicion with prompts judged to be for vascular calcifications by one or more subjects excluded.

The remainder of this chapter is concerned with the data gathered from the verbal protocol where the criteria for the reasonableness of prompts is explored in detail. This is done in the context of how subjects made use of the system and thereby defined its role in supporting the task of reading, and how they used system responses as evidence for determining both the scope of its abilities and its performance.

5.5 The role of the system

The system's 'role' refers to what the system is 'for', that is, what activities it is designed to support. Notionally a prompting system is designed to draw readers' attention to potentially significant features within the mammogram and in doing so to reduce the possibility of FN decisions due to cancers being overlooked. Thus the system is designed to assist film readers to make a more thorough or complete examination of the mammograms. During this exercise, there were occasions where the system was acknowledged by subjects as having fulfilled this role. However, subjects suggested additional roles that the system might have in supporting the reading process including: orientating readers to their task of interpreting the mammogram and providing evidence for classification decisions.

5.5.1 Accountability

Mammograms are information rich artifacts which are examined by trained observers for abnormalities that sometimes have a very subtle presentation. However, the majority of mammograms are normal, and the majority of radiological presentations that a film reader will encounter are likely to be benign. The number of screening tests performed is large, but the resources available within the screening programme to deal with this workload are constrained in a number of different ways. There are only a limited number of film readers, who are only able to invest a limited amount of time in the task of reading. [1] The cognitive resources available to individual readers are also bounded — attention is a limited resource, and human observers are prone to fatigue. [2]

It would be possible for human observers to approach the task of reading mammograms by exhaustively examining and analysing each part of the image. However, studies have shown that visual search is often incomplete [104] and that experienced readers are able to quickly 'zero in' on significant features [106]. The high information content of mammograms coupled with resource constraints effectively preclude an exhaustive search, instead attention is given to mammograms in a way that is dependent on their content (see Chapter 2).

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. Thus it is not necessary to attend equally to every mammogram if an accurate determination of how much attention is required can be made. Similarly, there can be variation in the degree of difficulty associated with the interpretation of individual features within a given mammogram. Some features may be obviously benign or malignant, others may be uncertainly so, either because 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 trained human observers to reading mammograms involves selectivity in the application of effort to produce some reasonable level of performance under particular resource constraints. Selectivity is mediated by

[1] A survey of film reading practice in the UKBSP programme reported by Wells and Cooke gives an indication of the resourcing difficulties faced by some screening centres [129].

[2] For a review on limits to attention see Kahneman [74]. In a study reported by Cowley it was found that readers' performance starts to decline after 70 or 80 cases have been examined without a break. Also, that performance is improved if more time is spent examining individual cases [30].

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'. Readers can be thought of as satisficing — there is too much information in a mammogram for all of it to receive an equal and detailed analysis, a more strategic approach is employed to render the reading task tractable.

However, there are pitfalls inherent in the strategy of selectively allocating cognitive resources that arise precisely because of the implication that not all of the mammogram is examined in the same way, and to the same level of detail. Firstly, not everything in the mammogram will be 'noticed' by the reader, making it possible for significant features to be overlooked. Secondly, noticed features are analysed by employing strategies that are commensurate with an initial assessment of their characteristics, and effort expended will be proportional to some experientially derived notion of 'worth'. Thus it is possible to misapply analytical strategies, or to too quickly dismiss a feature as benign. The degree to which performance is sacrificed by this approach will depend on the generality of the heuristics used, and the accuracy achieved in appropriately matching strategies to circumstances.

The examination of a mammogram can be thought of as a process of eliminating potential interpretations. When the initial examination is made, the set of potential interpretations is large as it includes all of the interpretations that it is possible for a film reader to consider. As the examination proceeds, the number of potential interpretations is rapidly reduced to include only those that are likely to apply given the appearance of the mammogram. It is possible for errors to occur if the number of actively considered interpretations is narrowed to quickly. [3]

If mammograms are not examined exhaustively, then it becomes important that mammograms and features that are worthy of attention do actually receive the attention they deserve, and that an appropriate analytical strategy is deployed in their interpretation. 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. [4] Readers also have to account for features in

[3]For example, see the discussion of 'Satisfaction of Search' in Chapter 2

[4]The notion of accountability outlined is similar to Garfinkel's – particularly in his discussion

a particular way — that is, according to the most appropriate strategy for analysing a given type of presentation. [5] Accountability to the process of interpreting a mammogram is bound by what an experienced reader might reasonably be expected to notice, the lengths to which they might be reasonably expected to go to establish the status of some noticed feature, and also by what analytic strategies it might be most reasonable to select given the type of presentation. Accountability demonstrates an approach to the fallibility of satisficing by prompting a continual series of reflections about courses of actions available and the certainty of any conclusions. The end result of this process includes both a decision and also a justification, or rationale for that decision.

On several occasions subjects appeared to be using the system 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." (B-1.42)
>
> "Ok, I thought it would get that bit. Yeah that's fair enough, I think it's not unreasonable to prompt you to have a look at that, to have a proper look at it." (C-2.32)
>
> "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." (A-2.33)

In each of these cases the prompted feature was judged to be benign. 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. Here the prompt is being used as an aid to their approach of selective examination by suggesting features in the breast that might be worthy of a greater level of investigation. The prompt is not necessarily bringing features to subject's attention that had gone unnoticed, instead the prompt is seen as suggesting that there is something about the feature that needs to be accounted for. Conversely, subjects express dissatisfaction with the system if the prompts are for regions that are

---

of decision-making in the Los Angeles Suicide Prevention Centre. Here 'decisions' are openly acknowledged to fall short, but at the same time they are recognised as being "adequate for all practical purposes". What is made observable by members accounts is the "rational adequacy" of their decisions — members' accounting attends implicitly to the pragmatics and contigency inherent in each and every investigation [56].

[5] For examples of informal means by which readers seek to maintain accountability see the discussion about the practice of annotation in Chapter 3

less significant, especially when there is a region of greater significance within the image that has remained unprompted.

In the examples given above, prompts for benign features are tolerated, and are even found to be desirable. This perceived utility is not due to any influence on the final decision — the prompts have neither drawn something to the subject's attention that had been missed, nor have they altered the subject's interpretation of the prompted feature. Subjects suggest that the system has proved useful in these cases because they have been encouraged to take a 'proper look', or to 'look more closely' at some feature in the mammogram. In doing so, it is possible that the prompting has a psychological benefit by reducing the anxiety a reader may have that a complete and thorough examination has been made. Use of a prompting system may also improve readers' capacity for self awareness and reflection. When confronted by a prompt it is natural for a reader to reflect their own interpretation — whether they saw the feature, how much attention they gave to it, and the interpretation they reached. By using the system readers can be made more aware of their own thoroughness. Use of the system may give a perception that mammograms are being examined with a greater discipline, thus appealing to readers' conscientious approach. Although the responsibility for the final decision still rests with the film reader, in some senses use of a prompting system implies that a reader's charge of the task of reading is reduced. The system can be thought of as a work partner, or collaborator, with a suggestion-making, or consultative role, that is able to share the responsibility for achieving a thorough investigation. One benefit of promoting reflection might be to prevent the reader from too hastily narrowing the possible interpretations that might be attributed to the mammogram.

Use of a prompting system can be seen as engendering a more systematic approach to the task of reading by providing a 'checklist' of features to be examined. As a consequence, a reader has less discretion over what in the image is attended and in what detail, entailing a narrowing of the reader's autonomy. The corollary of this is that the prompt itself is an accountable phenomena — the reader is bound to seek an explanation for the presence of a prompt.

5.5.2 Context

As stated earlier, subjects' views on the reasonableness of prompts appears to be highly contingent and dependent on interrelated factors. The suggestion that the interpretation of a mammogram by a human observer involves satisficing can offer a partial explanation for this observation. In addition to the importance of

a feature's character as an indicator of suspicion, the context of its presentation, or its situatedness, also plays a role in determining how much effort should be invested in its investigation. Prompts may be judged reasonable because they attend to contextual considerations, sometimes to the extent which they may be judged reasonable even where the feature prompted has little or no significance. In the following sections, the relationship between subjects' responses to prompts and their context is explored.

5.5.2.1 Region within breast

One of the ways readers' orientate themselves to the task of reading 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." (A-1.25)

Subjects occasionally judged the reasonableness of prompts against the criteria of location; Subject A was particularly keen for the system to prompt for features in the review areas, for example:

> "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." (A-1.3)

> "What, if you're going to pick a tiny area - why did it not pick that out? Because that's sitting on the back of the breast. I wouldn't necessarily have recall it - but I think it would be valuable. Mainly because of the site of it." (A-1.37)

In case 1.3, subject A 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. In case 1.37 subject A contrasts what the system prompted with her own view about what is significant in the mammogram. She implies that the prompted region and the region that she highlights differ in significance only because of their respective locations.

"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." (C-2.7)

Here subject C 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 an area to which they may not be inclined attend so readily. The region is in front of the lower portion of the pectoral muscle, so subject C's contention that the feature is 'at the edge' of his 'field of vision' appears counterintuitive (given the way the mammograms are arranged, this region is towards the centre of the reader's field of view). However, he is probably identifying the centre of each breast as the centre of his area of interest. One might draw a map of a mammogram and label it according to some notion of which regions a reader would consider worthwhile examining. All off breast areas (from the edge of the breast to the edge of the film) would receive a low score. Feature rich regions, such as the glandular disk would receive a medium score, and specific features that have some suspicious characteristics would be given a high score. The region identified by subject C as being at the edge of his field of vision is towards the edge of any interesting regions. It is also possible to imagine readers' particular attention to the review areas as re-labelling what are often feature poor areas as being of greater interest than readers' natural inclinations might suggest.

Subject C's interpretation of the system's response in case 2.7 is particularly interesting because although the prompt is for microcalcifications, it is entirely clear that there are no microcalcifications present (Figure 5.2). 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.

In contrast, subject B judges a prompt to be unreasonable because of its location:

"I wouldn't bring the woman back for that, I don't think that it's anything at all. It's either artifact, or at the most vascular. I don't think there would be any breast tissue as such down there, I don't think it helpful." (B-1.32)

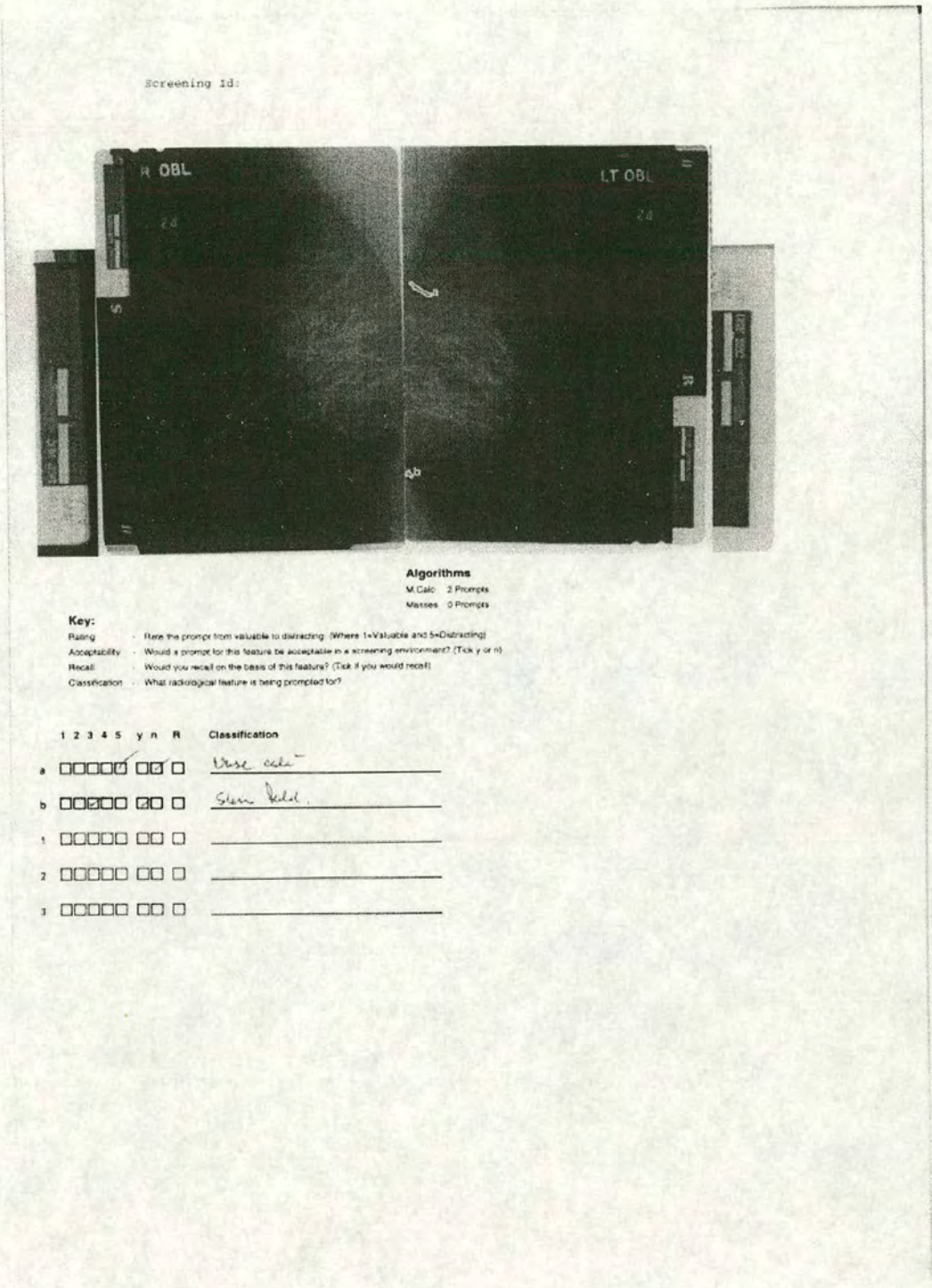


Figure 5.2: Case 2.7 with annotations by subject C. Subject B is referring to the prompt *b as being useful because of its location.*

Here the prompt is judged to be 'unhelpful' partly because it is in a region where tumour development is unlikely due to an absence of breast tissue. Thus the prompt is seen as irrelevant since it draws attention to a region of the mammogram that has little intrinsic interest for the subject. A parallel might be made with

subjects' responses to prompts that lie outside the breast, including occasional FP for film numbers and other markings.

5.5.2.2 Tissue type

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

> "These are very dense breasts, very difficult, and time consuming." (A-1.22)
>
> "Again kind of DY type breasts — much more difficult to interpret." (A-2.25)
>
> "Is back on to my nightmare... These are a nightmare, 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." (B-2.23)

Readers state that the task of examining dense, or 'glandular' breasts is 'difficult' and 'more time consuming', suggesting an orientation to the task of interpretation, perhaps involving a selection of strategies different to those invoked when interpreting lucent breasts. Dense breast tissue complicates the task of interpretation and readers are aware that their judgements may be less reliable — subject B in particular, appears to express frustration that her ability to detect cancers may be compromised. 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 women with lucent breasts.

Subject B also commented that a prompting system might have a useful role in addressing this difficulty:

> "These are very difficult mammograms because they are very difficult breasts. These are the ones where we are open for some help." (B-1.22)
>
> "The problem we have is that we are looking for something, we're really looking for something... it's mission impossible... we're looking for help in the really difficult ones where you do look at these things.

In a way it's a help because it makes you look again and try and..." (B-1.23)

"This is a difficult set of mammograms because they are very glandular — these are the ones that we need help with, and A has been annotated on the right round a blob — but it's not any different from any of the other multiple blobs on this patient's breast." (B-1.42)

"Well, this one's a nightmare, and these are the ones where it is helpful, you hope, to have things prompted. Because it's so glandular, and so dense, but however — we'll just see." (B-2.18)

"I mean if your prompting system is really good in an N1 breast, but isn't very good in a dysplastic type breast. I think that's important to know, and it might be important to run a few experiments to see if there's any... 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." (A-2.23)

Subject B is identifying situations where some assistance with reading could be of greatest benefit. She is, not unreasonably, suggesting a role for a prompting system in providing support in circumstances where the reading task is found to be most difficult. Here, the psychological benefit of reducing anxiety may be greatest.

In case 1.23 above, subject B suggests that the system may helpful in these cases because it 'makes you look again'. The system is perceived as helpful in ensuring that the difficult case is given the necessary scrutiny that its degree of difficulty demands. It is suggested that the prompts might be helpful in this situation because of this effect they have on the reader, rather than because what is prompted is significant. Again, the utility of the prompts can be explained in terms of improving readers' accountability, or in providing reassurance that the examination of the mammogram has been thorough.

In case 2.23, subject A is hopeful that the prompting system would not be prone to difficulties experienced by human observers when interpreting dense breasts. She expresses a desire for dependable assistance — particularly where the reading task is most difficult, and is implicitly suggesting such provision could reduce variation in reader performance. However, this may have implications for how prompting information is interpreted. If a prompting system is more able (than a film reader) to distinguish between benign and malignant presentations

under particular circumstances, then the film reader cannot rely on her own judgement of the mammogram to distinguish between TP and FP prompts. Instead, a judgement would have to be made on the track record of the prompting system.

Conversely, prompts might be viewed as being less reasonable if ascertaining normality is perceived to be relatively easy:

> "I know, it's the same sort of bit here. I'm sure that's what it's gone for, I don't think that's anything at all, I really don't. I would be very (...?) if it is vascular, ach it probably is vascular. [break] ...on a nice set of lucent breasts like this, which is what everyone should have before they are even allowed into breast screening program. This is normal, sorry — distracting." (B-2.28)

It may be that the utility of a prompting system 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.

When examining the prompt sheet for for case 1.9, subject A commented:

> "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." (A-1.9)

In case 1.9, subject A 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. Furthermore, the prevalence of a particular feature type may be indicative of some underlying process or susceptibility. For example, partially involuted breasts may be the cause of a 'patchy' appearance, or a susceptibility to cysts may result in the presence of multiple densities. Thus, the very presence of similar, multiple presentations, can sometimes suggest a way of accounting for

each of instance of a feature. There is the attendant danger of 'premature closure', i.e. precluding the consideration of a broader set of alternative explanations.

Subject A 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 re-evaluating the prompted features she is perhaps widening her analysis to include a broader spectrum of interpretations.

This discussion suggests that the significance attributed to a particular feature can depend on the context of its presentation. When examining case 2.42, both subjects B and C refer to a feature's situation when explaining the reasoning behind their decision:

> "...but A is a very reasonable prompt, a possible mass. [break] ...suspicious on these films." (B-2.42)

> "Nothing abnormal on the right, on the right side. The left side there's a stellate distortion in the lower part of the breast which is almost certainly composite, and there's a bit of asymmetry in the upper part of the left breast. Neither of them look particularly worrying, but perhaps if you saw the both of those on the one film you would just want to have another view of it." (C-2.42)

Subject B makes a direct reference to situatedness of the feature within the broader context of the general appearance of the mammogram. Perhaps the feature's appearance is unaccountable given the types of normal structure that she might expect to find in a mammogram with that general appearance.

Subject C identifies two potentially suspicious features, but does not believe each individually to be overly suspicious. However, his inclination to recall the case is increased precisely because they occur together in the same mammogram. His reasoning is concerned with probabilities:

> Observer - "Can it add to your suspicion then, where there's more than one..."

> C - "No, no, no, not really. Just that when they're... they don't make it additionally suspicious as such, it's just that if there are two bits, two areas, you know. Well there's double the chance that one of them might be something." (C-2.42)

In the following extract, subject B 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 B'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.

5.5.3 Accounting for prompts

For a system to be useful as a reading aid, readers have to be confident that the system is capable of detecting cancers, and that the system's specificity is sufficient to make its use worthwhile. If readers believe that the system can be useful, this implies that readers should attend carefully to its analysis.

"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." (B-1.40)

"Breast tissue. Just trying to see if there is a mass in here — but I cannot see a mass. Breast tissue." (B-2.63)

In the above extracts, subject B 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 she did not initially apprehend. In case 1.40, subject B does finally identify a characteristic of the mammogram as a possible reason for that region being prompted. In case 2.63, subject B re-affirms that the prompted region is nothing more than breast tissue.

"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, but if that wasn't there, then I wouldn't be impressed by that, so I wouldn't recall it for that. I don't think it has been helpful." (B-1.61)

"I don't what it's been getting at with c, I've got to say. I would be bringing her back anyway, so I suppose it's whether there's another opacity there, but... Probably benign I think. It's reasonable enough though, and I don't see anything else." (B-1.71)

Subject B suggests that her decision to recall the case would provide an opportunity to gather further information about the prompted feature. She is entertaining the hypothesis that the prompted feature may be significant, although she doubts this. She is maintaining 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 would further her understanding of the abilities of the system.

If a reason for a prompt is not readily apparent, then this can pose problems for readers who 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 quote demonstrates how the interpretation of false positive prompts for 'subtle' features can be problematic:

"I don't see anything that's worrying, and what's been prompted is... I'm not sure what's been prompted - possibly that - that's quite distracting because I'm saying to myself why have you prompted that? And I don't see anything else to worry about." (A-1.42)

"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." (A-1.65)

The prompt in case 1.65 presents subject A 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 film 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.

5.5.4 Influencing interpretation

The PROMAM system is designed to be a detection aid — its role is to ensure that features with malignant characteristics are not overlooked by a human observer. It is intended that the presence of a prompt should 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, rests with the film reader. Acting on a prompt by examining the prompted region is the least response that might be expected if the prompt is produced by a credible system. The greatest response would be to recall every prompted case. There is also an intermediate response — that the prompt is used as contributory evidence for making a recall decision in some cases.

In some senses it might be expected that a prompt would contribute to a reader's suspicion of a feature precisely because the presence of a prompt implies the possibility of cancer. As possible indicators of suspicion, prompts may be used by readers as evidence of suspicion when making classification decisions.

There were a number of occasions where subjects reported that their interpretation of a feature was affected by the presence of a prompt:

> "Big prompt, taking up most of the left breast. Just looking at it from the distance first of all. There's a very very vague suggestion of distortion. And it's only because of the prompt that I am looking at it. I think it's reasonable though, maybe there would be something on the CC. I think it has been useful, and in fact I would recall it. And I'm not sure what it is, so give it a C3." (B-1.6)

In case 1.6 the prompt covers a wide area of the breast. Subject B responds to the prompt by paying particular attention to the prompted region, and in doing so

she identifies, almost construes, a mildly suspicious feature. She further suggests that it is 'only because of the prompt' that this feature is being considered. Although the evidence from the image itself is weak, her suspicion is sufficiently aroused to recommend recall. In attempting to account for what superficially appears to be an uninteresting prompt, subject B makes the working assumption that there is some good reason for the presence of a prompt. In fact, subject B is mistaken in her interpretation of the system's response. The system has identified the whole of the prompted region as a large, potential ill-defined lesion — the prompt does not relate to the area of distortion identified by subject B. In searching for a possible, and rational, explanation for the prompt, subject B misconstrues the scope of the system's abilities. The system is not able to detect distortions, and therefore its responses cannot be used as evidence for the presence, or significance of distortion.

In the following cases the presence of a prompt appears to add to subject B's degree of suspicion about a feature:

> "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." (B-1.36)

> "It's probably composite. I mean, if I got prompted to that I would actually recall it because of the prompt. [break] [break] ...not anything. I would recall it." (B-2.40)

Common to subject B'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 C also reports heightened suspicion due to the presence of a prompt:

> "(...?) That 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." (C-2.40)

The features considered for recall by subjects B and C 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 demonstrating a minor degree of suspicion would overwhelm resources available for assessment clinics. In case 2.40, subject C 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 A suggests that the lack of a prompt can be significant:

> A - "It hasn't really picked up on the asymmetries but they're not worrying in any way..."
>
> Observer - "So would you rather it..."
>
> A - "...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." (A-1.33)

Here the lack of a prompt is seen as 'comforting', precisely because subject A equates the lack of a prompt as indicating that the system has assessed a region and found it to be benign. This type of reasoning is a potential cause for concern for three reasons. Firstly, the ill-defined lesion detection algorithm has both a 'feature extraction' and a 'feature classification step'. If a feature is unprompted, then this may be because the feature extraction step has failed, and if this is the case it cannot be said that the system has performed a 'complete' analysis of the feature. Secondly, the system does not target asymmetries for detection, so it is not reasonable to assume that the lack of a prompt is a good indication that a region of asymmetry is benign. Thirdly, the system does make false negative decisions, so the absence of a prompt is not a foolproof indicator of benignity.

In making the judgement that it would be inappropriate to prompt for these particular asymmetries, subject A is trying to define an appropriate confidence threshold for system responses.

Subject C suggests that the system might have a direct role to play in influencing film readers' confidence thresholds for asymmetry:

> "So it is not a bad thing that it is brought to our attention, perhaps we are no naturally, particularly at the right the threshold for detecting

asymmetry. So I don't know, maybe the prompts are right and we're wrong. Certainly quite a number of the missed, or false negatives are in asymmetrical breast tissue. However, I don't think that would come into that category." (C-1.40)

Again, this demonstrates an orientation to perceived weaknesses in film reader performance resulting in a suggestion that the system may be able to assist in this respect. If the system were to influence a reader's decision-making in the way that subject C suggests, then the system would be fulfilling a very different role to its intended one. Again, readers would have to sacrifice their role as arbiters of whether a prompt is a true positive or a true negative. They would be dependent on their experience of, or claims made about the systems capabilities, rather than on their own interpretative abilities.

5.6 The system's functional scope

It is not intended that a prompting system should duplicate film reading expertise in its entirety. The extent to which the behaviour of the system can be considered to be similar to that of human observers is that the system also 'notices' features of potential significance in the mammogram. However, even in 'doing noticing' the system's capabilities are relatively constrained. The system is only designed to notice particular types of feature, namely microcalcifications and ill-defined lesions — and not the complete gamut of presentations that can be indicative of cancer. Furthermore, the system can be thought of as a naive observer, in that it lacks the interpretative sophistication of a human film reader.

The verbal protocol reveals that subjects actively engage in making sense of the system's behaviour. This sense-making appears both to involve an assessment of the system's capabilities — what it might reliably detect, and what might be overlooked — and also finding a rationale for the system's responses — how the system might be expected to behave under different circumstances. The payoff for investing effort in learning about the system is that presumably this enables more efficient and effective use of the information it supplies. If readers can understand the reasons for a prompt, this would enable a quick decision to be made about the action it demands, thus lowering the cognitive burden of system use. Furthermore, if readers can understand the diagnostic implications implied by the presence (or absence) of a prompt, they can be in a position to make the most appropriate decision. In this latter case, readers are trying to understand what the prompt 'means', and how it should be 'interpreted'.

Strategies in evidence for making sense of the system include:

1. Comparing the system's response for similar types of feature.
2. Assuming purposeful behaviour.
3. What might be indicated by the prompts shape, size and location.
4. Explanations external to the operation of the system.

Each of these approaches are discussed in turn below. A further sort of sense-making concerns an orientation to the performance of the system (its sensitivity, for example) and involves subjects comparing their notions of significance with those exhibited by the system. This is discussed fully in Section 5.7.

5.6.1 Prompts for similar types of feature

Readers made comparisons between the system's responses to similar types of feature, either in the same breast, or between cases. This activity is frequently accompanied by readers puzzling over why one instance of a feature type has been prompted, but not another.

Subject's often found the tendency of the microcalcification detection algorithm not to prompt all benign microcalcifications to be a source of confusion, for example:

> "[...] It's interesting it's prompted the vascular calcification on the one side and not the other. So that gives me - I'm thinking the whole things inconsistent you know." (A-1.20)
>
> "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." (B-1.39)
>
> "What has indeed been prompted is some calcification but I think these are benign. What I'm saying to myself is why has it prompted for those when there's actually similar calcs all over the place? So - for that reason I found it a bit distracting. C2, why only these calcs?" (A-1.33)

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." (A-1.64)

"I can see why it's outlined this area but I do actually think there are areas on the other side which are just the same." (B-1.64)

Subjects' judgements for many of these cases is that a prompted feature often is not (diagnostically) different from other unprompted features within a particular mammogram. Their expectation is that the system, in noticing possibly suspicious features, should be consistent. Subjects are able to make generalisations — that features that have a different appearance within the mammogram are occurrences of the same sort of presentation. This system is unable to make such fine judgements, and may respond differently due to variations in image properties that may seem insignificant to a film reader, or that are difficult for a film reader to detect.

This case can be seen most clearly in the operation of the microcalcification detection algorithm where a simple clustering rule is the system's criteria for suspicion. Benign microcalcification is an almost ubiquitous phenomena and calcification of arteries within the breast is particularly common. The majority of FP prompts produced by the micro-calcification algorithm are due to vascular calcification. Calcification of arteries is progressive, resulting in unbroken calcification in the walls of the vessel giving a characteristic 'tram line' like appearance on the mammogram. In case 1.20 and 1.39 (above), subjects express confusion because of the system's differential response to seemingly similar regions of vascular calcification. However, their perception of inconsistency often arises because the operation of the system is far less sophisticated than they suspect. 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. In case 1.33, subject A identifies widespread benign calcification, but is confused as to why only a small number of particles have been prompted. However, the system is neither able to make global appraisal, nor can it recognise individual particles as being characteristically benign. Again, it merely prompts for a region where, by chance, particles are sufficiently close together to satisfy its clustering rule.

From the subject's point of view the system is drawing a distinction which they cannot see for features that are interpreted as being the same 'sort of thing'. Because they cannot account for this discrepancy, subjects were inclined to doubt

the system's reliability by stating that its behaviour is 'inconsistent'.

Apparent inconsistencies may not only have an effect on impressions about reliability, but also on subjects' beliefs about the system as an indicator of suspicion. If the system has made a distinction it might be felt that this is for a reason — that some characteristic of the prompted feature has been found to be significant in some way. If readers have this expectation, then they might try to account for the differences in terms of significance:

> "[...] These are C2's - these are benign calcifications, they're not in anyway polymorphic. So that's the kind of subtle difference between the two. That is something could be misconstrued as a cancer - this should never be misconstrued as a cancer - so our tolerance to see it prompted goes down a bit, you can say, well, you know: 'Why has it prompted that - I mean that's just so obviously benign'. [...]" (A-1.56)

In addition to informing beliefs about system performance, scrutiny of the system's differential response to similar types of feature can sometimes lead to rationales for 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 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. But, I mean it's the same vessel here, higher up that hasn't highlighted that. That's arterial." (B-1.59)

> "There's definitely arterial calcification, the same goes all the way along this very long bit here that's of course got breaks in it, which is why it has to be given different er... Right, so all the way down to... it's all vascular, it's all vascular..." (B-1.28)

> "And I think that this is a false prompt. It has gone for those, maybe because there are two dots that are slightly closer together. But, benign calc, C2. It's gone for it obviously because there is one so close to the other. I'm adopting a neutral stance about whether they are acceptable, yes it would be." (B-1.33)

"And we've missed out, we haven't annotated any of this stuff — but it's benign calcification. I suppose there aren't really clusters of it, that's why it hasn't (...?)." (B-1.72)

In the above cases subject B is able identify 'low level' differences between prompted and unprompted features — the fragmentation of vascular calcification (1.28), improved contrast between calcifications and background tissue (1.59), the presence or absence of clustering in widespread benign calcifications (1.33 and 1.72 respectively). However, it can be difficult to sustain these types of explanation and use them in a more generalised way:

"There's a non-cluster area, I don't understand why it's prompted that and not that. Benign calc. I've called it micro-calc, but I think it's... distract... See it's gone for that, I don't really understand why it has gone for that, and hasn't gone for that, or that." (B-1.69)

In case 1.69, subject B identifies a microcalcification prompt for an area of calcification which, in her opinion, does not form a cluster. This particular system response provides evidence that is contrary to the supposition that the action of the algorithm can be explained simply in terms of a clustering rule. It is possible that a microcalcification detection algorithm has mistakenly identified small bright densities (such as crossing linear structures) as additional calcification particles. Without the knowledge that the algorithm can generate false positives in this way, it becomes difficult to account for cases where false positives are produced for non clustered microcalcifications, and still rely on the general assumption that the system will detect clusters.

However, subjects A and C do identify situations where the microcalcification detection algorithm does produce false positive prompts in this way:

"...I think there is an appearance of a couple of flecks of calcium there, it could just be (...?). If it is calcium I don't think it is significant. It could just be a kind of mottled appearance where you're got a lot of stromal elements criss-crossing giving the appearance of dots. You know, there's an extreme example, of that, or that, where you are getting a vessel or a stromal line end on." (A-2.55)

"That's just some benign calcification, and what have we got over here? [break] ...that sort of bright bit. It's picked up (...?) and thought it was calcification, but it's not calcification. So then, a's not helpful. (...?) [break]" (C-2.52)

In the absence of detail about the operation of the detection algorithms readers may form partial explanations for the presence and absence of prompts. This process can be thought of as reverse engineering, whereby subjects attempt to formulate a rationale for the system's behaviour from their observations of its responses in specific situations. However, this is a difficult task since it depends both on the chance occurrence of examples that unambiguously illustrate a single aspect of the system's operation, and on the ability of the reader to build a unified model of rules and exceptions from these snapshots.

Occasionally, this process will yield results. For example, subject B was able to account for the tendency of the ill-defined lesion detection algorithm to produce false positive prompts for "linear increases in density" in a way that eluded the other subjects:

> "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." (B-1.24)
>
> "It's back on to this sort of linear appearance there, these are not technically good films, the right exposure is not very good at all. I would be technically recalling this women." (B-1.45)
>
> "We're back to our friend... It seems to do this quite a lot, pick up on, on, what seems to be linear masses, and they're not — it's just overlying fibrous strands. So there isn't a mass, it's called it a mass, but there isn't a mass, it's composite." (B-2.5)

5.6.2 Assuming purposeful behaviour

Much of sense-making done by subjects betrays an assumption of rationality and purposeful behaviour — they not only assume that there is some reason for the presence of a prompt, but also that there is some *good* (i.e. diagnostically relevant) reason. This can partly be explained in terms of accountability. A reasonable default assumption would be that the system has prompted something of significance, thus demanding readers to find some good reason why the prompt should be ignored. It can also be explained partly as a means of trying to understand system function. Given that subjects have little or no knowledge of the working of the detection algorithms, the assumption that the system behaves in a purposeful way provides an alternative means of understanding its behaviour. For example:

"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 (...?)." (C-2.25)

Subject C's presumption when attempting to understand the system's differential response to lymph nodes over a number of cases is that there 'must be' a good diagnostic reason for the majority remaining unprompted.

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

"[break] ...an elliptical prompt there round something that I' 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." (B-2.31)

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

It is interesting that the chance association of prompts with asymmetrical regions can give the impression that the system is effective at detecting asymmetries, precisely because in cases where the asymmetry is actually minimal the system might be thought of as being highly sensitive:

"(...?) asymmetries, it seems to be quite keen on asymmetries. I wouldn't have described that as asymmetric though. Wee block of breast tissue, I would pass that as normal." (C-1.9)

"It's acceptable. Wouldn't recall it — it's just asymmetry. It's quite good at detecting asymmetry isn't it? I mean, the trouble is, I think that perhaps, maybe too sensitive." (C-1.59)

However, the perception that the system does prompt for asymmetry can lead to unwarranted expectations with respect to performance:

> "And I would also want a be prompt for that. Surprised it didn't pick that, because that, presumably it's something to do with asymmetry. I thought that that was quite strikingly asymmetric." (C-1.7)

> "I think that it's interesting that they've not prompted for this area of asymmetry [...] and there is marked asymmetry there which has not been picked up there so I'll call that 1 and I'll classify that later on." (A-1.10)

Furthermore, when making judgements about asymmetry, readers will take technical factors into account that might explain a differential appearance — such as different exposures. Such factors can also be used to explain the action of the system:

> "I don't think there's anything to that at all. Presumably it is just picking up on the asymmetric because of the different exposures. Could it be? I don't know. I don't think that's particularly helpful." (C-1.45)

Memory for both previous system responses, and candidate explanations for those responses, form part of a complied 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 revision as the reader is exposed to additional cases. When seeking to comprehend current prompts, subjects may appeal to the idiosyncracies of individual mammograms in order to explain the action of the system (as suggested by subject C in case 1.45 above). This has the advantage of conservatism — new rationales need not be developed, nor current ones abandoned. Instead, exceptional circumstances are invoked in order to shoehorn the system's response into an existing rationale. However, if this becomes an increasingly difficult task, then radical reappraisal of a dominant rationale may be precipitated. Towards the end of the second session, subject C began commenting on cases that had been prompted bilaterally:

> "That's probably just breast tissue, but needs further investigation. e [break] ...can't have it both ways, that's dense and that's not dense — so that's not helpful. Just breast tissue." (C-2.49)

> "That's both of them. [break] I don't why you can prompt them both, I thought it was supposed to rely on asymmetry. So that doesn't help me to look at it more, I look at both breasts anyway, so that's no. I don't know, I just have to put a question..." (C-2.64)

The accumulation of such cases provides the necessary leverage to enable an alternative account of system behaviour to be contemplated:

> "What's that one? Well a first of all. Well that's nor... that's not very helpful, the asymmetry is minimal. Or is that picking up as a whole mass? Is that what the idea is? I don't know. Anyway it's not particularly helpful." (C-2.67)

5.6.3 Explanations cued by prompt characteristics

Occasionally the shape or size of a prompt is used as an indicator of what the system's response might mean. For example, the ill-defined lesion algorithm will occasionally falsely prompt the glandular disk:

> "For some reason, it has decided it's the whole breast. Now I don't see anything wrong with that, they are perfect films... well, the nipples not quite in profile, but they're lovely pecs and all the rest, and the the breast tissue's all on. I honestly do not know why it has prompted this, I wouldn't recall this at all. It's normal." (B-2.19)

In case 2.19, the prompt is apparently 'for' an entire breast. Consequently Subject B looks for an interpretation that might correspond with some global phenomena, such as poor technical quality. It seems that the prompt not only directs attention to a particular region of the mammogram, but also by its form give some initial indication of what to 'look for'.

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 micro-calcifications 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:

"Some calcification, judging by the shape of the prompt it's probably vascular." (C-1.33)

"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." (B-2.11)

On a number of occasions subject A suggested that the area covered by the prompt is too broad to be useful, for example:

"Now, what's been prompted are two large areas which, that's quite unhelpful really. I think what you're wanting to do is (...) helpful if you focussed in on smaller areas, and indeed this area here has not been prompted at all. So I don't think the appearance...although there are areas within that area that I have picked up that's not been very helpful at all." (A-1.7)

Although she has identified a feature that is potentially of interest within the region circumscribed by the ill-defined lesion algorithm, she is not convinced that the system's response was triggered by that feature. Other readers were more lenient, and willing to accept serendipitous responses:

"And that — I don't think it's round the calcs particularly, though it includes the calcs — but I don't think that's necessarily what it's gone for. It has done it, it would direct me to see the calcs, so it's — that's acceptable. That's BT plus probably benign. And I would recall just to see the calcs." (C-1.54)

Responses to case 1.61 provide an interesting example of how subject's interpretation of the meaning of a prompt can differ. It also demonstrates readers trying to make sense of the system and the difficulties they face in doing so in the absence of an accurate model of system function. The prompt sheet annotated by subject A is shown in Figure 5.3. The commentaries for this case, from each subject are shown in Figure 5.6.3

The system has produced three ill-defined lesion prompts, *b* circles a large region of dense tissue in the upper part of the right breast. Prompts *a* and *c* highlight smaller regions, with prompt *c* lying wholly within the region circumscribed by prompt *b*.

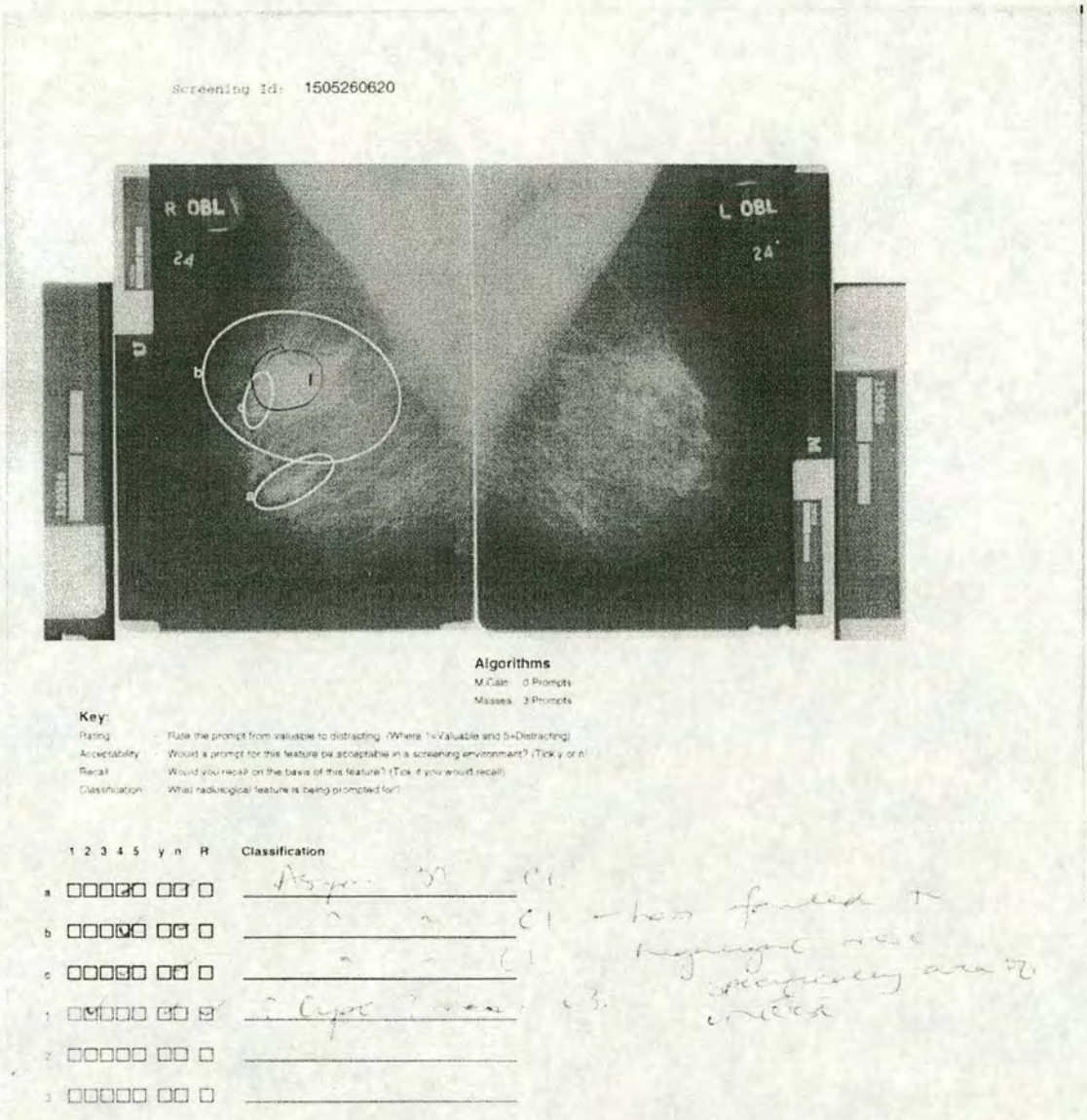


Figure 5.3: Case 1.61 with annotations by subject A

Subject A identifies a significant feature within the region highlighted by prompt *b*, but doubts the relevance of prompt *b* because it prompts a too wide an area — it does not concur with subject A's notion of significance. Subject A also dismisses prompt *c* as being for the wrong area, and proceeds to annotate the feature and labels this as '1'.

In contrast, subjects B and C feel that prompt *b* is relevant, but both have difficulty finding an adequate explanation for prompt *c*. Subject B presumes that the system has identified a 'separate mass', but is confused because the only likely candidate is the region annotated by subject A. Subject C believes prompt *c* to be redundant because prompt *b* 'mentions the whole thing, and not just one part

of it'.

From these commentaries it appears that the area of interest is smaller than the prompt might suggest, that is: as annotated by subject A, and indicated by subject B. Subject C's view on the precise region of interest is a little less clear, and he may view the entire prompted region as being significant. However, prompt *b* does refer to the smaller region. The discrepancy arises because the ill-defined lesion detection algorithm scales the prompts by a constant factor to prevent the prompted feature from being obscured. Taking this approach produces a larger gap between the edge of the feature and the edge of the prompt as the size of the feature increases.

Prompt *c* poses a problem for subjects B and C 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 A as significant (although by a wide margin), why is it that *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 dilemmas 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 — that is, the significance any given feature plays no part in determining the significance of any other feature. Prompt *c* is the result of the independent processing features from a smaller scale size, and is unrelated to any significance the system has associated with prompt B.

5.6.4 External explanations

When trying to make sense of the apparent inconsistencies in system behaviour, subject A resorted to a level of explanation external to the operation of the system:

> "Bit of overkill, we're also going through some - I'm slow on the uptake here but twice we've changed the sensitivity here." (A-1.27)
>
> "I mean when your getting one fleck of calcium you can't start - I mean obviously in this test you've varied the sensitivity and there's

Subject A

"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. It is probably distracting. And 1 would have been more helpful - query cyst, query mass..." (A-1.61)

Subject B

"Ah right, there you are. The first thing that strikes me is the big annotation around this up here. Now definitely an increase... so b, an increase in density and it does merit... it's helpful, and indeterminate but it needs... 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 women anyway so I will see what it's like, but if that wasn't there, then I wouldn't be impressed by that, so I wouldn't recall it for that. I don't think it has been helpful. C is in the middle of this bit. I'm not quite sure whether it thinks there is a second mass within... I suppose what's happened is that it's outlined this whole upper half and then within it... what was that? 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, and it's drawn it just round here. So I'm not quite sure what its getting at. I will give it a question mark. Now, having seen that, I actually am quite suspicious about this. I'm looking at the other side, and I don't see anything else for annotation." (B-1.61)

Subject C

"There's quite marked asymmetry in the upper part of the right breast, and I would certainly want to investigate that a little further. I don't see anything else. Yeah. a's a bit unhelpful. b's a bit that's a real, really interested in, and c's not particularly helpful. So a is a bit of a distraction really. b is useful, yes, yes — asymmetry — PB. And c, well no I don't think that's, well a mentions the whole thing, and not just one part of it, so c's a bit of a distraction. (...?) only got that in isolation it would be a distraction. So I would recall that. It's the whole thing, not just this one part of it." (C-1.61)

Figure 5.4: Subjects' commentaries on case 1.61

lots of inconsistencies you know - which is manifest." (A-1.35)

"And indeed, it's not picked up the vascular calcification. So why has it? That's you, you're fiddling it." (A-2.50)

Subject A's belief that operational parameters of the algorithm might have been changed for some experimental purpose indicates that she was unable to find a way of accounting for system responses solely in terms of its observed behaviour. Accountability for performance is transfered from the system itself to investigator. Furthermore, this subject was most closely involved with the development of the system, and knew those involved with development of the algorithm components. During this exercise she made reference to these people specifically as having some share of the responsibility for systems performance (fictitious names are used):

"A must be vascular because it's ignored the big dod there...so off we go, we don't like vascular, you tell Tom we don't like vascular calcification." (A-1.38)

"What's been prompted is the whole asymmetry. So Dick is not good enough. Tom's too good and Dick is not good enough." (A-1.45)

"Well, it is a bit of a let down. I'm thinking of Dr Dick Francis at this moment, and... [break]" (A-2.48)

In addition, subject A frequently referred to the actions of the system as being due to 'they' (the system developers?), rather than 'it' (the system), for example:

"Which makes it much more worrying in my mind - and then on the other side, they have prompted an asymmetry and it's not prompted for areas of calcification." (A-1.19)

"The rest I don't really see why they've prompted it. Because it seems very similar to surrounding tissue." (A-1.23)

Clearly the issue of accountability is important, and subject A in particular demonstrates a tendency to hold those responsible for developing the system accountable for its behaviour. Similarly, early on in this exercise, subject B expressed a naive (to a developer) view that system responses are somehow directly related to annotations made by radiologists: [6]

[6]Training and test cases for use during the development of the system were 'annotated' by film readers. This involved using a purpose-built computer package to delineate and label significant regions in digitised mammograms. Both subjects B and A were involved in this activity.

"I think that it's completely normal. I think that would have been whoever had annotated the films, which could well have been me. It's not, not, not the machine's fault." (B-1.2)

"Because I can't see a mass, and I can't see any microcalcification. So I am not terribly sure what A and B are about. It's a good demonstration of how you change your mind when you look at a film again, because I could have well annotated this." (B-1.9)

Here subject B is holding the radiologists who produced annotations to train the system responsible for its behaviour.

Subject B sometimes resorted to explanations that might be more appropriately used to explain the behaviour of a human observer, rather than that of a computer, for example:

"As for the reason for it doing — B's called the left breast, and C is the right breast. But I don't know — breast tissue, it's breast tissue. And please, let's not have these, because it's not helpful. It's not helpful to outline the whole breast. It's a bit of a cop out. I think that mean I can't be bothered doing any more somewhere around the end." (B-2.64)

Although a computer system obviously cannot become 'tired' or 'bored', it can be reasonably supposed to be 'defective' in some respect. When the system's response is obviously (to the reader) naive or inconsistent then there might be some doubt as to whether the system is in fact functioning correctly.

In deciding what constitutes a reasonable prompt, subject B often appealed to what she considered to be technically feasible. With respect to distinguishing vascular calcifications:

As I say I don't know quite if we're trying to get the computer to pick up micro-calc, I don't know how you can get round that. You can't, you'll just have to prompt it and you can dismiss it. So in an ideal situation it would be able to tell this is what it was, but I don't know that that can be helped. It's distracting to be honest. (B-1.27)

"Vascular calcifications are a problem. It's a problem for the machine, it's not a problem for us. So that's why, I don't think it, you know, I think it's acceptable, because you can look at it and say 'that's what that is'. I don't see how you could expect me... you couldn't expect the system not to pick that up." (B-1.65)

Detection of distortions in 'difficult' breasts:

> "And it hasn't been prompted. So I'll just put '1', query distortion. It's probably not, it's probably normal but just wonder... It's probably normal, but anyway I think it would have been helpful to be prompted but I think it's quite difficult to see how it could do it." (B-1.22)

Overall specificity:

> "I don't quite know how we sort this out. I'll accept it I think. I'll have to accept it's going to be prompting for far more things than..." (B-2.25)

Benign calcifications:

> "That's benign calc — C2. I think the way the prompt works you have to accept that these things will be prompted, but it is benign. It's of dubious helpfulness, however." (B-1.5)

In making sense of the system, Subject B appreciated that the system operated in a different way to human observers:

> "I think the prompt's problem is going to be sensitivity — or that's likely to be the case. It's logical thought that it can't, or maybe illogical thought, and tangential thought, that's maybe it — what it's missing." (B-1.20)
>
> "Oh sure, yeah, but that's (...?) I mean you can't expect it, you're not creating a robot for goodness sake, and a brain. I mean it's got far better brain than us, but it just can't quite... it's not just the same." (B-1.66)

Subject B tries to rationalise the system's shortcomings by drawing a distinction between the properties that might be appropriately attributed to human and computer agents. This reaction is similar to the way that subject A attributes responsibility for the system's behaviour to the system developers, rather than the system itself. In both cases, 'intelligent' and 'responsible' behaviour are conceived of as being in realm of human activity. The system's responses are viewed not only as result of the system's analysis of the image, but also as an end point of a causal chain that begins with the development and training of the system.

5.7 System and reader performance

This section is concerned with how subjects make judgements about both the capability and utility of the system. In particular, how subjects gain an impression of the system's performance, and how they perceive that use of the system affects their own. These perceptions are important because they have a bearing on subjects' views as to what degree it is appropriate to engage with the system's responses. The discussion is in three sections. The first concerns how subjects assess the system's performance. The second considers examples of where the system brought features to subjects' attention that they themselves had not noticed. Finally, occasions where subjects felt disinclined to examine the system's responses are discussed.

5.7.1 Assessing system performance

One conclusion drawn from the 'subjective responses to prompting' experiment described in the previous chapter was that subjects seemed able to make accurate judgements about the system's sensitivity despite the small number of actual cancers in the test sets. It was suggested that readers are able to extrapolate from the system's performance on the larger set of features that have suspicious characteristics but for which readers have a lower confidence of their being cancers. A prediction was made that readers would show tolerance for FP prompts corresponding to features that are actively considered for recall. In this exercise there were examples of cases where a feature was deemed sufficiently suspicious to prompt, but not suspicious enough to warrant recall, for example:

> "I can understand why they have prompted for that though, because it is an asymmetrical island surrounded by fatty tissue. So I don't think that's distracting, I think that's probably valuable, and I would accept that, I probably wouldn't recall it though, but it wouldn't be unacceptable." (A-1.13)
>
> "And b is, yes that's a fair reason to prompt that, and it's quite acceptable, I wouldn't recall it, it's just BT. It's just quite useful to have it brought to your attention." (C-1.64)

However, these reflections (and others discussed throughout this chapter) reveal that tolerance for prompts is often not solely judged on the significance of the feature alone. In cases 1.13 and 1.64, subjects show tolerance because they can

'understand why' or there is 'fair reason' for a prompt — indicating not only a degree of agreement with the system's assessment, but also a tolerance because the the system's responses are perceived as rational (or at the least comprehensible) in a way that is relevant to the task of interpretation.

In the extracts below, subject A did not recall case 1.30, but cases 2.42 and 2.50 were recalled by subjects B and A respectively:

> "But I think that's probably benign - it's picked that up I think that's useful. Would I recall that? No I wouldn't recall that but I think it's still been valuable because it's an area that caught my eye as well." (A-1.30)

> "Yep, I can see what it's going for. I'd be a bit suspicious about that too. It might end up to be composite, but... just a feeling that there might be a mass behind that. So that would be reasonable enough, that's the only thing that has been prompted." (B-2.42)

> "Right, so, this is looking good on the prompt because it's picked up the one which is least well defined there, which is good. I think that one probably also merits... although you could argue that's reasonably well defined. [...] 'Agree, do not recall left soft tissue opacities'." (A-2.50)

In the above extracts subjects draw attention to the consonance between the system's responses and their own. Not only are the system's responses rational, but they are also accord strongly with the readers' own view of what is accountable in the mammogram — i.e., features that might require closer inspection or further tests. Subject A's commentary for case 2.50 (Figure 5.5) is particularly noteworthy — by prompting the most significant of a number of similar types of presentation her perception is that the system has made an intelligent or competent decision under difficult circumstances. This can be contrasted with subject B's response to the same case:

> "Now this is one that I don't understand. Looking at this I don't understand why it hasn't annot... it seems to have completely ignored the left breast. [break] ...anyway, we'll concentrate on the right, what it has annotated. And it has annotated some things and not others. I think this might be the time it was getting on '15', it might be the same tired point." (B-2.50)

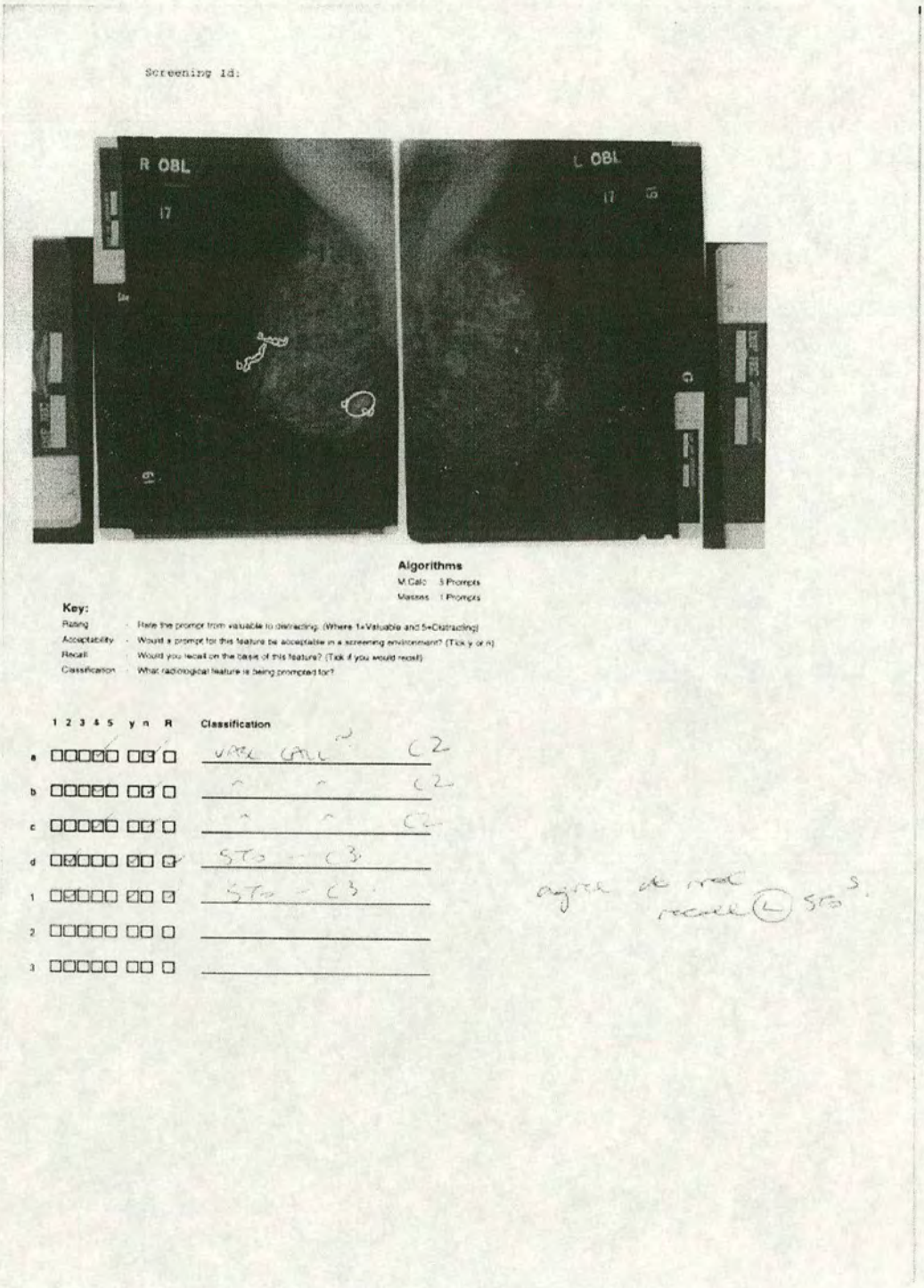


Figure 5.5: Case 2.50 with annotations by subject A

Subject B is inclined to believe that the lack of a response to features in the left breast is indicative of negligence rather than of intelligence. Although subject A also expresses surprise that features on the left breast remain unprompted (she makes a note of her approval on the prompt form itself), she views this as positive rather than as problematic. The distinction is perhaps between whether

subjects believe there has an omission, or considered decision not to prompt. The system itself supplies no information to suggest that the left breast was actually examined but that nothing suspicious was found. Given an understanding of the sorts of features that the system is likely to prompt, a possible explanation is that some system failure has occured with the result that the left breast has not been examined at all. Rather than simply attending to individual features in isolation, subject A appears to be forming an interpretation of the mammogram as a whole. Significance is not only dependent to the character of an individual feature, but also on the wider context of a feature's presentation. The system is interpreted as having made an analysis in a similar fashion — thus the absence of a prompt can be as important in evaluating the system's performance as the presence of a prompt:

> "I don't see anything else to outline, this is a lymph node — it hasn't outlined it, that's fair enough, that's a lymph node." (B-2.44)
>
> "And there's a big blob of very dense calcification on the right, but I think that it correct not have prompted that, because it's definitely benign." (B-1.18)

Prompts that correspond to areas that have some identifiably significant characteristic might reinforce a reader's confidence in the system, however, prompts that correspond to unambiguously benign features may have the opposing effect. Prompts for obviously benign features may lack both a rationale and the merit of an association with significance, so instead of appearing competent and rational, they may appear naive and arbitrary.

Judgements about system performance are probably useful in a number of different ways. For example: they may serve to establish the reliability of the system, that is, how much the system can be depended upon — which may in turn have an impact on a reader's own vigilance. They could be used to inform views about the balance of costs and benefits of system use, and thus impact on the attention given to the prompting system. Also, they could influence the degree to which the system is trusted — thereby determining how much credence should be given to the system's analysis where the reader is uncertain about an interpretation.

Subject C directly refers to how the system's behaviour influences his confidence in the system's abilities:

> "What's been prompted for is just this area - I think that is valuable - that's the sort of thing that should be prompted and I personally

would not recall that but I think that's valuable. But again, having something like that prompted I suppose in a way kind of gives you confidence in the system if that hadn't prompted for that I would be thinking Mmmm, you know." (C-1.18)

"[...] well that just catches your eye instantly. So a will be, well — it's valuable in the sense that god if it didn't (...?)." (C-1.8)

"And k, yeah well k is very useful but if it didn't prompt for that it shouldn't be doing the job I don't think. So yes that's useful, yes it's acceptable, and yes, it's a cancer." (C-2.17)

In the first of the above extracts, subject C believes the prompted feature to be benign, and in the second and third he believes the prompted features to be cancers. However, all three features are considered to be 'easy targets', and as such they provide an unambiguous benchmark for judging system performance. In fact, the feature in case 1.8 is judged to be so conspicuous that subject C suggests that a prompt would have little value as an attention cue because it is highly unlikely that the feature could be overlooked. The prompt has value only in that it reenforces the subject's confidence in the capabilities of the system. Subject C's response to case 1.18 is particularly interesting. The feature is judged to be benign, and the prompt therefore is effectively a false positive decision. However, the prompt is still viewed as desirable to the extent that the system would be believed to be less capable if the prompt were absent. Subject C is making this judgement based on implicit beliefs about the interpretive capabilities of the system — if a prompt were absent, then this would not be viewed as an interpretively acute decision, but as perceptually naive one.

Below are examples of subjects' commentaries where the system fails to prompt for regions that subjects believe to be significant:

"And I'm disappointed that that one hasn't been prompted I must say. Because I think that is actually reasonably suspicious. And I actually think that is a mass, well certainly, well, an opacity." (B-1.50)

"And it's missed it. I'm almost, you know - because I'm kind of looking at that at that I'm almost kind of dismissing that and hardly looking at that when it prompts me to look at that area because I'm more saying 'ach the system's rubbish - it hasn't picked up' - and when you look in fact there's a tiny wee cluster of benign calcification so that's been not a lot of help at all." (A-1.34)

"Oh no, that's not the same thing. Right that's not helpful, that's just breast tissue. That's distracting. That is not the area that I am interested in, thats just ... oh I see what that is, it's just picked up on this vague shape here. This bit of breast tissue. So I would actually not be, find either of these prompts useful." (C-1.21)

Subjects variously express disappointment, annoyance and surprise at the discord between the system's interpretation and their own. Subject A goes as far as to suggest that her confidence in the system is damaged to the extent that she is disinclined to examine what it is that the system has prompted. The system's response is judged to be irrelevant and inappropriate.

Subjects' responses imply an expectation about the capability of the system — occasionally these expectations were explicitly stated, for example:

"This looks like somebody has had previous surgery, there's a bit of distortion in the upper part of the left breast here. I presume that's post surgical, although I imagine it would prompt it to look at it, or would expect to be to prompted to look at it." (C-1.26)

"Right, there's quite a marked degree of asymmetry here, and the right side is even more dense than the left. And that I would have thought would be a useful prompt. if it's prompted, and I think there's sufficient asymmetry there that I you'd want to see that on another view, although there's probably nothing there." (C-1.30)

"Right breast is normal. There's some asymmetry in the lower part of the left breast, which I would expect to be prompted to look at more closely." (C-1.57)

"Similarly here again, one that strikes me is an asymmetry there, it's probably benign but these are the sort of things that I would expect it to throw back at me. What it's prompted is a tiny fleck of benign calcification so that's no use to man nor beast." (A-1.69)

"...I would have expected it to prompt at kind of density — I wouldn't recall it, but I would have expected that. [break]" (A-2.55)

Such expectations can be viewed as a judgement about what a competent system should be capable of detecting, thus providing a baseline for judging system performance. They can also be interpreted as indicating situations where the presence of a prompt is perceived to be useful because it would provide assistance in some way with the task of reading. It is interesting that again subjects suggest that prompts for benign features (that is, FP prompts) can be desirable.

5.7.2 Missed features

The intended function of the prompting system is to bring features to the reader's attention that they might otherwise overlook. In this exercise, subjects noted several occasions where the system did bring features to their attention that they did not identify for themselves. These events fall into roughly four categories:

1. Resulted in a recall decision.
2. The unnoticed feature was 'significant' but not recalled.
3. The unnoticed feature was 'insignificant'.
4. The subject had difficulty identifying the prompted feature.

Reports of a feature 'going unnoticed' can be ambiguous:

> "This, that again is a bit of asymmetry. I think that's actually quite a useful prompt, because it does draw your, I mean I didn't comment on that, I have to say, and there is some asymmetry, I would have to agree. So I think that is useful." (C-1.11)

There may be a subtle distinction between a reader overlooking a feature, and a reader noticing a feature but not comprehending its significance. In the above extract it is possible that subject C did notice the feature in his initial examination of the mammogram, but failed to realise an appropriate interpretation. It was suggested earlier that one role of the system might be to encourage readers to consider a wider set of possible interpretations. In addition to providing a psychological benefit by promoting a sense that a more thorough consideration has been made, it is also possible that FP's due to premature closure may be avoided. If used in this way, the system is not behaving either as an attention cue, nor as a classification aid, according to the accepted definitions of these roles. The feature has neither gone unnoticed, nor is there any protracted deliberation about its status. There is a failure of the reader to bind the appearance of the feature with an appropriate interpretation, but there is no intrinsic difficulty in doing so.

The following quotes are less ambiguous. The subjects appear to have overlooked significant features that the prompts then drew to their attention:

> "But I don't...there's some calcification, maybe that's what it is, that's what it is the calcification, that's a reasonable one. It looks benign

certainly, but yeah that's OK, useful, I didn't spot that, so that's useful. But would I recall it, I would probably recall on this that's been brought to my attention. I probably wouldn't have recalled it, well I obviously wouldn't have recalled it — I didn't see it in the first place. Having seen it, and having it brought to my attention would I recall it?" (C-1.19)

"I think however B has been, because I didn't pick up that, and it looks a bit more focal - so I think that has been valuable and I think it's not unreasonable to recall that. Query small mass, probably benign." (A-1.22)

"Now that is helpful because I've dismissed the... this micro-calcification here, but in fact, looking at the diagram trying to see where it is in relation to the more black bit, the lucent area's of the breast, in fact there is a tiny cluster of micro-calc. That has been helpful. And yes it does warrant recall on that. And I probably wouldn't have noticed it." (B-1.56)

Such occurrences probably play a role in assisting readers to form an opinion about the utility of the system by providing specific instances where the system has successfully fulfilled its role as an attention cue. Although the aim of a prompting system is to prevent cancers from being overlooked, it is also likely to draw readers' attention to features that are suspicious enough to recall, but that are not cancers, and thus it may have an impact on their specificity.

In the following cases, the system draws subject C's attention to unnoticed features, but he decides not to recall:

"(...?) a. Oh aye, there's a bit of calcification there, how did I miss that? It's a reasonable one, it's not (...?) would you recall it, no. It's still worth having your attention drawn to it I think. And b, I suppose it's fair enough." (C-2.37)

"There is some faint calcification there. All vascular again. In that particular case, if I didn't see it myself, so, it made me look at it so, acceptable I suppose yes it's acceptable (...?) but it is vascular." (C-1.65)

In case 2.37, subject C suggests that the prompted feature, although benign, is worthy of consideration. The feature is sufficiently suspicious to warrant attention and should be accounted for. In case 1.65, the significance of the feature is

dubious, however, subject C still rates the prompt as acceptable because of its effect in enabling him to make a more complete examination of the mammogram.

In the following extracts, subject C appears less convinced of the utility of having his attention drawn by the system to overlooked features:

> "Yeah, that's good. There's some micro-calcs been prompted as well with a, I think that looks more or less an artifact of the film. So a's... We've got to still look at it I suppose." (C-2.21)

> "If I got prompted for every bit of calcification as little as that, it would be a bit of a nuisance. So perhaps I shouldn't have said. (...?) or is it, I don't know. On the basis that I didn't see it when I looked at it, and I've now had a look at — and dismissed it, I suppose it's useful in the sense that it has brought me back to look at it. So, and therefore must be acceptable, but it's benign, tiny." (C-1.35)

Here subject C appears to be reaching a threshold in his tolerance, as if the less significant the prompted feature, the less importance is associated with his not noticing it.

In the following extract, subject A rated prompt B as distracting and not acceptable on the prompt form:

> "[break] ...(B?) is vascular calcification which I hadn't even seen. I have to be honest." (A-2.60)

Although the vascular calcification went unnoticed, she does not appear to show the same tolerance for the prompt as subject C does in similar circumstances. In fact, subject A's comment appears to suggest that the feature is 'beneath her notice'. It may be the case that features which appear more obviously benign demand less of an account than features that have an equivocal presentation, or appear more obviously malignant. Readers may feel a lesser duty to account for more obviously benign presentations. It was noted earlier that prompting for 'subtle' features may present readers with a problem of interpretation, in that readers may be concerned that the system has detected something of significance that they themselves are unable to perceive. Where a reader has difficulty actually identifying what the prompt is actually for, then there may be a considerable overhead associated with using the system:

> "Moving on to 1.15. Again there are isolated flecks of obvious benign calcification which I don't worry me in any way, and I don't see any

other worrying features. What's been prompted is (presumably)? a cluster of cluster of calcification posteriorly - I'm struggling to see it - I think there might be a vessel in that area - I think that probably has been quite distracting. I wouldn't expect that to be prompted and I wouldn't recall. I think it's probably vascular calcification - there - a tiny cluster - if its the (present?) at all" (A-1.15)

"I don't know whether it's maybe too sensitive. Because I can't in all honestly see where the calcification is supposed to be. I can't see... I can't see anything... I'm wondering if it's prompting a tiny wee linear line here, or what? Or there's some screen artifact. Because I mean there's some benign calcification on the right that it hasn't prompted for, but I would be ignoring anyway. And there's some other calcification on the left, which again is benign, so I'm not quite — it's certainly letting me down this time because I can't see that's it's calcification [...]" (B-1.15)

"Even with a prompt I can't see anything there. I can't actually see any abnormality, never mind something that's benign. I don't think there's anything there at all. Distracting, yes that's definitely distracting." (C-1.15)

A strong motivation to account for prompts may lead to a protracted and frustrating examination of the mammogram. Further to this, in prompting for features that do not require an account it is possible that the system disturbs a reader's orientation to the task of interpretation. When the system marks as significant regions of the mammogram that are uninteresting, its response may be discordant or disorientating. The reader would lose the benefit of effort reduction implied by selective examination as they are being encouraged to examine features that would have been implicitly discounted. Furthermore, any psychological benefits of FP prompts is likely to be minimal. Drawing readers' attention to the unequivocally benign will not add to readers' assurance that a more complete or thorough examination has been made. In fact, there may be an opposing effect:

"a is the left, there's benign calcification in there — I don't see a mass. I don't think it was very helpful. I takes my eye away from... And I wouldn't have recalled it to be honest." (B-1.7)

This effect may be be particularly strong where there are many irrelevant prompts, or where there are also features of interest within the breast that have remained unprompted:

"All of the left breast that has been prompted is vascular, there's probably a tumour here that we are missing because of all this crap." (A-2.68)

"a is actually distracting, because it's prompted you to look at that, you might ignore the bit you are really interested in. So it's actually quite distracting that." (C-1.21)

"I think that is distracting, because if anything I would have expected maybe this bit, or even that bit. But I wonder, you see now there is a kind of chunk of breast tissue gone from there. I just wonder if there has maybe been some surgery on that side. [break] So, I don't that's been helpful, it's probably distract you from the other features." (A-1.12)

In the above examples the prompts are perceived as distracting because they capture attention inappropriately. A similar phenomena may also be apparent in an unaided examination:

"[break] ...been prompted is the blob on the left and a little fleck of calcium which I didn't see on the left. [break] ...I think it's probably benign but... [break] ...similar to the other one, there's an example of where I missed it because I was looking at all the other blobs. So A is valuable." (A-2.37)

In case 2.37, subject A identifies a lapse of attention attributable to how she prioritised her examination of the mammogram. The system's response has the effect of broadening her focus by labelling simplicity discounted regions as significant.

In addition to inappropriately capturing attention, prompts might be distracting in more subtle ways — in the following case the ill-defined lesion detection algorithm has produced a prompt for a large area of apparently normal tissue:

"This is where we've prompted the whole of that, - the (issue)? of that whole breast has been of ductal prominence, but that's quite distracting. Because what you'll now do is you may forget to kind of go back and look at it closely and make sure there's...nothing." (A-1.36)

Subject A is concerned that in dismissing a large prompted region as benign, a reader may then fail to scrutinise the region closely for any smaller unprompted

abnormalities. Thus the prompt is perceived as potentially limiting the types of interpretation that might be considered when accounting for the content of the mammogram.

5.7.3 Loss of attention to prompts

A prompting system can only be useful if readers are sufficiently motivated to attend to the prompts. Subjects A and B commented that where there were a large number of prompts on one mammogram a degree of discipline is required to ensure sufficient attention is given to each individual prompt:

> "Another thing I'm finding I'm jumping around to find which is what prompt so thats quite unhelpful. I'm beginning to switch off now, there's so much now, I could easily - there might be..." (tape ends). (A-1.26)

The 'jumping around' that subject A experiences is an artifact of the protocol. Each prompt is labelled with a letter to identify it for the subject's written appraisal. If the subjects work through the prompts in alphabetical order this requires that they locate each prompt with the appropriate letter, but the prompts might not be presented in any logical order. However, subject A does note that her attention is wavering because of the number of the prompts, and then begins to articulate concerns about this loss of attention.

A large number of prompts on a single mammogram or mammogram pair is often due to the presence of widespread vascular calcifications:

> "I'm almost getting to the stage now, I've made up my mind and I'm classifying the prompts without going back to each one...which I think's important (...)." (A-1.27)

> "I think the danger actually is, that you get so... if you had a lot of these ones... and get so fed up being prompted by those, that you might not be quite as careful at making sure that are prompting only the arterial. I'm try to honest about it." (B-1.28)

Subjects A and B suggest there is a tendency in this situation to reduce the effort required to interpret the prompts by making the simplifying assumption that the prompts are all for the same kind of thing. There is a tension between subjects' motivation to account for each individual prompt, and the overhead associated with doing this when many FPs are produced.

"Yeah, you do have to worry a wee bit about that. I wouldn't have been any more helpful but... I suppose the other thing about (...?) when there's a lot of arterial calcification is the prompt going to be so coarse that it's going over an area where there's worrying microcalcification." (B-1.51)

Here the subject is concerned that significant calcification and vascular calcification might be grouped together within an individual prompt. Or that a prompt for significant calcification may occur within a group of prompts for vascular calcifications. In either case the concern is that the presentation of the prompts might be misleading. If limited attention is given to prompts that are judged to be 'more of the same thing', then such a significant cluster might be overlooked.

Subject A notices that her tolerance for multiple prompts declines as she works her way through the cases:

"So I'm interesting, I'm changing my mind - having said it was valuable at the start and now changing my mind. I think it becomes more distracting the more prompts that are on the film, if it's only a single prompt then you can easily say that's it, when you've actually got to start working your way through - here we're faced with four prompts - so that might be a factor." (A-1.17)

"It's interesting, I've changed my mind about vascular depending on how much (has been)? prompted. I wouldn't recall it." (A-1.20)

Similarly, subject C quickly reaches the conclusion that prompts for vascular calcifications would be irritating:

"And then a, microcalcs is it, I suppose there is some vascular calcs there, but that's a bit distracting. It's really. If it picked up on every case of vascular calcification, it would be really quite irritating." (C-1.20)

Subject A suggests that the prompts for irrelevant features of the breast might be acceptable depending on their context: if they are infrequent, are the only thing prompted, and are easily classifiable:

"There's an artifact down there - so I think it is an artifact. So that has been distract... well why am I calling that distracting - but the

other one I didn't in the infra-mammary fold? On it's own it's not a big problem (because you it)? erm, as I say - if there were about five other prompts you'd find it quite irritating so overall erm, not unreasonable but I think it's an artifact. Quickly classifiable." (A-1.32)

Subject C suggests a similar set of criteria (in this case a calcified node has been prompted):

> "They occur so infrequently that I wouldn't mind it coming up with this because it is very obvious what it is. There's no diagnostic difficulty. And they do occur rarely, and you put up with them. If you can pick up the microcalcs you could put up with the odd one of these, but you wouldn't put up with all the vascular calcs, that it's picked up." (C-1.65)

If false positives have the potential to be frequent:

> "Yeah, that's what it is, that's just artifact. So I don't want to be prompted for every time I see an artifact. [break]" (C-2.35)
>
> "Oh yeah, that's just...that's not helpful. [break] ...picking up on every time it showed... it saw something like that, so that's probably unacceptable, and it's just composite." (C-2.16)

5.8 Discussion

The notional goal of prompting is to reduce FN decisions by assisting readers to avoid errors of attention. However, the assumption that prompts can be used purely and simply as attention cues may be misplaced. One problem is that the design rationale for prompting systems assumes generic difficulties, for example, that readers may sometimes neglect to examine the entire mammogram. In practice, however, the difficulties associated with interpreting of any given set of mammograms are 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 may 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. In practice, there may often be a gap between the capabilities of a prompting system and the problem a reader is attempting to address.

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 understand the system as a 'black box', even if he or she is a highly skilled reader. On the contrary, this study shows that optimal use of a prompting system depends on its behaviour being accountable to its users. Readers maintain accountability in 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 PROMAM's behaviour, especially where erroneous prompts are simply artefacts of the methods used to analyse the image. For optimal use, an account of system behaviour is needed above that available by examining the prompts alone.

The findings in this chapter should, however, be viewed in the light of certain methodological reservations. Although the think aloud protocol generated a rich account of how subjects reasoned about prompts, the setting of its production was removed from usual clinical practice. For example, subjects were asked to reflect upon their interpretation, both of prompts and images, to a degree that is untypical of routine reading. Subjects' propensity to attempt to account for prompts will have been encouraged by the study protocol, since this is what they were being asked to do. Furthermore, subjects were examined prompted cases only and were denied the use of previous films or CC views. The following chapter describes pre-clinical trials of the PROMAM system where corroboration was sought for the findings of the think aloud protocol using conditions more reminiscent of actual screening practice.

Prior to the trial, subjects were presented with training material based upon the difficulties encountered by subjects of previous investigations. The material was designed to guide subjects in their use of PROMAM by giving details about its capabilities and to provide an understanding of its operation so that they might be better able account for its behaviour. During the trial, both qualitative and quantitative data were collected to examine again how subjects understood and made use of the prompts.

Chapter 6

Pre-clinical trials

6.1 Introduction

This chapter describes a trial involving five film readers reading 2002 archive cases in prompted and unprompted conditions. The trial was to serve as a final acceptance test of the PROMAM system prior to full scale clinical trials. Specific aims of the trial included:

- Examining the effects on readers' recall for assessment decisions to provide reassurance that any increase would be manageable by potential trial centres.
- Determining the level of agreement between prompted and unprompted readers so that more accurate estimate of the trial size could be calculated.
- Administering reading sessions in a manner similar to that planned for clinical trials to test protocols.
- Examining the effect of using the system on readers' detection performance [1].

In addition, the trial was used as a vehicle for further studying readers' interpretation of prompting information, and it is largely the results of this continuing qualitative evaluation that are reported in this chapter.

There are important differences between this trial and the prompting study reported in Chapter 4. Rather than testing the algorithms at differing operating points, a single operating point used. The performance of the ill-defined lesion algorithm had been improved in the intervening time, and it was able to operate at a sensitivity of 80% with 1 in 2 cases prompted. Subjects were given access

[1] This final aim was optimistic given the size of the trial, however, testing in this way would eliminate the possibility of gross positive or negative effects.

to previous films where appropriate — so the information available to subjects matched more closely typical screening practice. Finally, as a result of the investigations reported in Chapters 4 and 5 a training regime was devised to account for system behaviours that had previously been reported as confusing.

The study described in Chapter 5 provided a detailed account of responses to prompted cases in somewhat artificial conditions. In this study, while less detail could be elicited concerning subjects' perception of individual prompted cases, the data obtained pertains more directly to the interpretation of prompts in a screening context. This enables testing of whether the types of interpretation and usage strategies previously identified hold in practice.

6.2 Design of the pre-clinical trial

6.2.1 Test set selection

Two thousand and two cases were selected from the archives of women screened at the SESBSC between December 1995 and March 1997, including oblique and craniocaudal view mammograms. The set contained 1836 normal (i.e. non-recalled) cases, 64 cases previously recalled for assessment, and 102 pathology proven cancers.

6.2.1.1 Selection of Cancers

Not all the cancers presenting in the above time frame were selected. Cancers with an apparent diameter greater than 3.5cm were excluded. The rationale for excluding grossly obvious cancers was to increase the difficulty of the set. Of the 18 cases classified as architectural distortion, 14 of these were excluded because they were not targeted for detection by the system, and 4 were included to maintain the variety of cancerous presentations readers are likely to encounter. A further 28 cancers were excluded either because pathology information, or the cases themselves were not available from the archives at the time of selection. Table 6.1 shows the proportions of cancers in the test set compared with their expected rate of occurrence at the SESBSC.

The protocol for cancer selection was decided by the PROMAM project team. The protocol sought to achieve a balance between representing all types of cancerous presentation and increasing the likelihood of obtaining a measure of performance for the cancers targeted by the system. By applying selection criteria it is accepted that performance measurements would have limited generality. The qualitative data might also be biased – subjects would have little scope for wit-

nessing the system's performance on untargeted feature types and may develop a more favourable view of its performance than is actually warranted.

Although one of the goals of the trial was to obtain a measure of improvement in subjects' sensitivity due to use of the system, the small numbers of cancers in the set would preclude the detection of all but the most gross effects. The most important performance metric at this stage of testing was judged to be the effect of system use on subjects' recall decisions.

Lesion type	Natural proportions	Proportions used
Micro-calcification only	33%	31%
Mass only	40%	47%
Mass and micro-calcification	27%	22%

Table 6.1: Shows the proportion of cancer types in the test sets compared with the 'naturally occuring' proportions at the SESBSC.

6.2.1.2 Selection of normal cases

For each cancer selected, 19 normal cases were drawn from the same screening clinic. Depending on the size of the original clinics, normal cases were generally selected at equal intervals from the screening list (so with a clinic where 80 people had been screened, then every forth cases would be selected). A number of cases were excluded: if large films (10x12) were used, or if the case was unavailable at the time of selection. In these instances an alternative case was selected from the list. An attempt was made to select an equal number of cases screened by each of the radiographers on duty that day.

6.2.2 Processing

The 2002 cases were scanned and subsequently analysed by the ill-defined lesion and microcalcification detection algorithms. How operating points for each feature detection algorithm were chosen is described below.

6.2.2.1 Microcalcification algorithm

The operating point used by the microcalcification algorithm was decided after examining the results from several studies. The algorithm's sensitivity at various settings was measured using a database of 49 pathology proven malignancies and the algorithm's prompt rate was evaluated using three entire day samples from the SESBSC. These results along with the study into film readers' subjective

reaction to prompting (Chapter 4) were used to roughly determine a viable operating point. This point was subsequently fine tuned by adjusting the algorithm's clustering rule after obtaining a film reader's opinion on its performance on a set of interval cancers. As measured by these test sets, the operating point chosen produced a sensitivity of 90% with 1 in 4 cases falsely prompted.

6.2.2.2 Ill-defined lesion algorithm

An ROC curve for the Ill-defined lesion algorithm was established by measuring its sensitivity on a set of 92 pathology proven cancers (using a 'leave one out' train and test protocol) and its specificity on an unbiased day's sample of 119 asymptomatic cases, for different confidence thresholds. In conjunction with data from a study of sensitivity to interval cancer cases, and consultation with a film reader, an operating point was selected to try and achieve the best trade off between sensitivity and false positive rate. As measured on these test sets, the operating point chosen produced a sensitivity of 80% with 1 in 2 cases falsely prompted.

6.2.3 Subjects

All five film readers at the SESBSC were recruited as subjects in this trial. Subjects are referred to by the letters A to E, and feminine pronouns are used in order to preserve confidentiality. Table 6.2 shows the number of year's film reading experience in screening mammography that each subject had accumulated at the time of their involvement in this trial. Note that this is only a rough approximation to 'experience', as film readers might have different average reading loads depending on their responsibilities. In particular, subject C stated that typically, because of other duties, she tends to read fewer cases than the other participating film readers.

Subject	Screening experience
A	22 years
B	17 years
C	7 years
D	9 years
E	8 months

Table 6.2: Screening experience by subject.

6.2.4 Protocol

6.2.4.1 Subsets for reading

The set of 2002 cases was split into 20 subsets for reading, each containing approximately 100 cases. Cancers were assigned to subsets according to a binomial distribution about a mean of 5. Each subset was built up of 10 blocks of 10 cases, where the cases for each block of 10 were selected from the same day's screening. A block of ten cases could either consist of 1 cancer and 9 normal cases, or 10 normal cases. In this way, possible cues or patterns that might alert film readers to suspicious cases were minimised.

6.2.4.2 Allocating subjects to conditions

Each of the 20 subsets were read in prompted and unprompted conditions in a total of 40 reading sessions. Attempts were made to allocate subjects to reading conditions so that each subject saw an equal number of cancers in the prompted and unprompted conditions. In addition, it would have been been desirable for each subject to have read an equal number of prompted as unprompted sessions, and for each subject to have been paired with every other subject an equal number of times. However, due to time pressures and availability of subjects, the more stringent the constraints, the less they could be conformed to. A chronological breakdown of the trial sessions is shown in Table 6.4.

Some subjects read more prompted conditions than others so care has to be taken when interpreting some of the usability data — for example, attitude data will necessarily be richer for those film readers who have completed a greater number of prompted sessions. Furthermore, there is the potential for bias — for example, the opinion of subject E may well be over-represented in the interview data (Table 6.3).

Subject	Prompted	Unprompted	Total	Interviews
A	3	3	6	2
B	3	5	8	3
C	4	4	8	2
D	3	3	6	3
E	7	5	12	8

Table 6.3: Number of prompted and unprompted sessions completed by each subject.

6.2.4.3 Allocating cases to readers

Because cases for the test set had been drawn from the archives of the screening centre where the subjects for the trial had been recruited, there is the danger of subjects encountering and remembering cancers that they have seen before. Furthermore, some of the cancers used in this trial had previously been part of a set used to evaluate the system's performance, and had been 'annotated' by film readers serving as subjects. To minimise the possibility of memory effects, each of the cancers were given a scores according to each subject's previous exposure. Where it was not possible to show subjects cancers they had seen by them previously, cancers were allocated to film readers such that this contact score was minimised.

Session	Subject	Set	Prompted	Interview	Notes
1	E	1	No	Yes	1
2	E	2	No	No	1
3	A	1	Yes	Yes	1
4	C	2	Yes	Yes	1
5	C	3	No	No	
6	B	3	Yes	Yes	
7	B	4	Yes	Yes	
8	C	4	No	No	
9	E	5	Yes	Yes	
10	B	5	No	No	
11	E	6	Yes	Yes	
12	D	6	No	No	
13	C	7	Yes	No	
14	E	7	No	No	
15	D	8	Yes	Yes	
16	A	8	No	No	
17	A	9	No	No	
18	E	9	Yes	Yes	
19	C	10	No	No	
20	A	10	Yes	Yes	

Session	Subject	Set	Prompted	Interview	Notes
21	C	11	Yes	No	
22	E	11	No	No	
23	B	12	No	No	
24	E	12	Yes	Yes	
25	D	13	Yes	Yes	
26	D	14	Yes	Yes	
27	A	14	No	No	
28	E	13	No	No	
29	B	15	No	No	
30	A	16	Yes	No	1
31	C	16	No	Yes	
32	C	15	Yes	No	2
33	E	18	Yes	Yes	
34	D	18	No	No	
35	B	17	No	No	
36	E	17	Yes	Yes	
37	D	19	No	No	
38	B	19	Yes	Yes	
39	B	20	No	No	
40	E	20	Yes	Yes	

Table 6.4: Session details. 1 An additional observer was present at these sessions. 2 Due to shortage of time, this session was read in two sittings over two days.

6.2.4.4 Reading protocol

A modified version of the screening form usually used for reporting at the SESBSC was used for this trial. To emulate normal reading conditions, the form was attached to an empty film bag with a paper-clip. In the prompted condition, the prompt sheets were fastened beneath the reporting form so that the prompt sheet could only be examined by lifting the form. To facilitate blinding, a fresh reporting form, in a different colour, was supplied for the second reader. Each reader was told whether they were the first or second reader for a given session. Although case history and radiographers comments were not made immediately available, subjects were told that these would be made available upon request. A record was kept of any such requests. Previous films were made available to film readers where appropriate — however information from previous films was not used by the prompting system.

When reading a prompted condition, subjects were asked to observe the following protocol:

1. Examine the films
2. Examine the prompt sheet (by lifting the the reporting sheet).
3. Tick the box labelled 'Examined' on the prompting form to indicate that the prompting information has been considered.
4. Mark your decision on the reporting form as one of:
    - Routine recall
    - Technical recall
    - Recall for assessment
5. Move onto the next case

Subjects were asked to assume that they were a first or second reader using a blinded double reading system, and to assume also that recalls for assessment are made on a 'worst decision recalls' basis. For any cases recalled, they were asked to mark the position of the lesion on the breast schematic, and to give a description of the lesion type. If reading a prompted condition subjects were also asked to indicate whether the suspect lesion was correctly prompted or not by ticking a box on the modified reporting form.

6.2.4.5 Training

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

The developers of the microcalcification and ill-defined lesion detection algorithms each contributed prompted cases demonstrating aspects of the system's behaviour. Selection of examples was guided by the difficulties encountered by subjects participating in previously reported exercises (Chapters 4 and 5). Training sessions were conducted by the author, who gave a verbal description of each algorithm and then encouraged participants to read and comment on the selected examples.

The microcalcification algorithm was demonstrated first. Examples of TP prompts were shown, including examples of where multiple regions of malignant calcification had been prompted, and also cases containing both TP and differing categories of FP prompts. A FN example was given, showing how a cluster may be overlooked by the system if the calcifications are diffuse (not all of the particles are identified) and the distance between the remaining particles falls short of the critical clustering threshold. Finally, different types of FP prompt were demonstrated. Particular attention was paid to FP instances that might appear counterintuitive, including:

Vascular calcifications Describing circumstances where some vascular calcification and not others might be prompted. Continuous linear vascular calcification is typically not prompted. Regions where the calcification is discrete and in clusters may be prompted.

Edge of film artifacts Damage to films due to insertion in cassettes can cause bright spots down the edge of the film that the algorithm identifies as particles of calcification.

Edge of pectoral muscle Because the breast and the pectoral muscle have different noise properties, a different iso-precision scaling regime is applied to each. Sometimes the process that identifies the pectoral muscle is inaccurate resulting in inappropriate treatment for breast regions, and FP prompts may result.

Large benign calcification Although the microcalcification detection algorithm notionally targets clusters, sometimes large continuous regions of calcification maybe prompted, for example, calcified cysts. If the brightness across the calcification is uneven, then the algorithm can be 'fragmented' by the algorithm into regions that are then identified as calcification particles.

Each algorithm was run independently on example cases, so no microcalcification prompts were produced for the example ill-defined lesions, and vice versa. Training material for the microcalcification algorithm was presented first, giving rise to the problem that subjects then expected calcification in the ill-defined lesion set to be prompted. Once this was explained it did not pose a significant problem, however, with the benefit of hindsight it would have been better to have run both algorithms on this second set.

The ill-defined lesion algorithm examples included cases demonstrating circumstances where the algorithm might produce FN prompts, including:

- A reduced probability of prompting for cancers that lie across the edge of the pectoral muscle or that are bisected by skin folds. In these circumstances multi-resolution analysis will identify two candidate 'blobs' rather than the whole, thus rendering extracted parameters inaccurate.

- Lesions with an apparent diameter greater than 3.5cm are not targeted by the system. This is a change from the system configuration used to generate prompts for the 'subjective responses to prompting' experiment described in Chapter 4. Analysis at larger scale sizes produced a large number of FP prompts for only a small gain in sensitivity. It was felt that the reduction in sensitivity would be acceptable because large cancers are likely to pose fewer problems for human observers.

- Stellate features with no central mass are not targeted by the system — the system is not designed to detect the linear radiating spicules that characterise stellate lesions. However, where there is a mass present, the performance of the algorithm is similar to that for non-stellate ill-defined lesions.

- Lymph nodes and other clearly defined opacities will typically not be prompted. This is because the texture parameter enables them to be distinguished from malignant lesions.

In addition, examples were given of cases where FN decisions by the system were less clearly defined. Typically these correspond to an error on the part of

the classification stage of the ill-defined lesion algorithm. That is, a 'mistake' was made when trying to partition benign and malignant occurrences while maintaining a reasonable specificity. These examples included cases where the lesion had an irregular shape and low brightness, where there was overlapping linear structure, or where lesion had an elongated shape. Examples were also included of prompts where it was difficult for the system developer to account for the system's error.

A number of examples of TP prompts were given. These included examples where there were also false prompts on the same case in a manner similar to the training material developed for the microcalcification detection algorithm.

A number of typical false positive prompts were given, including:

- Dense regions of breast tissue.
- Composite breast tissue.
- The whole glandular region (if small).
- Skin folds.
- Points where blood vessels cross the pectoral muscle.

Finally, general limitations of the ill-defined lesion algorithm were explained. No use is made of previous films, nor of asymmetry between views and the system does not detect architectural distortion or skin thickening.

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

Summary of important details from the training sessions were produced on two A4 sheets (one for each algorithm) and made available to subjects at all times during the trial. These summary sheets and examples taken from the training package itself are included in Appendix C.

6.2.5 Data collection

There were four primary sources of usability data, namely:

Observation of trial sessions The reporting of each of the trial sessions was observed. The duration of each session was recorded (to the nearest minute).

A note was made of any additional information requested, and of any other remarks made by subjects, during reporting. The reporting style of each subject was recorded.

Post-session questionnaires A questionnaire was administered after each *prompted* session. This included a 20 point Likert scale questionnaire, designed to establish each subjects' 'attitude' towards the system as the trial progressed. To similar ends, each subject was asked to give the system a rating from 0 to 100. Subjects were also asked to state whether the system, or components of the system, would be useful in screening, and to rate the system's sensitivity and specificity. There were questions concerning how easily prompted features could be located on the mammogram, and how easily prompts could be interpreted. Finally, five free form response questions were asked with the aim of revealing issues not otherwise addressed.

Post-session interviews Where there was sufficient time, subjects were interviewed, usually on completion of *prompted* sessions. Subjects were asked if their decision-making had been influenced, and how they coped with the FP burden of the system. The interviews provided opportunities to follow-up remarks made on post-session questionnaires, and for film readers to highlight issues not otherwise addressed.

Pre- and post-trial questionnaires Each subject was asked to complete a questionnaire before undertaking any of the trial sessions, and after they had completed all their allocated sessions. The pre-trial questionnaire comprised of nine sets of questions — all of which were repeated in the post-trial questionnaire. These questions were of a more general nature than those asked in the post-session questionnaires and were largely concerned film reader perceptions of their own abilities and the abilities of possible prompting systems. An additional four sets of questions were asked in the post-trial questionnaire concerning film readers' assessment of their own performance and of the performance of PROMAM.

The results of this study are presented below in the following order: timing data, post-session questionnaires, interview data, pre- and post-trial questionnaires and finally, data concerning recall decisions. In addition, data is presented from a questionnaire given to film readers in 5 of the 6 centres studied as part of the work practice investigation. This questionnaire was completed by 16 respondents and contained questions in common with those asked in the post-trial

questionnaire. Details of these questionnaires and subjects' instructions are included in Appendix C.

6.3 Time to complete sessions

Figure 6.1 shows that, in general, subjects took less time to complete prompted sessions as the trial progressed — however, a similar trend is apparent for the time taken to complete unprompted sessions. This would indicate that these changes are due more to familiarisation with the trial protocol, rather than, in the prompted condition, with becoming more adept at interpreting/dismissing the prompts.

The reading protocol was tightly specified for the prompted conditions. Subjects were asked to examine the films, then examine the prompt sheet by lifting the reporting form, and then to record their decision. Similar guidelines were not given for the unprompted sessions and subjects typically employed one of two reporting styles, presumably reflecting the methods they adhere to during routine screening. Subjects either attended to the paperwork as they examined each case, or they deferred the paperwork until they reached a case they wished to recall, and only then marked their decision on the intervening cases. Table 6.5 shows the style of reporting used predominantly by each subject when reporting unprompted conditions.

Subject	Batch cases when reporting
A	Yes
B	No(?)
C	No
D	No(?)
E	Yes(*)

Table 6.5: Reporting styles used in the unprompted conditions. The question-marks indicate that for most sessions it is not recorded whether the subject batched cases while reading. (*) During their first unprompted session, this subject initially reported the cases individually, then later in the session switched to the batch style of reporting.

Subject E, in her first unprompted session, initially started to report cases individually, but later in the session switched to a batch style of reporting. The dramatic decrease in the time taken between subject E's first and second session in the unprompted (and to a lesser extent, in the prompted) conditions, was probably due to initial unfamiliarity with the trial protocol.

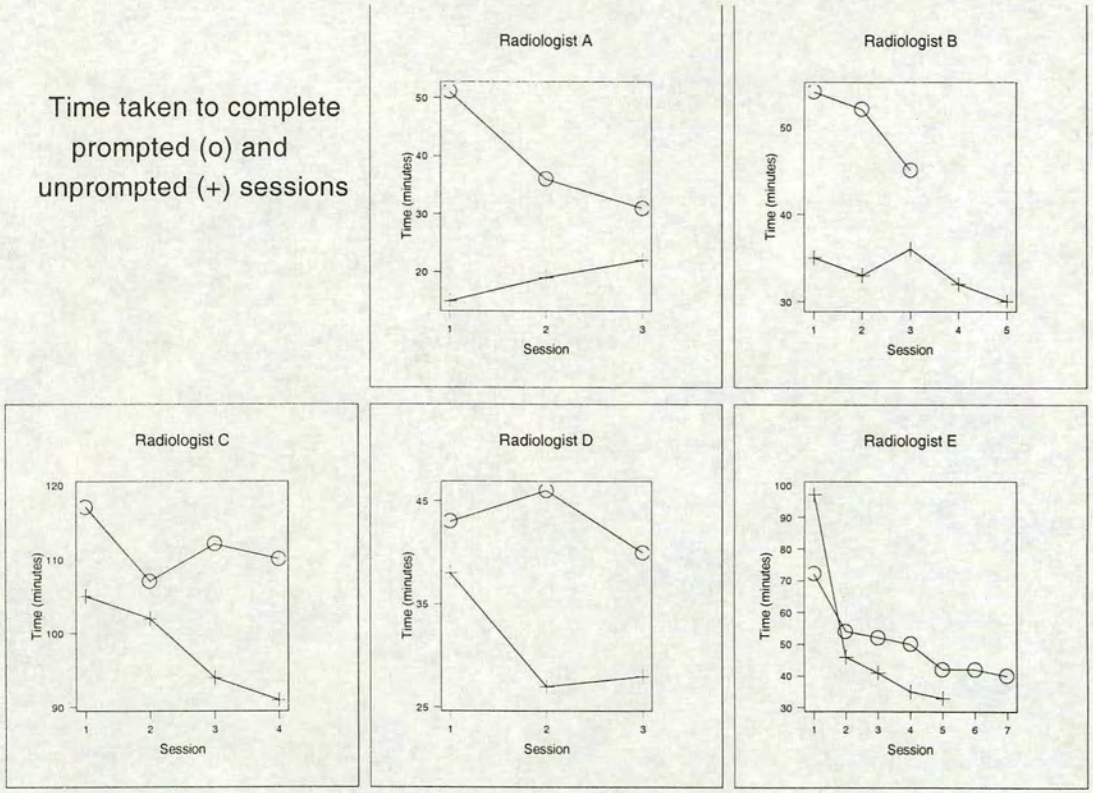


Figure 6.1: Time taken to complete prompted and unprompted sessions.

6.4 Post-session questionnaires

6.4.1 Attitude scores

A Likert test was administered to each subject after they completed each prompted condition, the results are shown in Figure 6.2. A higher rating indicates a more favourable response towards PROMAM — a description of how these scores are calculated is given in Appendix A.

The test was similar to that used in that used in the experiment described in Chapter 4, except that subjects were instructed to take into account their total exposure to the system, rather than only considering the session they had just completed. The richest source of Likert data is from subject E, who had been exposed to the most prompted sessions. The data for subject E indicates that she became less well disposed to the system during her second and third prompted sessions, after which her opinion remained fairly constant for the remaining sessions. A similar trend is shown for subject C. The initial lowering of opinion for these two subjects could be due to overly high expectations prior to participating in the trial, which could have been due in part to the training process. It is difficult to draw conclusions for the responses given by subjects A, B and D because of the scarcity of data. However, one possible conclusion is that subjects form a consistent opinion of the system relatively quickly, after completing three or fewer sessions (after seeing 300 or less cases).

Subjects were also asked to give the system a rating on a scale of 0 to 100 to indicate its overall usefulness (with 100 being the best possible score) after each prompted session (figure 6.3). These results show a pattern similar to that given by the Likert data. Figure 6.4 shows the Likert data plotted against the rating score. The correlation between the Likert and the rating score is 0.839, which is statistically significant at the 1% level — indicating that the Likert test is in fact providing a measure of attitude towards the system.

6.4.2 System appraisal

This section considers the responses to three of the questions asked in the post-session questionnaire:

1. Do you believe that the system overall, and each of the system's components would be useful to you in a screening context as they currently stand? (Yes/No).

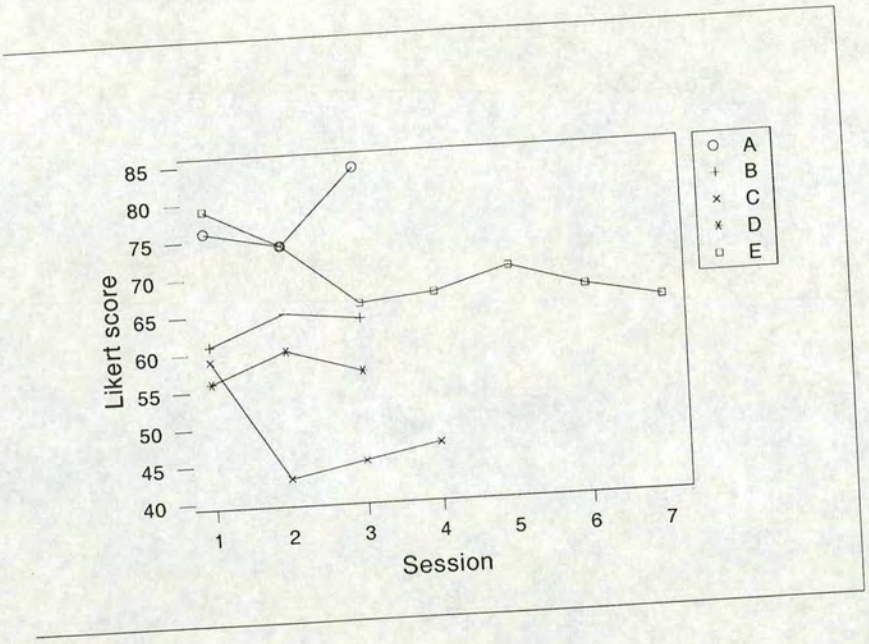


Figure 6.2: Shows the Likert score by 'session', where 'session' refers to consecutive completed prompted sessions. The higher the Likert score, the more favourable the view of the system.

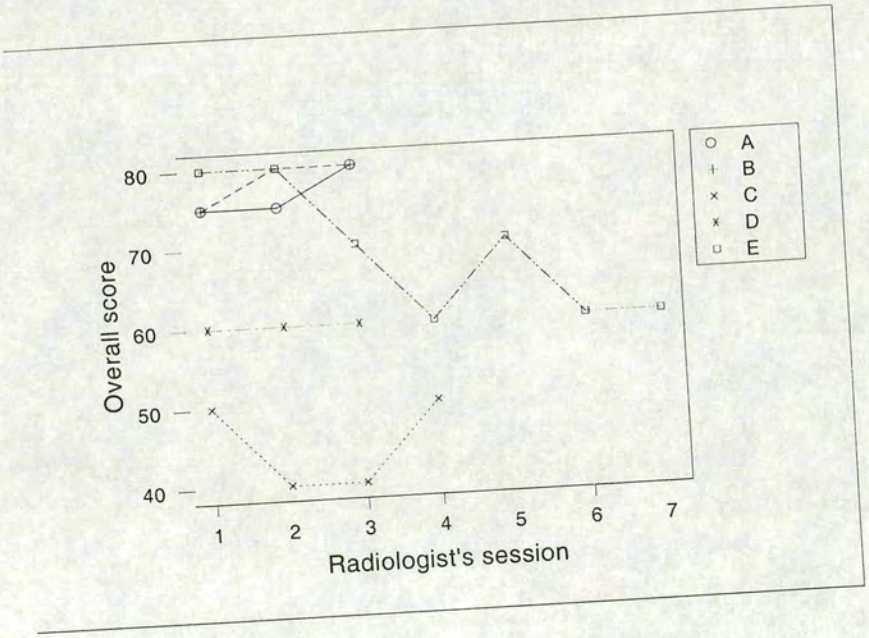


Figure 6.3: Shows the subjects' rating of the system, on a scale of 0 to 100 (the higher the score, the better the rating) after each prompted 'session' — where 'session' refers to consecutive completed prompted sessions.

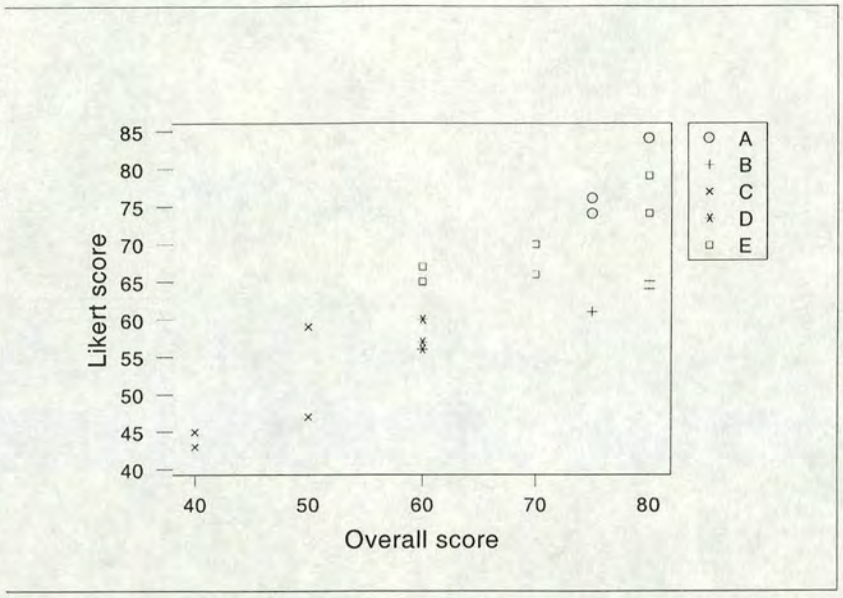


Figure 6.4: Shows the Likert score plotted against the subjects' rating of the system.

2. Please rate the sensitivity of the system overall and of each of its components as either 'Too sensitive', 'Just right' or 'Not sensitive enough'.

3. Please rate the specificity of the system overall, and of each of its components as either 'Too specific', 'Just right' or 'Not specific enough'.

6.4.2.1 Is the system useful?

Opinions of the system's usefulness appear to remain fairly constant as the trial progresses (Figure 6.5). There is a consistent consensus amongst four of the subjects that the micro-calcification algorithm would be useful as it currently stands. Initially, only two subject find the ill-defined lesion algorithm useful, this total decreases to one after each subject has completed three prompting sessions. The view that the system overall is useful increased marginally after the first session.

One subject believes that the system overall would not be useful, but at the same time believes that the calcification algorithm could be of use. There are two possible explanations for this view. Either the output from the ill-defined lesion algorithm detracts significantly from the overall usefulness of the system, and/or the subject believes that prompting only for micro-calcifications does not provide enough support to make use of a prompting system seem worthwhile.

It is worth considering the reasons why a film reader might not believe that the system, or its components, could be useful in screening. It is possible that their view might be based on a belief that the performance of the system is in some way inadequate. However, it is also possible that they have not developed an accurate understanding of what the system is able to achieve, and lack appropriate strategies for dealing with its shortcomings.

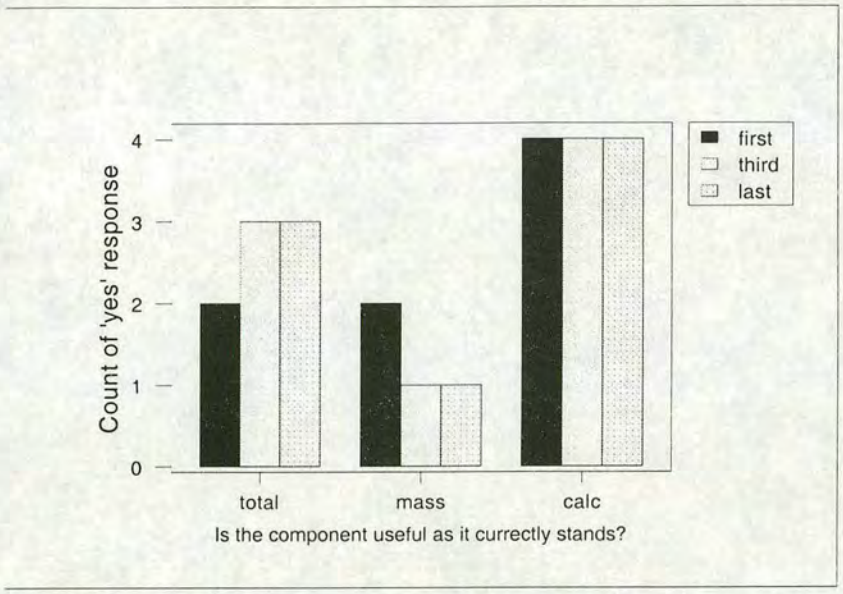


Figure 6.5: Subjects' responses when asked if the system overall and if each system component would be useful to them in a screening context as they currently stand. Yes/No responses were allowed — the chart shows the number of Yes responses given (a maximum of five Yes responses is possible). These questions were asked after each subject completed every prompted session, but for comparison, the chart only shows the responses after each subject had completed their first prompted session, after each subject had completed three prompted sessions, and for the final prompted session completed by a each subject.

6.4.2.2 Rating sensitivity

Subjects' assessment of the sensitivity of the micro-calcification algorithm appears to remain unchanged throughout the trial, with two subjects believing it to be 'Just right' and three believing it to be 'Too sensitive' (Figure 6.6). There is an apparent tendency for subjects to increasingly believe that the system is too sensitive as they are exposed to more prompted conditions. Assessment of the sensitivity of the ill-defined lesion algorithm appears to be split between those who believe that it is too sensitive, and those who believe that it is not sensitive

enough. One problem when asking this type of question is that film readers may sometimes talk about sensitivity in terms of specificity and vice versa. So if they report that an algorithm is too sensitive, they might be indicating that it is not specific enough — i.e. that it has too high a FP prompt rate. Bearing this in mind, it is possible to conclude that:

- the calcification algorithm is marginally not specific enough.
- the mass algorithm is neither sensitive nor specific enough, and that
- overall the subjects are increasingly of the opinion that the system as a whole is not sufficiently specific.

6.4.2.3 Rating specificity

Subjects appear to believe that the system overall and its components are not specific enough — i.e. that the FP prompt rate is too high (Figure 6.7). Despite this, it is interesting that subjects generally maintain their view that overall the system is useful as it currently stands (Figure 6.5). This may indicate that, although subjects believe that the FP rate could be usefully improved, they are able to cope adequately with this particular shortcoming. This is probably more true for the micro-calcification rather than the ill-defined lesion detection algorithm.

Subject B, in her third (and final) prompting session, stated her belief that the ill-defined lesion detection algorithm is too specific. It is possible that she is referring obliquely to a view that the sensitivity of ill-defined lesion is insufficient (see above).

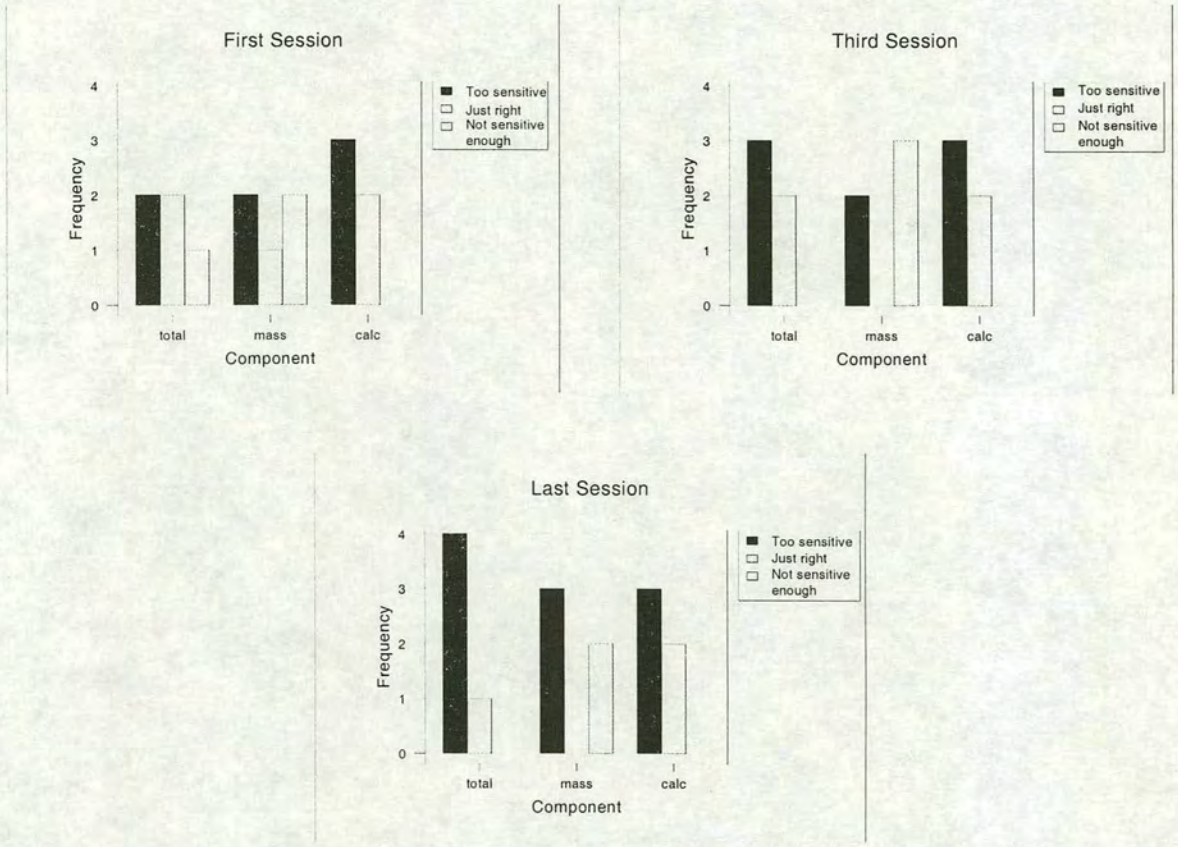


Figure 6.6: Subjects' responses when asked if the system overall and if each system component was 'Too sensitive', 'Just right' or 'Not sensitive enough' in the post-session questionnaire. The chart details responses from the first, third and final prompted sessions completed by each subject.

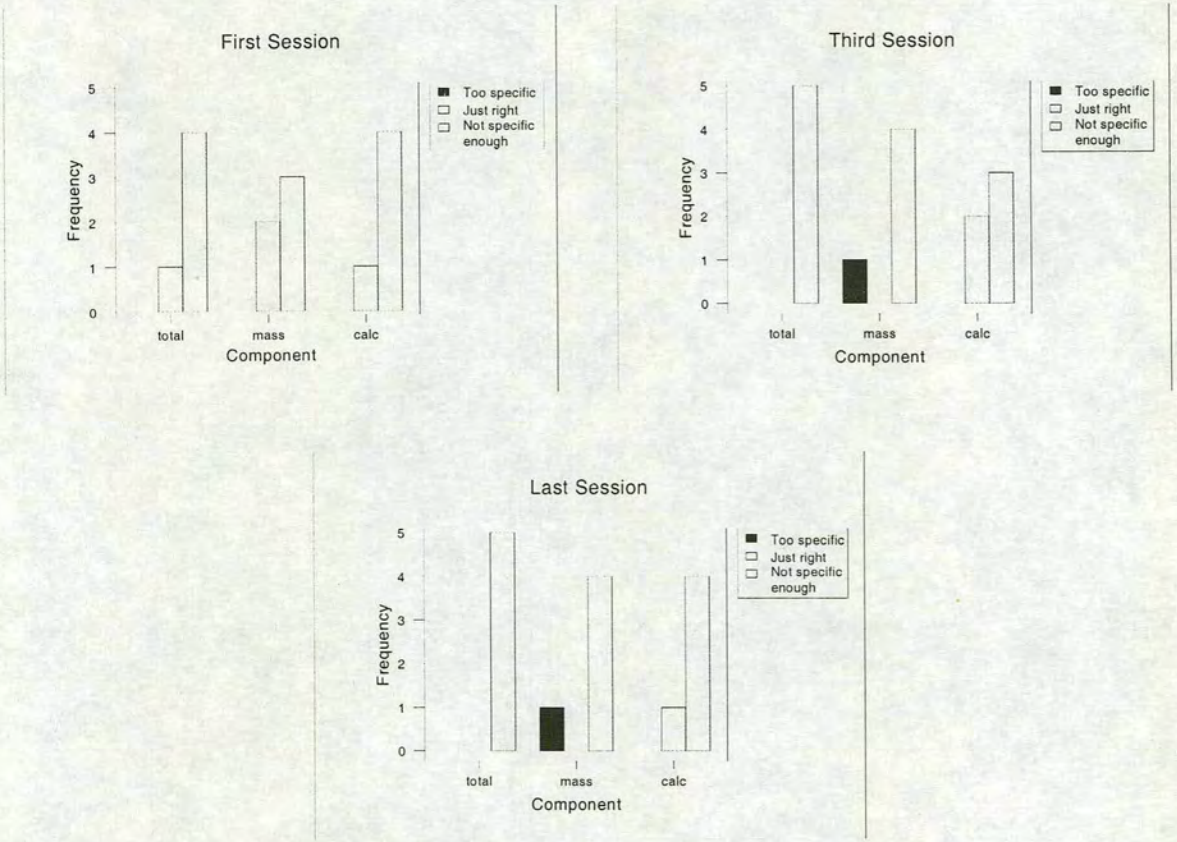


Figure 6.7: Subjects' responses when asked if the system overall and if each system component was 'Too specific', 'Just right' or 'Not specific enough' in the post-session questionnaire. The chart details responses from the first, third and final prompted sessions completed by each subject.

6.4.3 Understanding and locating prompts

After each prompting session subjects were asked to indicate roughly what percentage of time they had difficulty:

1. locating the prompted region on the mammogram, and
2. understanding why the system had prompted for a particular area.

When answering both questions subjects were asked to select from the following percentage ranges: 0%-20%, 21%-40%, 41%-60%, 61%-80% and 81%-100%.

Figures 6.8 and 6.9 show that in the main that subjects were able both to locate the prompted region on the mammogram, and to understand why a feature had been prompted for the majority of prompts. Subjects were also asked to list any instances or categories of prompts that they found particularly difficult to interpret. The four responses given to this question are shown in Table 6.6 Where some difficulty explaining prompt is expressed, this seems to be dependant on particular sessions and on particular subjects. Tables 6.7 and 6.8 detail the number of responses in each percentage range given by each film reader for all the completed prompting sessions.

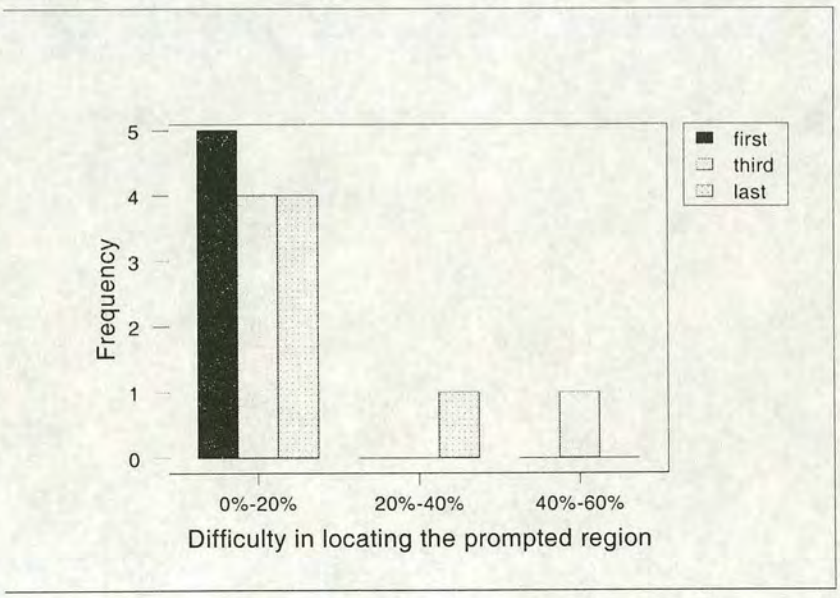


Figure 6.8: Subjects' responses when asked to approximate the percentage of prompted regions they had difficulty locating on the mammogram. The chart details responses from the first, third and final prompted sessions completed by each subject.

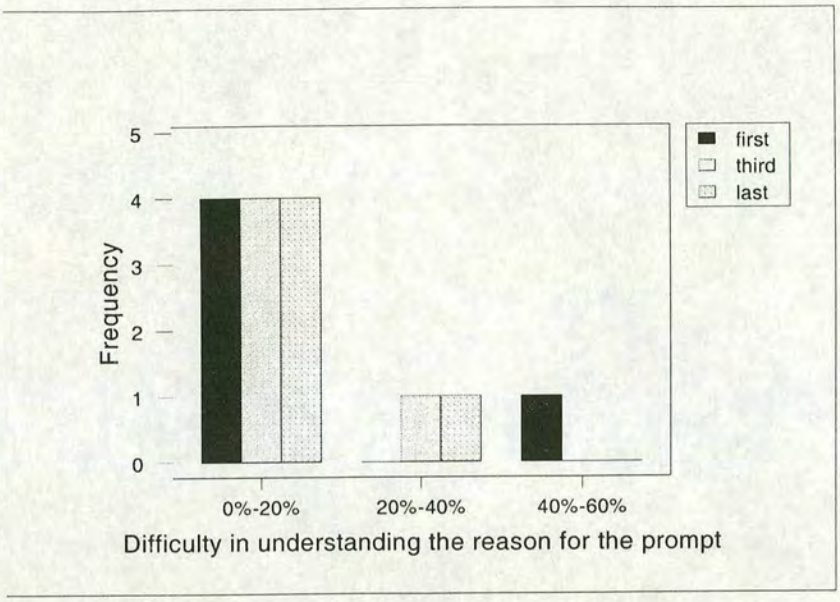


Figure 6.9: Subjects' responses when asked for what percentage of the prompts they had difficulty understanding why the system had prompted for a particular region. The chart details responses from the first, third and final prompted sessions completed by each subject.

Subject	Session	Comment
A	20	Several calc prompts where no calcs seen.
C	13	A lot of the "microcalc" prompts today I have been unable to see or 'account for' by anything else.
C	21	Elongated prompts for masses along the pectoral - where there is no obvious crossing vessel.
C	32	Some of the prompt diagrams on todays diagrams have been very dark (?printer problem) and therefore hard to see the prompted area.

Table 6.6: Comments made by subjects in the post session questionnaire about particular prompt types they had difficulty explaining. (Underlining is subjects' own emphasis.)

Subject C was the only subject to indicate that she had difficulty locating 21% or more prompts on the mammogram, and seemingly she had the greatest difficulty understanding why the system had prompted for particular areas (for 21%-40% of prompts in 3 out of her 4 prompted sessions). Subject C made the greatest number of written comments concerning being unable to understand or

Subject	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
A	3	0	0	0	0
B	3	0	0	0	0
C	1	2	1	0	0
D	3	0	0	0	0
E	7	0	0	0	0

Table 6.7: Have difficulty locating prompted regions on the mammogram

Subject	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
A	3	0	0	0	0
B	2	1	0	0	0
C	1	3	0	0	0
D	2	0	1	0	0
E	7	0	0	0	0

Table 6.8: The percentage of prompts for which the subjects have difficulty in understanding why the system has prompted for a particular area.

locate prompts, which also provide some account of the difficulties she experienced. In particular, subject C identifies micro-calcification prompts that cannot be attributed to features in the breast, and spurious ill-defined lesion prompts that cannot be explained with respect to the information given in the training sessions. Subject A also alludes to problems accounting for micro-calcification prompts, as does subject D in the free form response questions given after session 25 - "High prompt rate, particularly for calcification which I can't identify".

Although three of the five subjects have given some indication that the micro-calcification algorithm will produce unaccountable prompts, there is a considerable variation in estimates of what percentage of prompts subjects are not able to explain. Subject C, for example, indicates that this problem is more prevalent than the responses from subjects A and E might suggest. It is possible that some film readers have a greater tolerance for spurious prompts, and that with exposure to the system they are satisfied with "this is something that the occasionally does system does" as an 'explanation'. Other film readers might demand a more complete account. Unaccountable prompts might present film readers with particular difficulties (especially when they are for subtle features such as micro-calcifications) because they would have to choose from the following possibilities:

1. "The prompt is correct, but I have failed to identify the region prompted for on the mammogram". This demands a more detailed examination of both prompt and film to be certain.

2. "The prompt is correct, but I have failed to see the feature even though I am looking in the right area". This demands a more detailed search of the film to be certain.

3. "I know that there are cancers cannot be detected by radiology — but perhaps the prompting system is picking up something that I can't see.". A naive view of system capabilities — but there is an example of this type of reasoning in the discussion of the interview data (Section 6.5.2.1).

4. "The prompt is spurious, the system occasionally does this sort of thing". This explanation requires that possibilities (1) and (2) are discounted first. Acquiring the habit of using this explanation too readily might lead to 'correct' prompts being occasionally neglected.

It is possible that subject C's confidence threshold for eliminating possibilities (1) and (2) above is higher than for the other subjects. Training prior to using a system should encourage film readers to maintain an acceptable balance between the time and effort required to eliminate (1) and (2) and the necessity of assuming (1) and (2) for effective use of the system.

6.4.4 Free form response questions

The post-session questionnaire included five free response questions, namely:

- What do you think the system's strengths are?
- What do you think the system's weaknesses are?
- What irritated you most about the system?
- What aspects of the system did you find most useful?
- Can you suggest how the system might be improved?

Full details of the responses to these questions are given in Appendix B. The responses are discussed below.

In total, 10 positive statements were made about the sensitivity of the micro-calcification algorithm, compared with only three for the ill-defined lesion algorithm. Subject A suggested twice that the mass algorithm occasionally missed some lesions, whereas none of the subjects commented that the micro-calcification algorithm had made any omissions.

Suggestions for improving the algorithms included detection of stellate lesions / distortions (suggested by two subjects) and use of previous / additional films (suggested by two subjects).

Vascular calcifications and prompts for areas where no calcifications could be found were identified as particular problems for the calcification algorithm. In addition, one subject suggested that 'too many' prompts were being used to highlight an area of calcification when only one would be sufficient. Overall, subjects indicated that both the sensitivity and specificity of the system could be improved.

Subject E suggested that the system was easy to use on three separate occasions, where as subject B indicated twice that she felt the system slowed the reading process down. Both subjects B and E indicated that they felt that the effort required to dismiss multiple prompts on normal structures slowed reporting a little.

Subjects also draw attention to the dangers of over-prompting:

> (Subject B S36) Could over do it or give a false sense of security — ie by ignoring prompts if there are too many.
>
> (Subject B S38) Danger of over-prompting therefore making each less valuable.

There are perhaps two issues here: a high overall prompt rate might reduce the value of individual prompts generally, and the presence multiple of prompts on a single image may reduce the value of individual prompts in that particular case. Subject E suggested that 'multiple prompts' were both a 'system weakness' and 'irritating'.

A number of the comments indicate that the system had affected decision-making:

> (Subject B S38) Could consolidate suspicion of a particular area. Could draw attention to the 2nd lesion.
>
> (Subject C S4) It found one small cluster of m/c [2] that I hadn't spotted.
>
> (Subject C S32) Occasionally spotting a "mass" or small cluster of calcs I hadn't noticed (or dismissed — makes me think again).
>
> (Subject E S9) Potential to falsely reassure.

[2] An abbreviation for 'microcalcification'

(Subject E S11) Highlighting of areas to review, affected my decision, in a few cases, to recall.

(Subject B S38) Firming up on some questionable areas. Negative prompts could be reassuring (? over-reassuring).

The statements made by subject C provide some evidence that the system is useful as an aid for correcting detection errors. However, the responses given by subjects B and E indicate that prompts have also been used to assist with classification decisions, and they express some concern that this mode of use is potentially dangerous. Further evidence for these types of usage is available from the interview data, and is discussed in greater detail in the following section.

6.5 Interview data

A total of 18 interviews were conducted, 16 after prompted sessions and 2 after unprompted sessions. For the purposes of discussion, the interview data has been organised according to three topics:

1. effects of the system on decision-making,
2. system usage, and
3. subjects' views on system performance.

So that direct quotes and discussion of particular statements can be examined in context the transcripts are referred to directly using the following notation: (xx.yy) — i.e. 'Statement yy made during interview xx'.

6.5.1 Effects of the system on decision-making

In each of the post-prompted session interviews subjects were asked either if the system had drawn a feature to their attention that they had not themselves noticed, which they then recalled, or if the system had affected their decision-making in a more general way. Table 6.9 shows how often subjects reported one or more occasions where the prompts had effected their decision-making. Out of a total of 16 interviews held after prompted sessions, subjects indicated that their decision-making had been affected one or more times in a total of 11 of those sessions. This section discusses in more detail how subjects believed the system had contributed to their decision-making.

Subject	Decision making affected?		Prompted sessions	No. Interviews
	Yes	No		
A	2	0	3	2
B	2	1	3	3
C	1	0	4	1
D	0	3	3	3
E	6	1	7	7

Table 6.9: Shows the number of occasions subjects claimed that prompts had affected their decision making one or more times. 'No. interviews' refers to the number of interviews conducted with that subject following *prompted* sessions. Because of constraints on subjects' time, it was not possible to conduct an interview after every prompted session.

6.5.1.1 Not affecting decision-making

In order that subjects may accurately assess whether prompts have had an effect on their decision-making they need to be consciously aware of occasions where this has been the case, or have access to some other type of feedback. Perhaps the evidence most easily available to subjects during the course of this trial would be instances where the prompts had caused, or had been used as evidence during, some conscious deliberation about the status of some feature. The most obvious examples of this would be if a subject had overlooked a feature that the prompt subsequently brought to their attention, or if the presence (or absence) of a prompt had otherwise made some significant contribution to their decision to recall. However, it is also possible that the prompts may affect decision-making in ways that are not available to introspection, and therefore in ways that might go unreported in response to questions posed during interviews. Possible types of bias that prompting might 'silently' introduce into a film reader's decision-making include:

Search bias Areas that are prompted might be given a more thorough examination than other parts of the image. Thus film readers might make fewer detection errors for targeted feature types, and more for untargeted feature types. (Where 'untargeted feature types' refers both to features outside of the operational scope of the algorithms, and to the sorts of False Negatives (FNs) that the algorithms might generate).

Confidence bias The presence of a prompt might subtly increase, and the absence of prompt subtly decrease, a film reader's suspicion of an area.

In addition, the 'accuracy' of subjects' responses to interview questions will depend on their ability to take a dispassionate and objective view of their own behaviour. Subjects might be inclined to underrate the effect of the prompts if they believe that any effect is at odds with the integrity of the objective application of their skill. Conversely, they might be inclined to overrate the effects of the prompts if they believe that this outcome is of particular interest to the person conducting the interview.

Subject D was the only subject to consistently report that she was 'not aware' of any occasions where the prompts had affected her decision-making (9.5, 14.2, 15.12), but she did state that the prompts (particularly those for micro-calcifications) had made her go back and look again (9.8, 14.4). So, even though subject D believed that the prompts had had little direct effect on her decisions, they were obviously having an effect on her *behaviour*. Similarly, on the one occasion that subject B reported that she was not aware that the prompts had affected her decision-making, she stated "There were cases where it made me look again, I don't think it actually made me change my mind. But it did make me look back again".

6.5.1.2 Correcting errors of attention

The primary aim of a prompting system such as PROMAM is to correct for errors of attention. In ten of the interviews, subjects reported that on one or more occasions during that session their attention had been drawn to features that they had over-looked. These events fall into three categories:

1. features that subjects had failed to detect, which they then decided to recall,
2. features that subjects had failed to detect, which they then decided were normal, and
3. features that the subjects had failed to detect, for which the prompt seemed to contribute as much to their decision to recall as the appearance of the feature itself. (This is really a special case of (1) above, but events of this type are worthy of discussion.)

There were four reported occurrences of events in category (1). In her first session, subject A reported that the prompts had drawn her attention to a couple of recalls that she had not seen (2.15-18), although what these features were is not recorded. In subject C's first prompted session, when asked if the system had drawn her attention to anything that she had recalled, she stated that:

(3.10) C — "There was one, just at the end there, a cluster of calcium on the edge there, probably when I was beginning to lose attention."

In her forth prompted session, subject E stated that she:

(13.2) E — "[I] missed some micro-calcification (...?) the prompt, and I recalled it on the basis of the prompt. I don't know why, I think I just [suffered a] lapse of concentration."

Again, in her sixth prompted session, subject E indicates that the system had drawn her attention to some micro-calcification that she had missed:

(19.2) E — "Yeah, one, on micro-calcifications, this time actually, that I didn't see and then I brought back."

It is not known whether the recalls alluded to above were actually for cancers, however, this does not detract from the potential the system shows for drawing subjects' attention to features that they have missed. It is expected that only a subset of missed features are likely to correspond to cancers, and that the greater proportion of missed features that are suspicious enough to warrant recall will turn out to be benign. The slight increase in recall rate this entails is an unavoidable consequence of the desirable effect of assisting film reader's to make a more complete assessment of relevant features.

In the majority of cases where film readers' attention is drawn to features that they have missed it is likely that the features will either be of no consequence (adding to the overhead of system use) or of some consequence, but it will transpire that they can be dismissed as benign. The three occasions where subjects specifically commented on this type of event (events that fall into category (2) above) are discussed below. In her first prompted session subject C stated:

"I don't normally write down every time I see vascular calcification, in fact, I don't even see it, you know, it's been a couple of times when I've seen these things drawn all over the prompts, and I've looked back and I've said 'right enough, there is vascular calcifications', but I've not actually noticed it." (3.8)

When asked after her second prompted session if there were instances in the set where the system might have affected her decision-making, subject E stated:

(7.6) E — "...there was times when the prompts made me go back and look and I still decided to stay with what I thought..."

When asked in her second prompted session if the system had drawn her attention to something that she hadn't noticed before, subject A stated:

> (11.6) A — "Yes, there were a couple of cases, I think they were calcs and they were unaltered from previous."

Subject C's comment indicates that film readers may involuntarily dismiss some features that are of no consequence, and it is possible that errors in this process may contribute to errors of detection. However, it would be inefficient from a film reader's point of view if, when using a system, their attention is continually drawn to features that they have already involuntarily dismissed. A system should have sufficient specificity to dismiss the obviously benign. The consequences of prompting for benign lesions and the strategies that subjects evolved to deal with this type of prompt are discussed further in Section 6.5.2.1.

The comments made by subjects A and E indicate that there are examples of cases where the system is drawing subjects' attention to features that they have missed that are worthy of consideration, but are not sufficiently suspicious to recommend recall. This type of event is probably beneficial, as it demonstrates the system's capabilities, as well as providing film readers with the reassurance that they have made a more complete examination of the mammogram.

Reports of type (2) events seem sparse in the interview data since it is expected that the majority of missed features brought to the film readers' attention would be of this type. However, these events might be under-represented as they are possibly 'less interesting' to subjects than missed features that resulted in a recall. It is possible also that they interpret the question 'did the system draw your attention to something that you had missed' with the implied meaning 'did the system bring anything *interesting* to your attention that you had missed'.

Subject E reported four events in category (3) above (6.2-4, 7.8, 10.14, 18.3). That is, where the system had drawn her attention to a missed feature which she then recalled and where the presence of the prompt had seemingly made a significant contribution to her classification decision. Examples of subjects using prompts to assist with classification decisions are discussed in the next section, but it is worth noting that the presence of a prompt might trigger the initial suspicion, as well as contributing evidence to decision-making for a feature that a film reader identified themselves.

6.5.1.3 Assisting with classification decisions

As stated earlier, the primary aim of a prompting system such as PROMAM is to draw film readers' attention to relevant features in the mammogram to reduce the

possibility they might overlook something of significance. Systems like PROMAM are not designed to assist film readers to make classification decisions; specifically, PROMAM is not designed to:

1. help achieve a lower recall rate by assisting film readers to assess more accurately whether a feature is benign or malignant, or to

2. improve a film reader's TP rate by correcting for classification errors.

If the absence of a prompt is used as evidence of normality, then it is possible that cancers missed by the prompting system might be misclassified, thus eliminating any potential performance gains due to the assistance provided for detection. A greater degree of efficiency is possible if the actions of the system and the film reader are complementary. That is, for the film reader to pick up cancers that the prompting system has missed and vice-versa. PROMAM only gives a binary response (prompted or not prompted) — it does not indicate the likelihood that a given prompted feature actually is a cancer — thus a prompt is only a very coarse indicator of suspicion. Furthermore, PROMAM is not very specific in comparison with a film reader, so it has little value as an aid for correcting film readers' own classification errors.

If the presence of a prompt is used as evidence that a feature is a cancer this may have the effect of improving detection performance, but it would also reduce specificity. Much the same effect could be achieved without the aid of a prompting system if film readers simply lowered their confidence threshold.

During training sessions given prior to participating in the trial, subjects were advised not to use prompts as contributory evidence in their decision-making. Subjects were told that they should only use a prompt to direct their attention to areas of the mammogram which should then be evaluated using *only* the evidence available in the image. However, both the questionnaire data (see sections 6.4.4 and 6.6.4.1) and responses given in the post-session interviews indicate that subjects are inclined to use prompts to give assistance with classification.

In the interview data subjects B and E refer to occasions where they had found the absence of a prompt 'reassuring', for example:

> (5.21) Interviewer — I noticed [in a] previous questionnaire that ... you said that it was quite (comforting?) when you had cases that didn't have any prompts on, and that you were were thinking that [the lesion was benign].
>
> (5.22) B — Yes ... I think that that is reassuring. It might just be falsely reassuring sometimes.

(19.7) E — ... Sometimes that's been something that I wondered whether it would pick up, and actually I thought well I'm not got to recall that anyway, and it hadn't prompted it, so it hadn't thought it was significant either, and that's quite reassuring — reassuring a bit. Yeah, it is reassuring.

The quotes above indicate that the absence of a prompt is viewed as 'reassuring' only — the prompts merely confirming a decision that has already been made. However, quantifying the degree of influence a prompt has with respect to a particular decision is difficult, as the following quote indicates:

(5.13) Interviewer — Did any of the prompts this time make you change your mind?

(5.14) B — No. There was one that it perhaps firmed up a decision.

(5.15) Interviewer — ... Ok, so that's one that you decided to recall then?

(5.16) B — I think that I would have recalled it anyway, but it's an artificial situation here. And I'm not so sure if I would have done — I don't know...

The problem seems to be that the subject cannot form an objective view, for comparison, of what decision they might have reached in the absence of prompting information. In cases where the presence of a prompt has seemingly made a subject more inclined to recall there appears to be more certainty as to the role the prompts have played:

(20.1) Interviewer — Was there any incidences in that set where the prompts affected your decision-making?

(20.2) B — There was one where I was undecided, and it was prompted (...?) 'I will bring it back, yes'.

(20.3) Interviewer — You [decided to] bring it back, because it was prompted?

(20.4) B — ...because it was prompted yes, because otherwise I probably would have said 'oh, forget it'. Whether that's right or not I don't know.

(7.5) Interviewer — Was there instances in this set again where the system might have influenced your decision?

(7.6) E — Yes definitely. (...?) but I wanted to recall it when I saw the prompt, I could see that there was enough, there was times when the prompts made me go back and look and I still decided to stay with what I thought, but the decision to change was well, it just pushed me a little bit (further?), there's a little bit more there, go with the prompts.

It is possible that although only subjects B and E have reported these effects, other subjects may be also prone to the same effects, but are unaware that this is the case. Evidence from the post-trial questionnaire data that other subjects also use prompts to assist with decision-making is presented in Section 6.6.4.1.

Subject B recognises that this type of influence might be counterproductive (quote 5.22 above), and subject E expresses similar concerns:

(7.11) Interviewer — On your last questionnaire you said that some of the problems with the system might be that it could be falsely reassuring, in what way?

(7.12) E — I think because there was a few occasions where you would be looking at something and it's in that sort of gray area in the middle where [there's] not enough, for me anyway, to say definitely come back. And then, I was just slightly worried last time that there was you know occasions when it wasn't prompted it reassured me. Where I was obviously swaying, thinking well "I don't really need to recall that" it backed that up as well ... I don't know whether that is potentially dangerous but I guess if you know how well it performs — but not knowing how well it performs yet, I don't know.

The above quote also suggests that the presence or absence of a prompt is most likely to influence a decision when the evidence available from the image alone is ambiguous. Ensuring that all relevant features in every case are given sufficient scrutiny is not the only difficulty a film reader might encounter while reading films (although it is specifically this difficulty that prompting is designed to address). On a case by case basis film readers may encounter a variety of further difficulties, for example, they might be uncertain whether a particular feature is sufficiently suspicious to warrant recall, they might have difficulty in being certain that a woman with dense breasts has a normal mammogram, or

they may be aware that they have greater difficulty in detecting or interpreting particular feature types. It is possible that in these situations film readers will attempt to use whatever evidence that is to hand, including prompts, to resolve any ambiguity. Some of the statements made by subject E illustrate the potential for this type of usage:

> (10.10) E — Maybe it was highlighting something that I wasn't seeing in a dense breast, so that's why it needed confirmed. ... I (...?) with it you go with the prompt.

> (18.4) E — No, No, difficult area (...?). Interesting to know ... within a dense breast if it picks up an extra density — it would be very useful.

However, despite recognising that the prompts had played a role in her analysis of features, subject E also stated: "...But the majority of the time the prompts, it was easy to dismiss the prompts and my (clinical/initial?) impression dominated the proceedings" (6.2).

Subject E also draws an analogy between heightened suspicion when another film reader asks her to examine a case, and when a case is prompted by a computer system:

> (7.8) E — ...it's awful, because it's like when someone shows sets of mammogram and they'll say, [have a look at this], it's always nice for someone ... [not to] point out what they are worried about, because if [they] do, then immediately you [have a] heightened suspicion because someone else is suspicious about it. So in a way it kind of subtly increase your (way of?) looking again because you're reading such subtleties ... it can push you just a little bit if it is just borderline...

Prompting not only contributes to a subject's analysis of features of which she is already suspicious, it may also initiate suspicion of features that have either been overlooked, or dismissed as benign, for example:

> (6.3) Interviewer — Was [it] that that you hadn't examined that area, [or] that you hadn't actually identified something in there to be worried about, or that you'd seen something and then the prompt had changed your mind?
>
> (6.4) E — No, I hadn't really seen it ... it was just a little increase in density. I think it was probably normal actually, but on reflection I thought maybe it ought to be looked at.

Although it is unlikely that using a prompting system such as PROMAM to assist with classification decisions will yield any performance gains, there may other positive benefits for film readers from this mode of use. In the quotes above, subjects talk about the prompts giving 'reassurance' that a decision is appropriate. It may be that there is some psychological advantage to be gained from reducing any anxiety felt when deciding ambiguous cases.

6.5.2 System usage

6.5.2.1 Dismissing prompts

Given that the ill-defined lesion and micro-calcification algorithms prompt for many more cases that a film reader would typically recall it follows that a film reader using the system would have to 'dismiss' the majority of these prompts. Ideally film readers should give all prompts equal consideration, and only dismiss prompts after careful examination of the prompted region on the mammogram. However, it is clear that subjects develop strategies to determine the significance of prompts based on properties of the prompts themselves (their location, shape and frequency) as well as on the properties of the prompted region. For example, subject D states:

> (15.2) D — Not really. ...the more I'm getting used to the prompts you certainly do find [that you] you [get] to dismiss [them] much quicker. 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. Because obviously they tend to have multiple vascular prompts.
>
> (15.3) Interviewer — Do you sort of become more skilled in the way of following down and saying those are all lines...
>
> (15.4) D — Well you (much as?) see that there's kind of linear lines, and you can see there's lots of vascular calcifications you ignore it fairly quickly.
>
> (15.5) Interviewer — Ok, that's interesting...
>
> (15.6) D — And also I think masses again, actually, in certain situations — remember like in that problem area over the pectoral muscle and at the back — again you tend to dismiss those quickly.

Subject B states:

(20.13) Interviewer — Do you think [that you] are you able to look for any opinion about what the prompt might be for from their shape, or their size, or their position. Does that give you any information before you even look at the [mammogram]?

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

Subject E states:

(19.8) Interviewer — Is there anything in the shape or distribution of prompts that might give you a clue, as to what they might be for, and whether you can make any judgement on that?

(19.9) E — ... the ones that happen so frequently at the bottom at the edge of the film, I was thinking that it would be awful if there was a lesion there one day because sometimes it's crying wolf at that point all the time... Because sometimes you don't even bother looking — you have a quick glance down, and [it's] easy to spot the ... vascular calcifications ...

(10.6) E — Well not always, you tend just to go back and have a quick look. You have a quick look at the (mammos?) once again after you've seen the prompt, [though] not always ... If there's lots [of prompts] in particular I tend to go back just in case... which might be probably a good thing, but ... usually you don't find anything.

The quotes indicate that using this type of strategy eliminates some of the burden of false positive prompts, by reducing the amount of effort expended re-evaluating the mammogram. Subjects E and D indicate that that they might not look back as carefully, or at all, depending on their initial assessment of the significance of the prompt. As subject E remarks, there is clearly a danger that TPs might go unattended if they happen to correspond with regions or prompt types that film readers might learn to habitually dismiss. Clearly film readers have to discover an appropriate trade-off between making a priori assessments of significance, and only dismissing prompts after careful consideration of the image. That is, one that reduces the effort required to use the system without having an undue effect on their ability to detect additional cancers. Although this issue might be addressed in training, a more satisfactory solution would be to reduce

the FP rate — particularly for FP types that have regular characteristics (for example, vascular calcifications, edge of film artifacts, etc).

A further consideration is the ease that a film reader can dismiss a FP prompt when they have decided to examine the image more closely. There is some variation in how easy subjects found this task for different feature types: subjects B and E found it easier to dismiss prompts produced by the micro-calcification algorithm (10.8, 20.18) rather than those produced by the ill-defined lesion algorithm — where as subject D found the reverse to be true (9.7). For subject D the difference is in the amount of attention that needs to given to the image:

> (9.7) D — I think you can dismiss masses (it becomes a?) distance work, clearly on micro-calcifications you can always miss subtle calcification.

Given that micro-calcifications can be difficult to see it might be expected that this would be the prevailing view. Subject A indicates that some micro-calcification prompts require particular attention to dismiss:

> (2.12) A — No ... You can dismiss that very easily, you can more easily look at the false prompts for vascular calcification and dismiss it for what it is than you can for composite shadows, because you've really got to look very closely at those to make sure that that's what they are before you let them go. (From context: Meaning micro-composites that might mimic calcification particles.)

However, subject E seems to find dismissing mass prompts more problematic because of the difficulty she has sometimes in finding the correct interpretation:

> (10.8) E — The calc ones — easier to dismiss, actually — I think. The mass ones, similar actually. Some of the mass ones today I wasn't quite sure why it was prompting on — it was beginning to worry me a bit.
>
> (10.9) Interviewer — Why was it worrying?
>
> (10.10) E — Maybe it was highlighting something that I wasn't seeing in a dense breast, so that's why it needed confirmed. ... I (...?) with it you go with the prompt.
>
> (10.11) Interviewer — Were there many incidences in this set where [the prompts affected your decision]?

(10.12) E — Yes, I did call back a couple. Although I wasn't very convinced with the prompt. But because the prompt had highlighted it and I could see what it was saying, [and] I couldn't explain it particularly in the ways that you had (told us?) that it was prompting. So I thought that maybe it is prompting a mass and I wasn't convinced it was a composite shadow so (...?) to bring it back...

It seems that because some of the prompts lacked a convincing explanation in terms of FP types described in the training material, subject E was inclined to believe that the system was detecting something that she herself was unable to perceive. The subject is aware that dense breasts make detection of ill-defined lesions particularly difficult for film readers, and is therefore unable to be certain that that the mammogram is indeed normal. In reality, the prompting system will also have difficulty detecting subtle features in dense breast tissue, and so it is not necessarily advantageous to defer to the prompts under these circumstances.

6.5.2.2 Accounting for prompts

It is important that film readers are able to explain FP prompts by relating the information in the prompted part of the image with their understanding of the properties of the algorithms. In responses to interview questions subjects indicated that they were able to provide an explanation for the majority of prompts (Section 6.4.3), and similar claims are made in the interview data. Subject A stated after her first session that:

(2.20) A — ...there is always a visible reason for a prompt, it may not be the right reason, but you can see what the thing's picking up on. It's not picking up on nonsense.

After her second session subject A's comments indicated that she was able to rationalise FP prompts in terms of information given during training.

(11.1) Interviewer — There was a couple of times there when you seemed to be having trouble spotting what some of the calc prompts were?

(11.2) A — That's right, I've made a comment on that. Yes, there were several calc prompts there that I couldn't see calcs.

(11.3) Interviewer — Was there anything there that might have been in the training — some rationale for false positive or was it that you couldn't see anything at all?

(11.4) A — Probably just summation of densities I would think, but I wouldn't go too hard on that, that's my impression.

Similarly, other subjects were able to account for a majority of the prompts, although they appeared to be less convinced than subject A that an explanation was available in all cases. For example, when asked if she found it easy to discard FP prompts, subject B stated:

(4.8) B — To rationalise them? Yes, yes, I didn't find that a problem, I could see what it had done.

and later added:

(4.10) B — Sometimes, once or twice I just wondered what on earth it was picking up. But most of them it was fairly obvious what it is it has picked up, and therefore it is easy to discard it.

When asked if she could explain some or most of the prompts, subject E stated:

(6.14) E — Most of the time, there were a couple where I wasn't quite sure why there was a prompt here. But the vast majority I could understand.

Responses given in the post-prompted session questionnaires revealed that subject C was unable to find an explanation for a greater number of prompts than other subjects (Section 6.4.3). The interview data for subject C is sparse — she was interviewed after the first of her four prompted sessions, but she only expressed this difficulty in the questionnaires administered after her remaining three prompted sessions. In the interview following her first prompted session she said:

(3.4) C — I don't think I've explained 100% of those prompts, but I can explain most of them. I mean the ones that I, that don't turn out to be anything.

It is quite possible that set variations could account for subject C's change of opinion in later sessions. However, set variations do not necessarily account for the greater difficulty subject C experiences in explaining prompts when compared with other subjects. As discussed in Section 6.4.3 it is possible that film readers may view particular explanations as being more or less satisfactory, and in the

case of spurious micro-calcification prompts, they may be more or less inclined to believe that the prompt is spurious after giving the films only a perfunctory search. In support of these arguments the interview data reveals that subject C is particularly keen to account for prompts:

> (3.5) Interviewer — I saw that you were writing down lots of stuff on the record sheets, were you writing down a lot more than usual?
>
> (3.6) C — Yeah. I don't normally write down vascular calcification very much, I was doing that just so that I could show you that I had checked it and that's what I thought it was due to, the prompts. Maybe I don't need to do that, I won't bother then.

and later in the same interview:

> (3.20) Interviewer — If you were using it [the prompting system] in a screening context ... if you were using just day to day, would you feel that you would be writing notes to say why you ignored or accepted the prompts? Would you be doing that for other film readers, or for yourself?
>
> (3.21) C — Hmmm. That's a difficult one. I'm in the habit of writing down everything that I notice I think that somebody else might notice, even if we think it's something we should dismiss, like benign things .... Whether with the prompt I would want to write down that I'd explained what the prompts were, I don't know, I haven't really thought about that yet. This is the first time I've come across it.
>
> (3.22) Interviewer — As I was saying, next time we do it, you don't need to do that for this exercise, it's not a requirement for us.
>
> (3.23) C — I'm just wondering if medical-legally it might be.

It appears that that subject C is keen to explain each FP prompt — partly because of a misinterpretation of the protocol (she believed that this was expected of her), and partly because of a more general desire to provide an account of her decisions. Some of the motivation for this type of behaviour is evident in her concern about medical-legal requirements. If a film reader overrules a prompt which later transpires to be for a cancer, and there is no account to demonstrate that the film reader did examine the prompt sheet, or of their reasons for overruling the prompt, then there could be legal consequences. Furthermore, subject C also seems to require a more complete account of the evidence in the mammograms

themselves than did other subjects, she made by far the most number of requests for historical details concerning cases, and also took the longest time to complete conditions.

It is clear that any account given of system behaviour in the training material should be sufficient to meet the needs of the most demanding film readers. Although a description of FP micro-calcification prompts due to overlapping structures in the breast was given during training, either subject C did not remember this, or did not understand the implications, or perhaps she felt that it was not a satisfactory explanation for some of the prompts she encountered. It is likely that prompts will be produced that cannot be readily explained by an account of system function given by system developers, especially as the set of cases analysed by the system increases. It would seem necessary to explain to film readers that the system will sometimes produce unaccountable prompts, and also to provide a mechanism to ensure that film readers report and discuss such cases with each other and with system developers so that a rationale in terms of image properties and algorithm function can eventually be supplied.

6.5.2.3 Anticipating prompts

Subjects reported that they were able, to a degree, to anticipate which features in the mammogram would be prompted, and that these predictions could be used to reduce the number of occasions that the mammogram had to be re-examined due to FP prompts. Subjects seemed able to develop this skill relatively quickly, after her first prompted session, subject C volunteered that:

> (3.2) C — I think that I'm beginning to get so that I can guess what's going to be prompted for.

Subject B made a similar comment after her first prompted session:

> (4.11) Interviewer — Were you generally looking back to the films to check to discard it, or where you sort of remembering what was there on the films and looking at the prompts and saying 'ah it's got that'?
>
> (4.12) B — A bit of both. I sometimes look at the films and say 'I bet it's going to prompt for that' And if it did...

After her first prompted session, subject E indicates that the prompts can be predictable:

(6.6) E — Yeah, a mixture, I tended to look back at the films again but it was obvious, you know, things like (...?) calcification, vascular calcification, things like that. And some of the areas, when I looked at the films, in the areas where there was sort of confluent sort of lines coming I expected there would be a prompt there and (wasn't?) looking back. So a mixture of things really but predictable, I wasn't surprised by them. And particularly a lot of things on the edge (...?) it had prompted a lot of the time on the edge of the film. At the chest wall.

In a later session subject E suggests how this predictability is of use:

(19.4) E — 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?almost?) without looking again.

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

(15.7) Interviewer — Do you also sometimes anticipate what might be prompted?

(15.8) D — No. Well yes and no actually. Sometimes you will be surprised what it is prompting, sometimes then you're surprised that it hasn't prompted something. There were one or two bits where I thought that it would have several prompts, (for?) masses, and it didn't, (...?) getting zero...

(15.9) Interviewer — So it's swings and roundabouts on...

(15.10) D — But overall I think that you can anticipate some of the prompts, yes.

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

(20.8) B — I suppose in a way, I mean, there are certain [features] that I recognise that it prompts for that are ...

(20.9) Interviewer — So is that beforehand, you are thinking...

(20.10) B — Well both, I find myself sometimes thinking 'well, I bet it's going to prompt for that', 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.

(20.11) Interviewer — How often do you think that you are right when you say 'I think it's going to prompt for that'?

(20.12) B — I don't know, about 50% of the time.

Compared with making an a priori assessment of the significance of prompts based on their shape, location and frequency, anticipation is probably the better strategy for dismissing FP prompts the film reader has actually made an assessment based on the evidence in the mammogram. There is probably some additional cognitive burden associated with anticipation; in addition to the mental effort required for reading, a film reader must also expend effort in forming an opinion about the behaviour of the prompting system. However, repeatedly checking whether the system's output meets with expectations is in all likelihood a natural activity — it is probably by this mechanism that film readers make an initial appraisal of the effectiveness of the prompts.

Thus it would be beneficial if film readers are, to some degree, able to predict which features might be prompted on a given image. The property of predictability is dependent on the system behaving in a consistent fashion — i.e. having a high probability of prompting for particular feature types, despite small differences in appearance and/or location within the image. Consistency, in turn, is dependent on the robustness of the algorithms — i.e. that the algorithms will respond in similar ways despite small variations to their input.

Previously it has been suggested that acceptable FP prompts are those that correspond to 'candidate features' — features that readers might consider for recall (Chapter 4). If a system consistently prompted for all candidate features, and only for candidate features, then it would most likely be a highly predictable system by virtue of the high level of agreement between the system and the film reader about which features constitute a cause for concern. It is possible that systems with other, consistent, performance characteristics could, in principle, be equally predictable, if the relationship between system function and mammographic appearance is known and understood by film readers.

However, in both these scenarios, consistency of the system is dependent on the robustness of the algorithms. Image processing algorithms are deterministic

— but this does not imply that a given algorithm will respond in the same way to features in different images that film readers would classify as being the same sort of thing. This is partly because film readers are able to make use of more evidence than an algorithm in reaching a conclusion, and partly because algorithms may be sensitive to small variations in appearance. Thus different results may arise due to variations that are difficult for a human observer to perceive.

One of the goals of training is to supply a useful account of how system function relates to mammographic appearance, and in particular to highlight circumstances where system behaviour might be counter-intuitive to film readers. If an algorithm is sensitive to small changes in image properties then it is unlikely that a usefully complete account can be given, therefore a particular design goal for system developers should be to ensure that algorithms are robust in this respect.

6.5.3 Subjects' views on system performance

The questionnaire data reveals that subjects are critical of the specificity of both the ill-defined lesion and micro-calcification detection algorithms, and of the sensitivity of the ill-defined lesion algorithm (Sections 6.4.2.2 and 6.4.2.3). This section details comments made in the post-prompted session interviews that pertain to aspects of system performance.

6.5.3.1 Specificity

Two subjects commented that the system was producing 'unnecessary' or 'unreasonable' prompts:

> (4.4) B — ... I thought it prompts for a lot more than ... I think was reasonable. Not reasonable, I don't mean reasonable, [than is] necessary. ...

> (9.1) D — ...but I found that it seemed to be prompting a lot of things that I would easily dismiss. There was a lot of unnecessary prompts.

Although it is encouraging that this subject finds she can easily dismiss prompts, the view that prompts are 'unnecessary' suggests that the system is producing FPs for features other than those which readers feel they are accountable for. That is, for features that a reader would consider for recall, or would be required to scrutinise before they could be dismissed.

Subject A was involved in the interval cancer study (Section 6.2.2), and thus had prior exposure to the system:

(2.4) A — ... Another thing that puzzled me was that it seemed to me on that set that the micro-calc algorithm was throwing up more false positives than I...

(2.5) Interviewer — that you expected...

(2.6) A — expected from my limited experience so far on things that as far as I could see were not calcifications.

(2.7) Interviewer — What sort of things was it throwing up for...

(2.8) A — Mostly linear shadows crossing. You could get the region where the prompt was and you could see something there but in a few instances they were not calcs, now, all right we only saw a few in the false negative interval cancer group, but it seemed to me that the calcs prompt was more specific on that set than it was on this.

In a later session, subject A reiterated this view:

(11.2) A — That's right, I've made a comment on that. Yes, there were several calc prompts there that I couldn't see calcs.

(11.3) Interviewer — Was there anything there that might have been in the training — some rationale for false positive or was it that you couldn't see anything at all?

(11.4) A — Probably just summation of densities I would think, but I wouldn't go too hard on that, that's my impression.

and subject D made a similar claim:

(14.6) D — ... a lot of the calcification I couldn't even identify actually, but still you went back and looked at it.

Subject E suggests that the micro-calcification algorithm is over-sensitive to benign clusters, but that this is mitigated by its sensitivity to significant clusters:

(19.2) E — ... It seems very sensitive to micro-calcifications and clusters actually. I know it's sort of over-sensitive to you know sort of benign micro-calcs, but it's usually, but also it doesn't seem to miss out on the significant clusters.

6.5.3.2 Prompting for features that can be dismissed with the aid of previous films

Comments made by three of the subjects indicate that a significant proportion of FPs from ill-defined lesion algorithm correspond to features that the subjects are able dismiss with the aid of previous films. In her first prompted session subject C states:

> (3.14) C — ... What had struck me most about the whole exercise was the benefit that I have from looking at previous films — a huge benefit there.
>
> (3.15) Interviewer — Did you think you saw cases then where if the system was able to take previous films into account it would have thrown away some false positives?
>
> (3.16) C — Yep, over and over again. I think it's terribly important actually. So for the first screen it may be more helpful to us than on subsequent screens because we've got previous films. But if we don't obviously we want our attention drawn to things that we would normally perhaps dismiss if we'd had previous films, that sounds a bit Irish but see what I mean.

Subjects E indicated that if she did not have access to previous films there were cases where a prompted feature would have resulted in a recall:

> (21.2) E — ... Interesting, I was thinking at one point, you know, having not had these films on one or two occasions I might have brought back a mass that it suggested, but with the benefit of previous films it clearly was normal tissue. ...

Conversely, subjects A and B were disappointed that 'ill-defined lesions' they were only able to dismiss with the aid of previous films went unprompted:

> (5.4) B — There were one or two that I wondered why it didn't prompt.
>
> (5.5) Interviewer — Were those things that especially were of interest to you, that you would like recalled...
>
> (5.6) B — There was one particularly I noticed that I think if I hadn't had previous films then I would have recalled.

(11.8) A — there was a mass lesion I felt that it should have prompted for — but I know that the system hasn't been looking at previous, the mass was unchanged from the previous examination — but if it had been a first time off it should have been prompted.

These comments indicate how the task of interpreting a prompting system might become much more complicated if previous films are used. If some interesting feature visible on previous films goes unprompted, then subjects are currently able to determine that this is an omission by the system. However, if the system had taken information available from previous films into account, then for any omission it would be difficult to decide if:

1. the system had decided the feature was normal based on the evidence within the single view, or if

2. the system considered the feature to be suspicious on the single view, but was heavily influenced in favour of normality by the evidence in a previous film.

It is likely that this type of problem will arise whenever a classification step has the ability to discard potential prompts. Such behaviour would further mitigate against the use of systems such as PROMAM to assist with film readers' classification decisions. For example, in the current implementation of the ill-defined lesion algorithm, regions corresponding to potential lesions are identified and are fed into a classifier — those regions deemed to be sufficiently suspicious are prompted for. However, the film reader has no way of telling if the system has made an omission because a region was incorrectly identified, or because the classifier decided that the region was normal. It is erroneous to believe that an absence of a prompt implies that a system has decided a region is normal based on a complete treatment of that region.

Subject E expresses a similar concern about a system making use of previous films for additional evidence:

(21.6) E — It's probably a dangerous thing to do because ... over a long period of time, something that is a cancer can change so slowly, but just in a small way, and for some reason you just decide to bring it back — something has subtly changed in it, so you couldn't do that I think. Because sometimes we review an opacity that has sort of been benign before, and something very slightly might have changed with

it — it's size for instance very slightly and (...?) I think that wouldn't be a good idea.

There are many technical difficulties associated with engineering an algorithm that can make a meaningful comparison between current and previous round films. Subject E also suggests that this might not even be a good idea in principle. Her argument is that the blanket application of a rule of thumb (that if a potential lesion were there before, it is not significant), might mislead film readers in cases where the exception to the rule is important (if there are 'subtle' changes). This indicates a particular view about which roles belong to the prompting system and which to the film reader — specifically subject E is suggesting that this type of discrimination should be left entirely to a film reader. The problem with this argument is that it is difficult to see what the objective effects of such a strategy might be until it is tried. A counterargument can easily be made: it might be the case that significant features that change only subtly between rounds are rare and that not prompting for such features has only a slight effect on their non-detection — but using previous films may lead to a huge gain in the *overall* effectiveness of the system due to a reduced FP rate for a given TP rate. This would allow operating at some tolerable FP rate with a higher sensitivity. Furthermore, it is clear that film readers do not believe that the ill-defined lesion algorithm has a sufficiently good TP or FP rate as it is currently implemented, and it is not entirely clear how its performance might be improved without resource to some of the additional information that is available to film readers from previous or multiple films.

6.5.3.3 Sensitivity

In addition to the claims made in the questionnaires that the ill-defined lesion detection algorithm lacked sufficient sensitivity, during interviews subjects related occasions where the system failed to prompt for what they believed to be significant opacities. Some of these are described above in the discussion of where the system had failed to prompt for lesions that the subjects had been able to dismiss with the aid of previous films, the remainder are discussed below.

Subject A stated in her first prompted session that the ill-defined lesion algorithm "...missed two masses..." (2.22) and in her second prompted session that "There was one case where there was a mass that had enlarged from the previous examination that I recalled and it hadn't picked up" (11.8). After the first of these claims the subject was asked if she expected prompts for the missed features, given the operational scope of the algorithm as described in the training

material. She responded by stating that they were masses (as opposed to spiculated lesions or areas of distortion) and that while one was close to the upper, and one close to the lower size limit for detection by the ill-defined lesion algorithm, she believed the sizes to be within the algorithm's operational scope (2.23-26). Similarly for the second claim, she again expressed the belief that the system should have detected the lesion by stating that it was within the appropriate size range "...it was still within the 3.5cm by the way. It was only about 15 mil" (11.8).

Subject D stated that "There was one (case) actually where there was definitely a cancer on the left and (questionable?) cancers in the right and it did(n't?) actually prompt the right".

It seems reasonable to infer that there are categories of features, which may be more or less suspicious, that subjects strongly expect to be prompted for — both because they believe that this category of feature should be brought to their attention, and because that feature type is within the operational scope of the system. (These 'missed features' were not always particularly suspicious, for example, subject A described one missed feature as "...a benign mass, ah a probably benign mass, but you couldn't tell that from the films" (2.24)). The quotes indicate that at least some of the time subjects are using the content of the training material to inform their judgement about the operational scope of the algorithm, and that these system 'failures' are used as evidence to inform their overall view of the system's performance.

6.6 Pre- and post-trial questionnaires

This section details responses given in the pre- and post-trial questionnaires. The questions asked fell roughly into four categories:

1. the desirable attributes of a prompting system,
2. the perceived ease of detection and interpretation of different feature types,
3. the potential role of a prompting system in screening and
4. an appraisal of subject and system performance during the trial.

These categories are dealt with in turn. In addition, reference is made to similar questions asked of film readers at five of the six potential trial centres administered as part of the study described in Chapter 3.

6.6.1 The desirable attributes of a prompting system

6.6.1.1 Responses to particular FP types

Subjects were asked to rate nine possible FP types on a five point scale from Useful (1) to Distracting (5), in both the pre- and post-trial questionnaire. The results are shown in Figure 6.10.

Changes in opinion seem largely subject dependant, with subjects A and D tending to be more favourably disposed towards these FP types after completing the trial, and subjects C and E becoming less well disposed. Changes in subject B's opinion were more evenly distributed. This result could be due to differing initial expectations, and/or differing tolerances to FP prompts. The only responses that appeared to be dependant on feature type rather than on subject were those for 'well-defined masses' and 'composite shadows'. Subjects tended to view these FP types as more distracting than useful after the trial compared with before — perhaps indicating that they contributed to the FP burden to a much greater extent than initially expected.

Subjects were also asked in both the pre- and post-trial questionnaire to prioritise the same nine possible FP types in the order in which they should be removed from a prompting system. The results are shown in Table 6.10 and the FP types in ranked order are summarised in Table 6.11. Overall, changes in priority are not statistically significant, however, eliminating prompts for composite shadows and glandular structure were given a higher rating after subjects had been exposed to the system. This could indicate an increasing recognition of ill-defined lesion FP prompts, or prompts for calcification clusters where nothing can be identified on the mammogram, as problematic. These results clearly show that eliminating prompts due to vascular calcifications and artifacts to be a priority. There are particular dangers associated with prompting for vascular calcifications and artifacts because of the frequency of the former and the regularity of the latter (See discussion in Section 6.5.2.1).

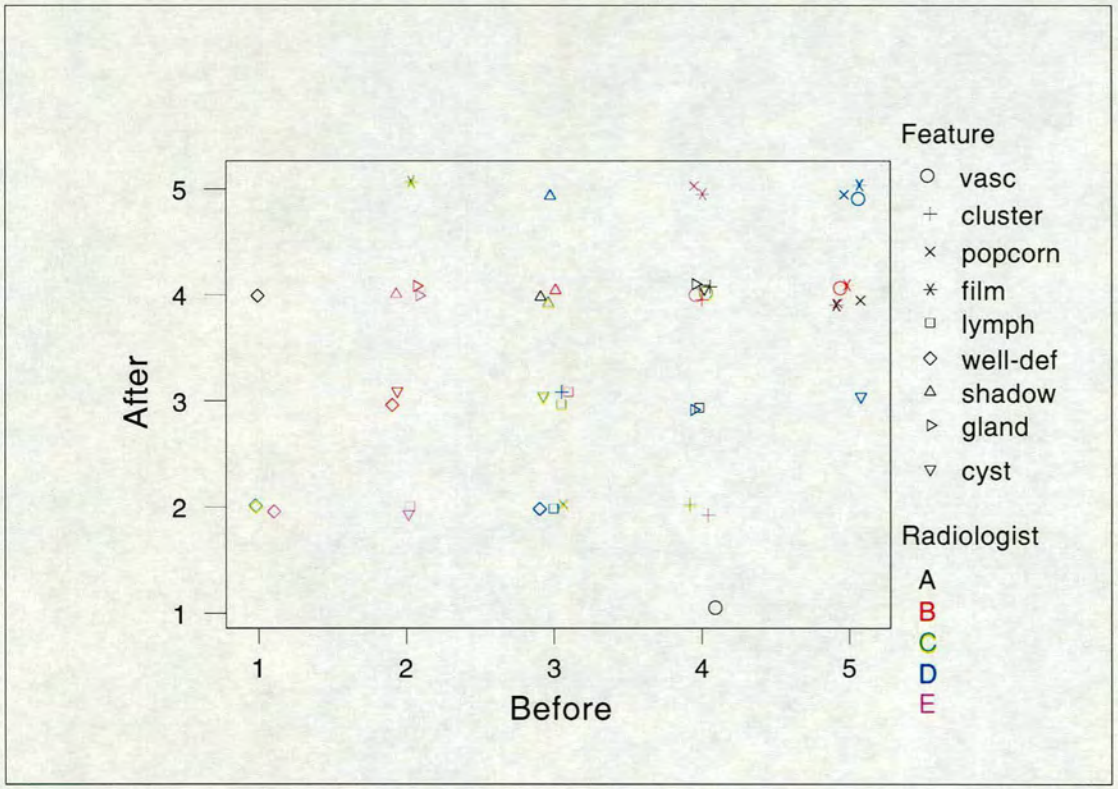


Figure 6.10: Subjects' rating of types of FP prompt, before and after the experiment, on a five point scale from Useful (1) to Distracting (5).

Feature type	Mean priority score		Ranking	
	Before	After	Before	After
Vascular calcification	1.6	2.4	1.0	1.0
Benign clusters	6.0	5.7	6.5	7.0
'Popcorn' Calcification	3.6	4.8	3.0	4.0
Film artifacts	2.0	3.4	2.0	2.0
Lymph nodes	4.8	5.6	4.0	6.0
Well defined masses	8.0	6.0	9.0	8.5
Composite shadows	6.4	5.2	8.0	5.0
Nodular glandular structure	5.4	4.25	5.0	3.0
Cysts	6.0	6.0	6.5	8.5

Table 6.10: The priority for removal of particular categories of false positive prompts given by subjects both before and after the experiment.

Rank	Before	After
1.0	Vascular calcifications	Vascular calcifications
2.0	Film artifacts	Film artifacts
3.0	'Popcorn' calcification	Nodular glandular structure
4.0	Lymph nodes	'Popcorn' calcification
5.0	Nodular glandular structure	Composite shadows
6.0		Lymph nodes
6.5	Benign clusters / Cysts	
7.0		Benign clusters
8.0	Composite shadows	
8.5		Cysts / Well defined masses
9.0	Well defined masses	

Table 6.11: The priority for removal of particular categories of false positive prompts given by subjects both before and after the experiment.

6.6.1.2 Desirable system functions

Subjects were asked to rate six possible functions a prompting system might possess as either 'Essential', 'Useful', 'Doubtful' or 'Of no use' — the results are shown in Figure 6.11. Figure 6.12 shows the results from the same questions asked in the 'clinic survey'. The results suggest that there was little change in opinion before the trial compared with subjects who had completed all conditions.

Film Readers participating in this trial seem to rate the usefulness of prompting for ill-defined lesions and calcification clusters more highly than film readers from other clinics. It is possible that the former have a bias in this respect because of their involvement in the development of the PROMAM system.

There appears to be a broad consensus between all film readers that prompting for distortions would be of particular use.

Subjects in this trial, and other film readers, were also asked to prioritise the same six possible prompting system functions in the order that they should be developed. The results are shown in Tables 6.12 and 6.13, the priorities given are summarised in Tables 6.14 and 6.15.

Overall, the ranking of possible system functions by subjects is not significantly different before compared with after exposure to the system. As with the previous question, film readers in other clinics tend to rate prompting for distortion more highly than those participating in this trial. Responses given by subjects and other film readers show a large difference between the three most highly rated and the three least highly rated functions. That is, prompting for distortions, calcifications and ill-defined lesions appear to be viewed as being more important than prompting for asymmetry or having some classification capability. It is not surprising that prompting for asymmetry is given a low priority as it is rare that features are picked up on the strength of asymmetry alone. However, it does indicate that film readers are more favourably disposed towards aids to improve sensitivity rather than specificity.

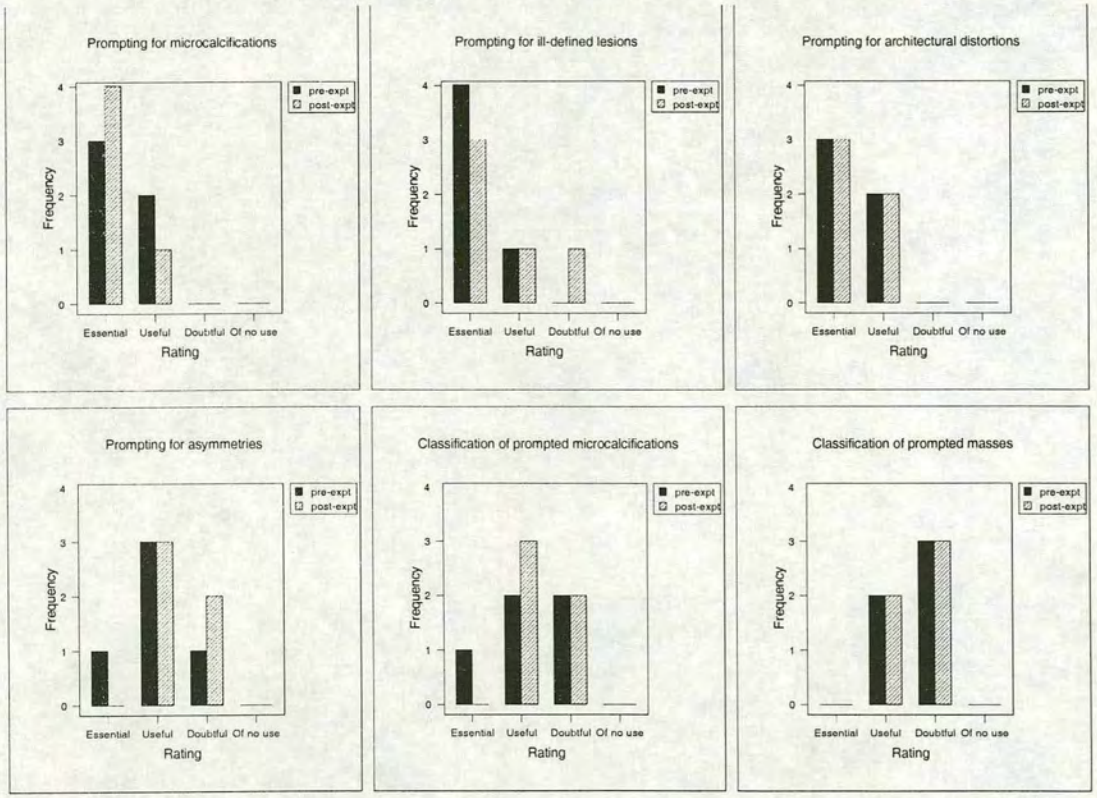


Figure 6.11: Subjects' rating of possible properties of a prompting system, before and after the experiment, on a four point scale from 'Essential' (1) to 'Of no use' (4).

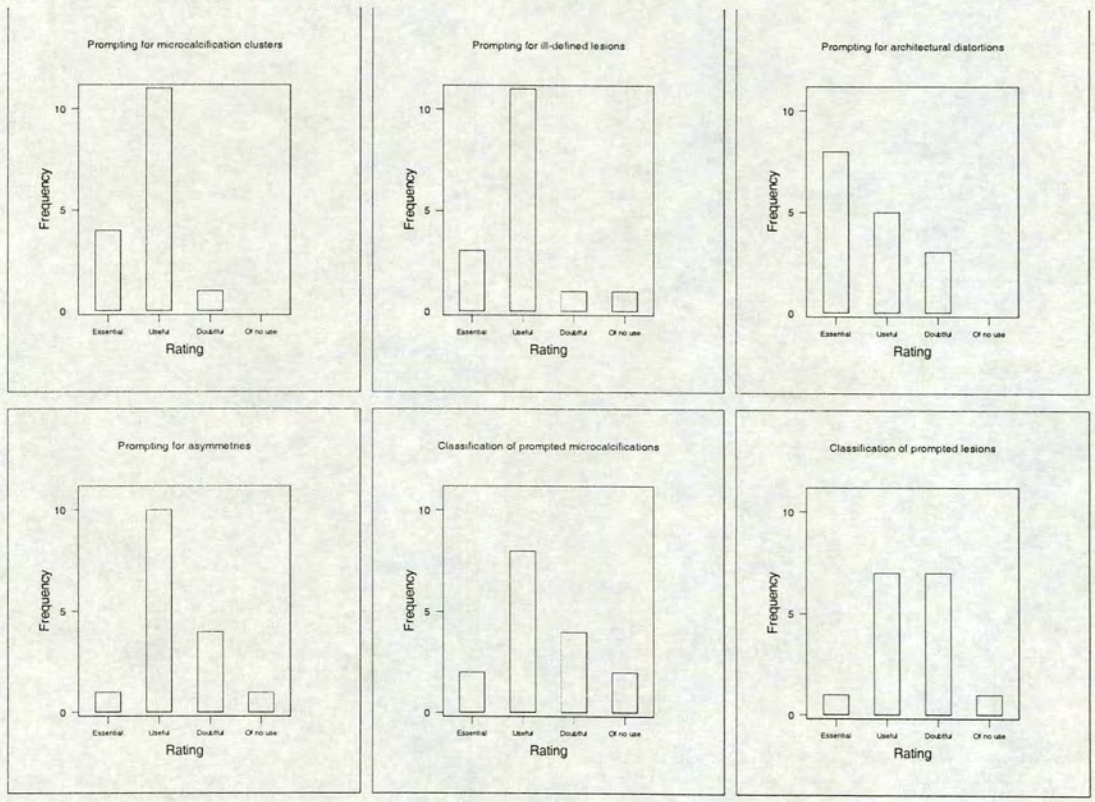


Figure 6.12: Radiologists' rating of possible properties of a prompting system on a four point scale from 'Essential' (1) to 'Of no use' (4). Clinic survey.

Capability	Mean priority score		Ranking	
	Before	After	Before	After
Prompting for micro-calcifications	1.8	1.8	1.0	1.0
Prompting for ill-defined lesions	2.4	2.4	2.0	3.0
Prompting for distortions	3.4	2.0	3.0	2.0
Prompting for asymmetry	4.8	4.6	5.5	4.5
Classification of calcifications	3.8	4.6	4.0	4.5
Classification of masses	4.8	5.6	5.5	6.0

Table 6.12: The priority given by subjects for the development of possible prompting system capabilities, both before and after completing the experiment.

Capability	Mean priority score	Rank	Responses	Adjusted mean	Rank
Prompting for micro-calcifications	2.938	2.5	16	2.938	2.5
Prompting for ill-defined lesions	2.938	2.5	16	2.938	2.5
Prompting for distortions	2.125	1.0	16	2.125	1.0
Prompting for asymmetry	3.937	5.0	16	3.937	5.0
Classification of calcifications	3.733	4.0	15	3.844	4.0
Classification of masses	5.200	6.0	15	5.219	6.0

Table 6.13: The priority given by radiologists for the development of possible prompting system capabilities. Clinic survey. Some radiologists completing this questionnaire left some statements unscored. In this circumstance the unscored items were given a score corresponding to an average of the remaining ranks. The 'Adjusted Mean' was then calculated using these derived rankings.

Rank	Before	After
1.0	Prompt for micro-calcifications	Prompt for micro-calcifications
2.0	Prompting for ill-defined lesions	Prompting for distortions
3.0	Prompting for distortions	Prompting for ill-defined
4.0	Classification of calcifications	
4.5		Prompting for asymmetry / Classification of calcifications
5.5	Prompting for asymmetry / Classification of masses	
6.0		Classification of masses

Table 6.14: The priority given by subjects, before and after completing the experiment, for the development of possible prompting system capabilities.

Rank	System capabilities
1.0	Prompting for distortions
2.5	Prompting for ill-defined lesions Prompting for micro-calcifications
4.0	Classification of calcifications
5.0	Prompting for asymmetry
6.0	Classification of masses

Table 6.15: Importance of prompting system capabilities in order of radiologists' ranking

6.6.1.3 Desirable performance characteristics

Subjects were also asked to rate statements describing five possible prompting system configurations, each having have different performance characteristics, on a five point scale from 'Most useful' (1) to 'Least Useful' (5) in both the pre- and post-trial questionnaire. These possible systems are outlined below:

1. 'High prompt rate, where most of the features prompted for are benign, but with a high probability that any malignancies will also be prompted for' (Figure 6.13).

2. 'Low prompt rate, where few of the prompts are for benign features, but with a high probability that some malignancies will be missed by the system' (Figure 6.14).

3. 'A system which is designed to prompt for micro-calcification clusters (whether malignant or benign) but not other types of calcification (eg vascular calcification, popcorn calcification)' (Figure 6.15).

4. 'A system that will prompt for all types of calcification clusters, rather than one that tries to discard those with benign appearance' (Figure 6.16).

5. 'A system that will prompt for opacities that can usually be dismissed by film readers with the aid of previous films or multiple views (eg composite shadows), as well opacities that are the result of a malignant process' (Figure 6.17).

The responses to the first statement (Figure 6.13) indicate a shift of a opinion following exposure to the system towards the belief that a system with a high sensitivity and a relatively poor specificity would be useful. Also, there is more agreement between subjects on this issue in the post-trial questionnaire.

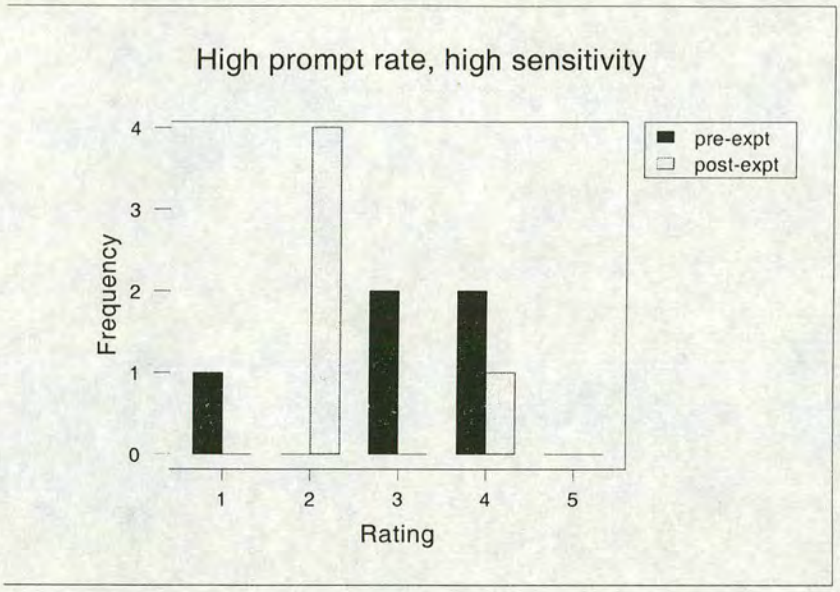


Figure 6.13: Rating of the question 'High prompt rate, where most of the features prompted for are benign, but with a high probability that any malignancies will also be prompted for' on a five point scale from Most useful (1) to Least useful (5).

Similarly, responses to the second statement (Figure 6.14) indicate a broad consensus that a system with a low sensitivity but a relatively high specificity would be less useful. The responses to both these statements agree with the model we have for the design of a prompting system: that it should provide a high sensitivity at the expense of specificity, leaving classification decisions to the film readers.

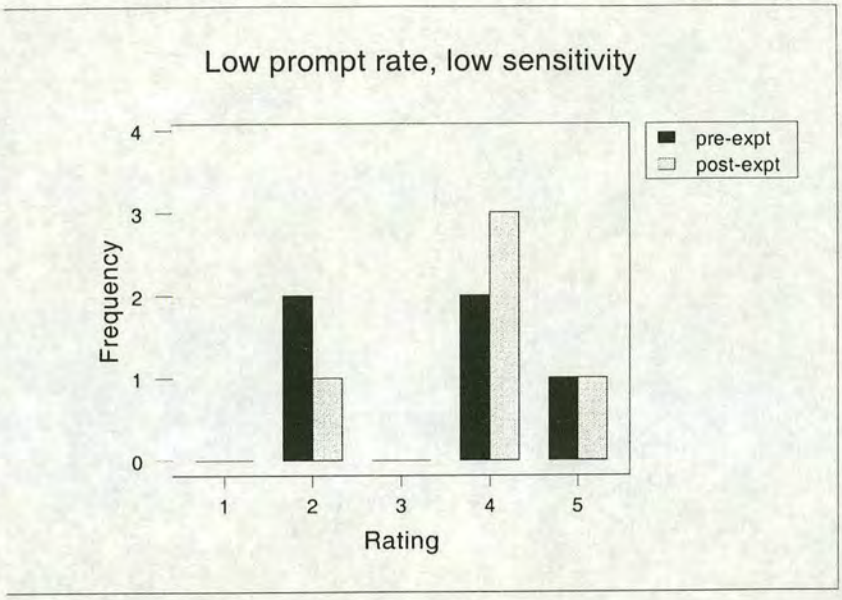


Figure 6.14: Rating of the question 'Low prompt rate, where few of the prompts are for benign features, but with a high probability that some malignancies will be missed by the system' on a five point scale from Most useful (1) to Least useful (5).

The response to the third statement (Figure 6.15) indicates that subjects would be keen for a system to have a high specificity in particular respects — ie it should not prompt for features that film readers would classify as obviously benign.

Responses to the fourth statement show a similar trend (Figure 6.16) — in the post-trial questionnaire subjects are more favourably disposed towards a micro-calcification prompting system that provides some degree of interpretive capability. When compared with responses to the third statement, subjects' opinions are more widely spread, however, a system that distinguishes between obviously benign calcification (vascular and popcorn calcification) and less obviously benign calcification (clusters) could be viewed as less specific than one that attempts to discard prompts for some of the less obviously benign calcification (some of the clusters). While subjects might prefer a system that ignores obviously benign calcifications, they may want to maintain their prerogative of relying on their own judgement on less obviously benign cases.

Responses to the fifth statement (Figure 6.17) show a shift of opinion in the post-trial questionnaire towards the view that it would be useful not to prompt for opacities that could be dismissed with the aid of previous films. However,

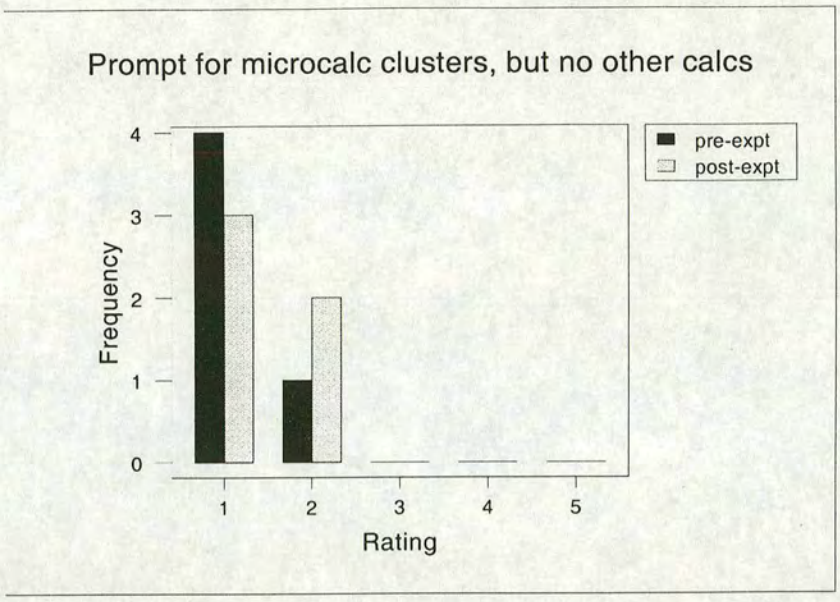


Figure 6.15: Rating of the question 'A system which is designed to prompt for micro-calcification clusters (whether malignant or benign) but not other types of calcification (eg vascular calcification, popcorn calcification)' on a five point scale from Most useful (1) to Least useful (5).

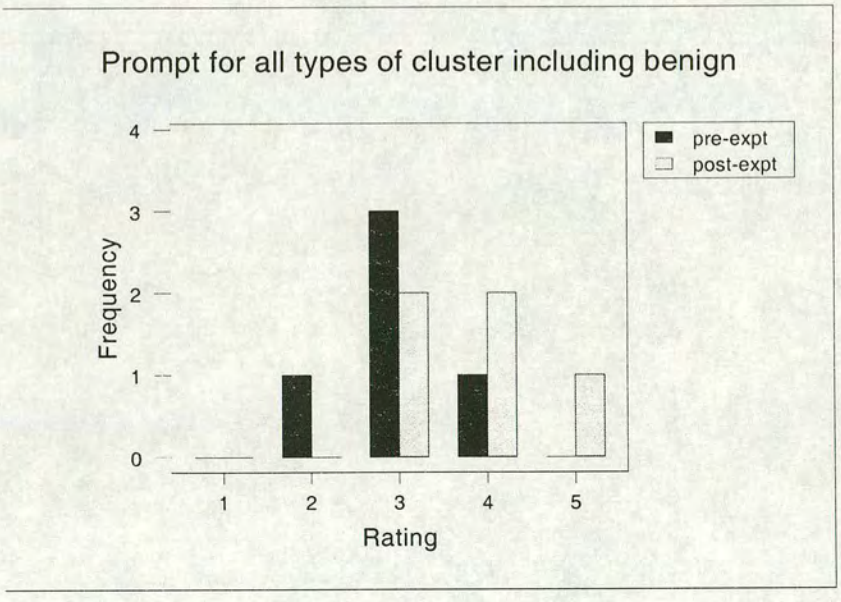


Figure 6.16: Rating of the question 'A system that will prompt for all types of calcification clusters, rather than one that tries to discard those with benign appearance' on a five point scale from Most useful (1) to Least useful (5).

although opinion has shifted, the distribution of opinion after exposure to the system shows an overall equivocal view as to the usefulness of a system that does not discount this type of feature. This change in opinion possibly reflects the observation made by some of the subjects that a large number of FP prompts produced by the ill-defined lesion algorithm can be discounted in this way (see discussion in Section 6.5.3.2).

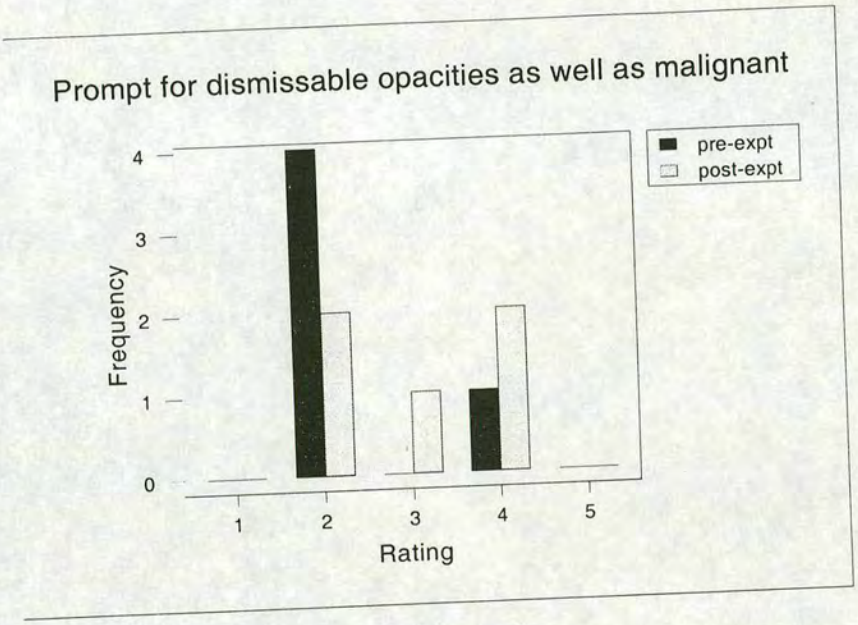


Figure 6.17: Rating of the question 'A system that will prompt for opacities that can usually be dismissed by radiologists with the aid of previous films or multiple views (eg composite shadows), as well opacities that are the result of a malignant process' on a five point scale from Most useful (1) to Least useful (5).

6.6.2 The perceived ease of detection and interpretation of different feature types

Subjects were asked in the pre- and post-trial questionnaire to rate the following reading tasks as either 'Very easy', 'Easy', 'Neither easy nor difficult', 'Difficult', and 'Very difficult':

- Detection of micro-calcification clusters
- Detection of ill-defined lesions
- Detection of architectural distortions
- Detection of asymmetries

- Classification of micro-calcifications
- Classification of ill-defined lesions

The results are shown in Figure 6.18. This question was also asked of film readers completing the clinic survey, these results are shown in Figure 6.19.

Generally the results show an agreement between subjects' responses in the pre- and post-trial questionnaire and also between subjects and film readers completing the clinic questionnaire. Classification of micro-calcifications stands out as being viewed as a relatively more difficult task by all film readers.

6.6.3 Potential roles of a prompting system in screening

Subjects were asked to prioritise six possible 'roles' a prompting system might perform in a screening practice, the results are shown in Table 6.16. Film Readers completing the clinic survey were also asked to do this and their responses are shown in Table 6.17. The rank order of each of the roles are summarised in Tables 6.18 and 6.19 for film readers participating in this trial and the clinic survey respectively.

On visual comparison there are no overall differences in the responses given by the film readers in the pre-trial questionnaire compared with their responses in the post trial questionnaire, or with the responses given by film readers in the clinic survey. However, there is a large difference in the mean scores for the three most highly ranked roles compared with the three least highly ranked roles — a similar pattern is evident in responses given in the clinic survey. This would indicate that film readers are primarily concerned with the effects a system might have on detection performance, as opposed to possible roles in addressing resource limitations, training, or improving specificity (the latter was least highly rated).

One particular difference in the priorities given by subjects after exposure to the system was that 'Reducing interval cancers' moved from being ranked first to third, effectively changing places with 'Improving the detection performance of a single reader'. If, when using the system, subjects feel more confident about their performance, they might, on reflection, view this as a considerable benefit in itself. (Improving the performance of a single reader in a double reading practice does not necessarily entail a reduction of interval cancers.)

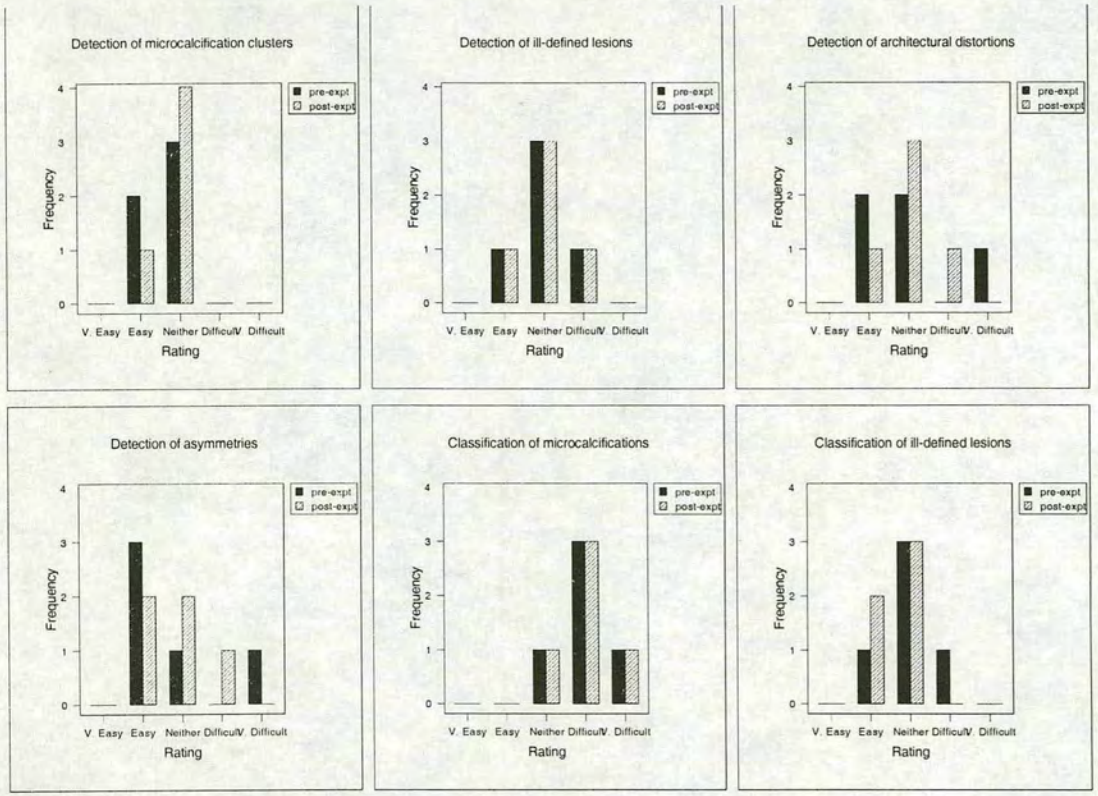


Figure 6.18: Subjects' rating of different reading tasks, before and after the experiment, on a five point scale from 'Very easy' (1) to 'Very difficult' (5).

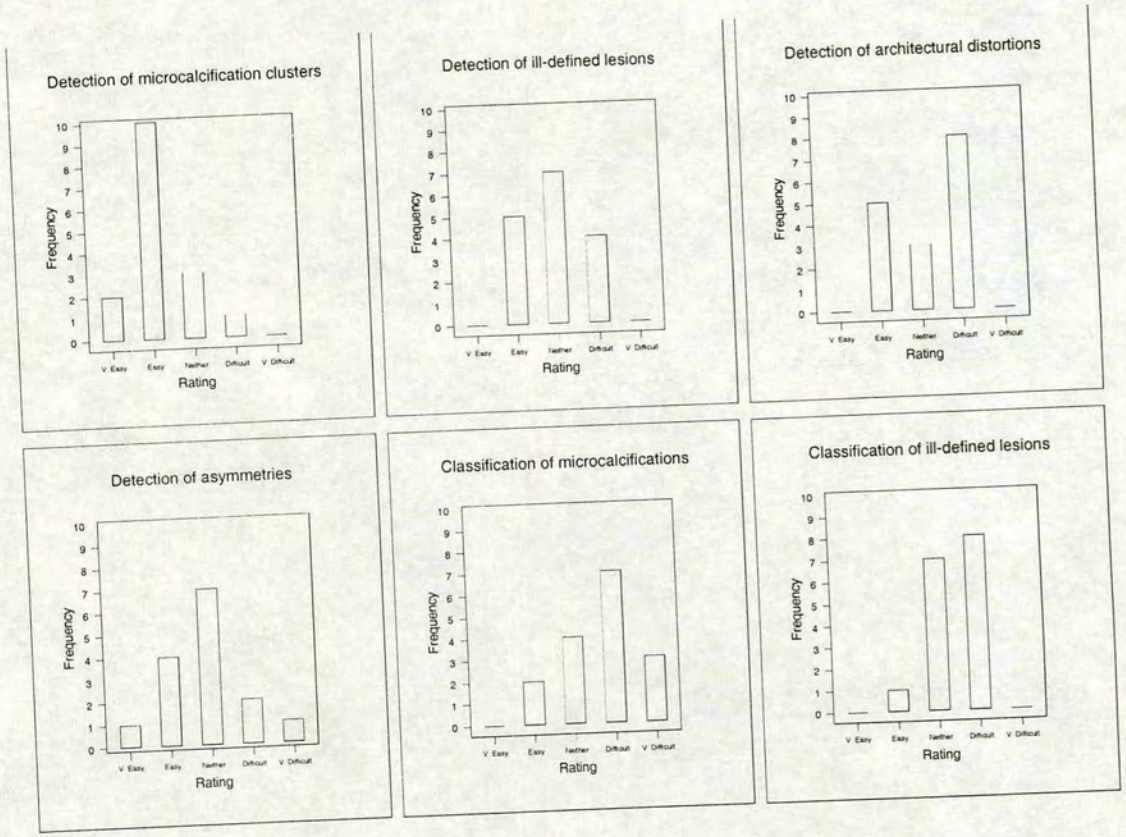


Figure 6.19: s' rating of different reading tasks on a five point scale from 'Very easy' (1) to 'Very difficult' (5). Clinic survey.

Problems prompting might address	Mean priority score		Ranking	
	Before	After	Before	After
Reducing the number of interval cancers	2.0	2.8	1	3
Improving the detection performance of a single reader	2.8	1.8	3	1
Improving the consistency of reading	2.4	2.4	2	2
Supporting inexperienced radiologists	4.6	4.2	5	4
Addressing resource limitations	4.6	4.6	5	5
Reducing recalls	4.6	5.2	5	6

Table 6.16: The priority given by subjects, before and after completing the experiment, to possible roles a prompting might address in a screening practice.

Problems prompting might address	Mean priority score	Rank	Responses	Adjusted Mean	Rank
Reducing the number of interval cancers	1.857	1	13	2.063	1
Improving the detection performance of a single reader	2.286	2	14	2.437	2
Improving the consistency of reading	2.857	3	14	2.938	3
Supporting inexperienced radiologists	4.643	5	14	4.500	5
Addressing resource limitations	4.462	4	13	4.406	4
Reducing recalls	4.667	6	12	4.656	6

Table 6.17: The priority given by radiologists, as to possible problems a prompting might address in a screening practice. Clinic survey. For an explanation of the 'Adjusted Mean', see table 6.13

Rank	Before	After
1.0	Reducing the number of interval cancers	Improving the detection performance of a single reader
2.0	Improving the consistency of reading	Improving the consistency of reading
3.0	Improving the detection performance of a single reader	Reducing the number of interval cancers
4.0		Supporting inexperienced radiologists
5.0	Supporting inexperienced radiologists / Addressing resource limitations / Reducing recalls	Addressing resource limitations
6.0		Reducing recalls

Table 6.18: The priority given by subjects, before and after completing the experiment, to possible roles a prompting might address in a screening practice.

Rank	Problems prompting might address
1	Reducing the number of interval cancers
2	Improving the detection performance of a single reader
3	Improving the consistency of reading
4	Addressing resource limitations
5	Supporting inexperienced radiologists
6	Reducing recalls

Table 6.19: Possible problems a prompting system might address in order of readers' ranking. Clinic survey.

Subjects and film readers completing the clinic survey were also asked whether, in their clinic, they envisaged a prompting system being used to enhance double reading, or to replace it with a single reader using a prompting system. The results are shown in Figures 6.20 and 6.21 respectively.

This question was asked in clinics where cases are predominantly double read, so it is unsurprising that the majority of respondents believed that prompting should be used to enhance rather than replace double reading — it is unlikely that they would be willing to sacrifice their 'best practice' until a system is proven to their satisfaction. Some respondents qualified their position by stating that the system should be used to enhance double reading 'at first'.

6.6.4 Appraisal of subjects' and system performance during the trial

Subjects were asked a series of questions concerning how use of the prompting system might have affected their decision-making, and on their view of the system's performance.

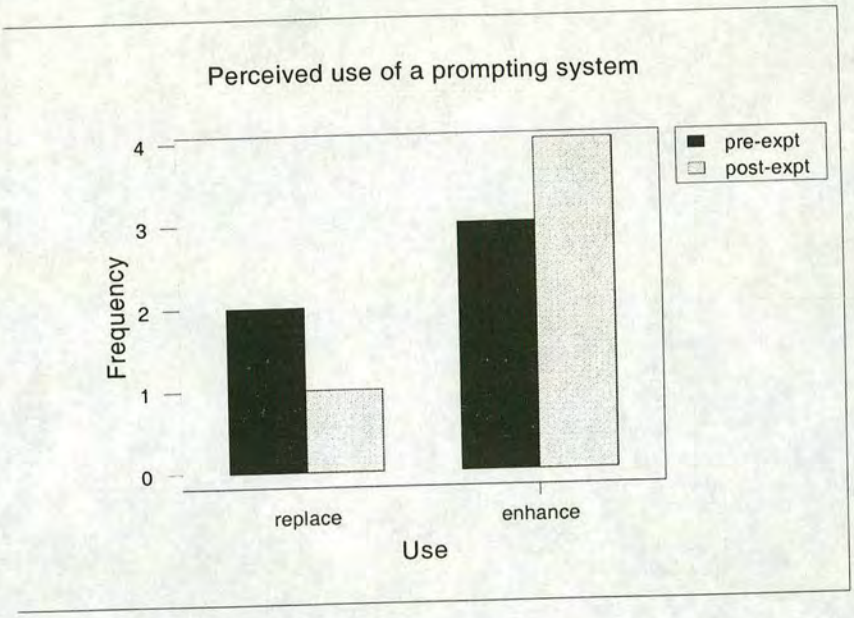


Figure 6.20: How subjects envisage a prompting system being used in their clinic, either to replace double reading with a single prompted reader, or to use the system to enhance double reading.

6.6.4.1 Effects of prompting for ambiguous features

Subjects were asked to rate their agreement with the following questions as 'Strongly agree', 'Agree', 'Uncertain', 'Disagree' or 'Strongly disagree' in both the pre- and post-trial questionnaire:

In cases where you are unsure, do you believe that:

1. The presence of a prompt will make you more likely to recommend recall?

2. The absence of a prompt makes you less likely to recommend recall?

The results are shown in Figures 6.22 and 6.23 respectively.

There is broad agreement with the first question both before and after exposure to the prompting system, with only one subject changing their opinion from 'Uncertain' to 'Agree'. This is perhaps not remarkable — if there is uncertainty in interpretation it might be expected that the default position would be to recall.

In the post-trial questionnaire, there is a consolidation of opinion in response to the second question, with subjects being more likely to believe that the absence of a prompt might influence their decision-making. This might be viewed as counter-intuitive if subjects are inclined to recommend recall when they are uncertain, however, it is entirely consistent with the hypothesis that subjects are using prompts to assist with their classification decisions.

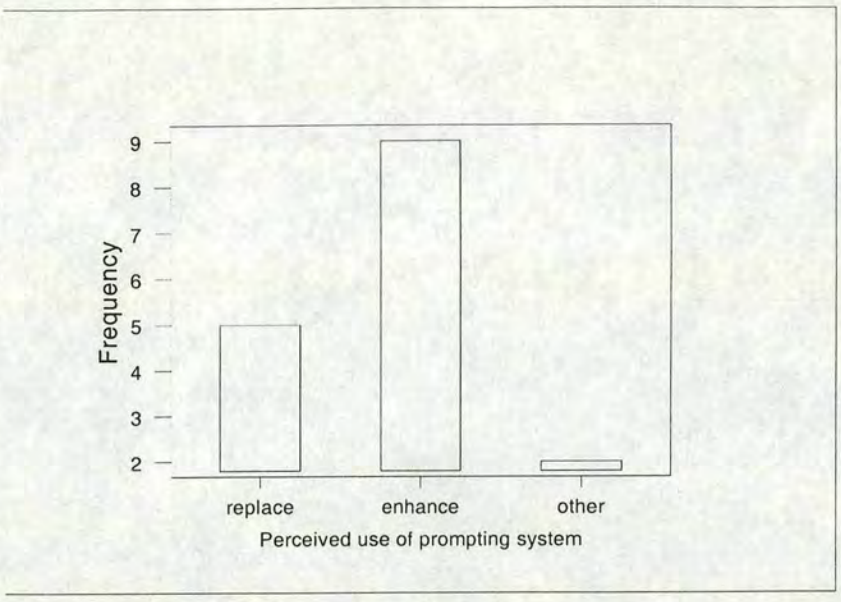


Figure 6.21: How radiologists envisage a prompting system being used in their clinic, either to replace double reading with a single prompted reader, or to use the system to enhance double reading. Clinic survey. Two respondents indicated 'other'. A follow-up question asked them to specify what 'other' role a prompting system might fulfil. One suggested that use of a prompting system might encourage a move towards single reading, but in the short term it would be used to enhance double reading. The second indicated that as double reading in their practice had a training function they were unsure as to the role of a prompting system in this respect.

6.6.4.2 Perceived effects on sensitivity and specificity

In the post-trial questionnaire subjects were asked to state whether they believed that their sensitivity and specificity with respect to ill-defined lesions, micro-calcifications and overall was better, worse or the same in the prompted conditions compared with the unprompted condition. Their responses with respect to sensitivity and specificity are shown in Figures 6.24 and 6.25 respectively.

Generally, subjects believed that their sensitivity for calcifications was improved and that their sensitivity for ill-defined lesions remained unchanged. Positive beliefs concerning the effectiveness of the micro-calcification detection algorithm may stem from actual cases where calcifications had been missed, but were subsequently brought to their attention by the prompting system. This result reflects the more favourable assessment by subjects of the micro-calcification algorithm compared with the ill-defined lesion algorithm given in response to other questions. Analysis of readers decisions showed that there was no signi-

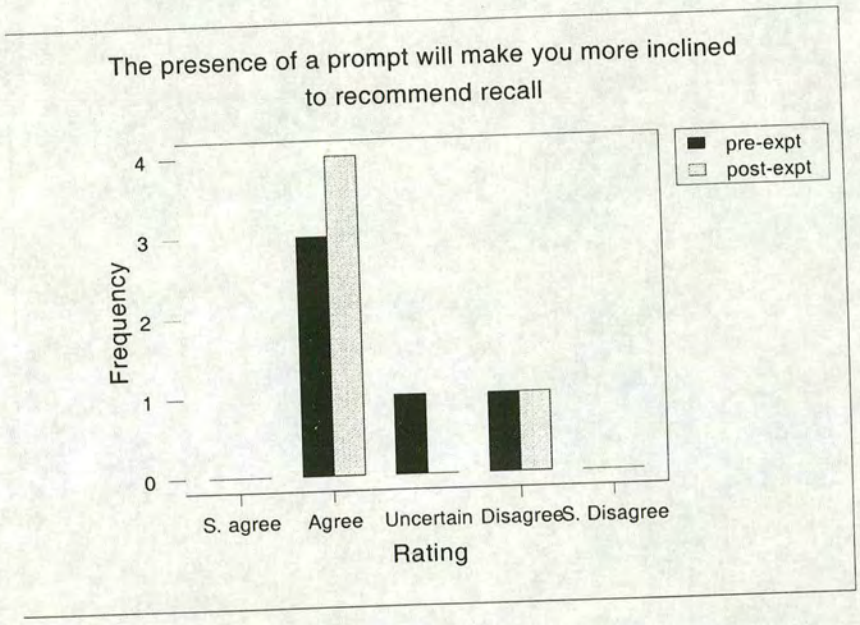


Figure 6.22: Agreement with the question: 'In cases where you are unsure, do you believe that the presence of a prompt will make you more inclined to recommend recall?' on a five point scale from 'Strongly agree' to 'Strongly disagree'.

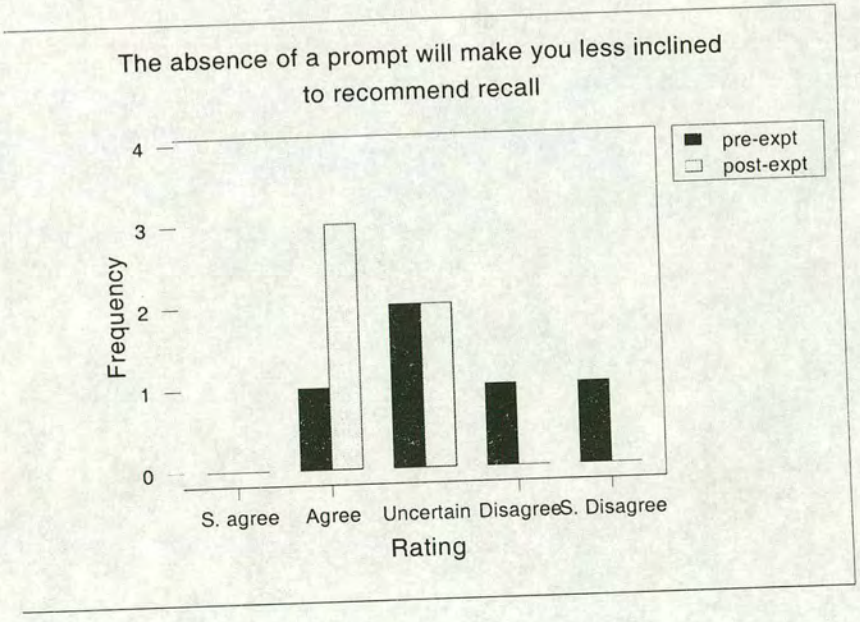


Figure 6.23: Agreement with the question: 'In cases where you are unsure, do you believe that the absence of a prompt will make you less likely to recommend recall?' on a five point scale from 'Strongly agree' to 'Strongly disagree'.

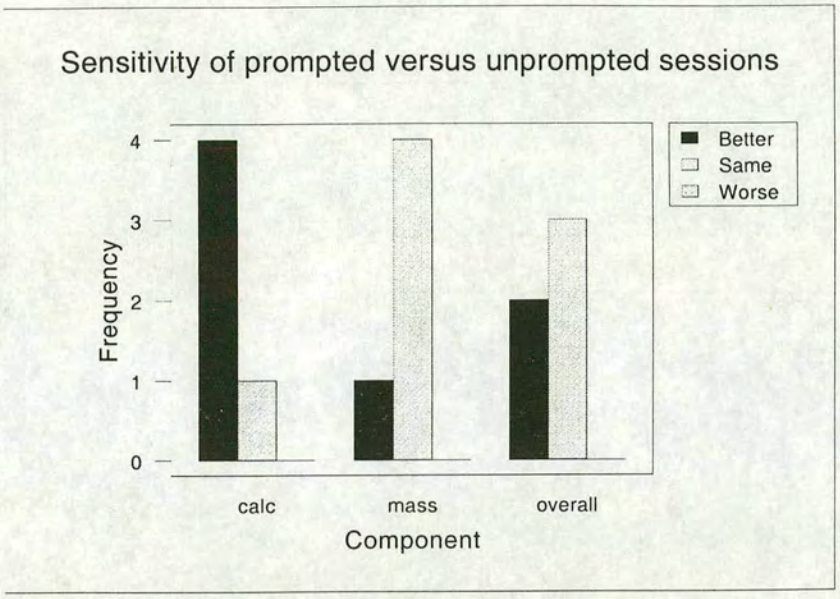


Figure 6.24: How subjects believed the system effected their sensitivity for targetted feature types.

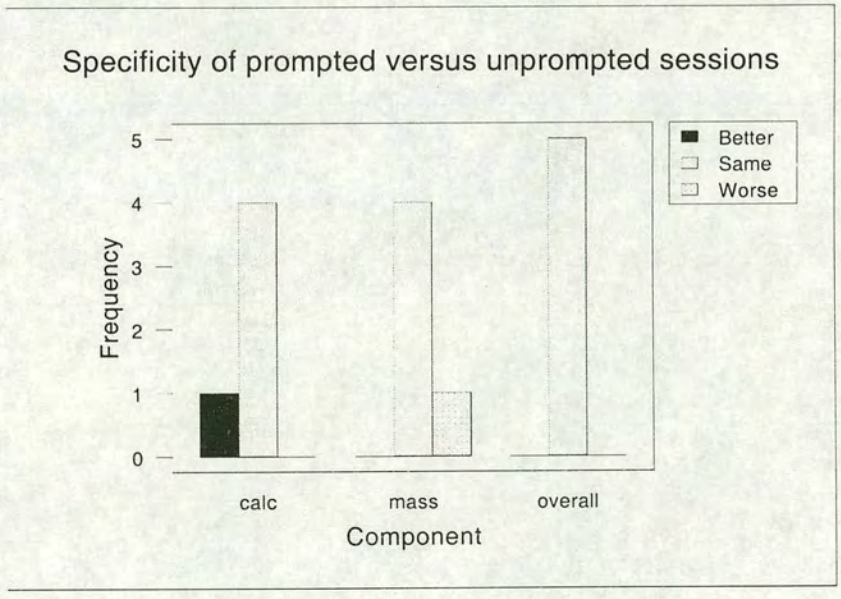


Figure 6.25: How subjects beleived the system effected their specificity for targetted feature types

ficant difference between subjects' sensitivity in the prompted compared with the unprompted condition. If fact, fewer cancers were recalled in the prompted condition [131].

Overall, subjects believe their specificity in prompted conditions is comparable with their specificity in unprompted conditions. The exception is subject E, who believed that her specificity had improved with respect to micro-calcifications, but had worsened with respect to ill-defined lesions. It is possible that subject E believes that the system is lowering her confidence threshold for ill-defined lesions (that the presence of a prompt will make her more likely to recommend recall) and raising her confidence threshold for micro-calcifications (the absence of a prompt making her less inclined to recommend recall).

The use of a prompting system is supposed to leave a reader's specificity unchanged, however, during the course of the trial a reduced recall rate was shown (though not significantly so) in the prompted condition [131]. This corresponded with a decrease in the number of cancers detected in the prompted condition (again, not significant), suggesting that use of the system might have increased subjects' confidence threshold for making positive decisions. This contrasts with subjects' assertions that their specificity was unaltered, but is in accord with responses to earlier questions about the perceived effects that the absence of a prompt may have on decision-making.

The perceived effect of prompts on subjects performance might be at odds with objective performance effects, although better quantitative data would be required for confirmation. If this is the case, it would be worth exploring where the discrepancy arises. The subjects in this experiment had been involved in the development of PROMAM and may be inclined to respond either positively, or according to the model of prompting as they understand it. Alternatively, there might be a bias arising from their interaction with the system, for example, additional detections may be more salient than occasions where their confidence has been altered.

6.6.4.3 System performance

In the post-trial questionnaire subjects were asked to estimate the sensitivity of the system's components, and of the system overall, and to rate their confidence in these estimates. The results of the sensitivity estimates and of the confidence ratings are shown in Figures 6.26 and 6.27 respectively. Averaged sensitivity estimates are shown in Table 6.20, along with the system developer's estimates for the sensitivity before the trial (subjects were told of these during training). The actual sensitivities achieved by the system on the set of cancers in the trial test set is shown in Table 6.21.

Table 6.20 shows that the mean of subjects' sensitivity estimates corresponds

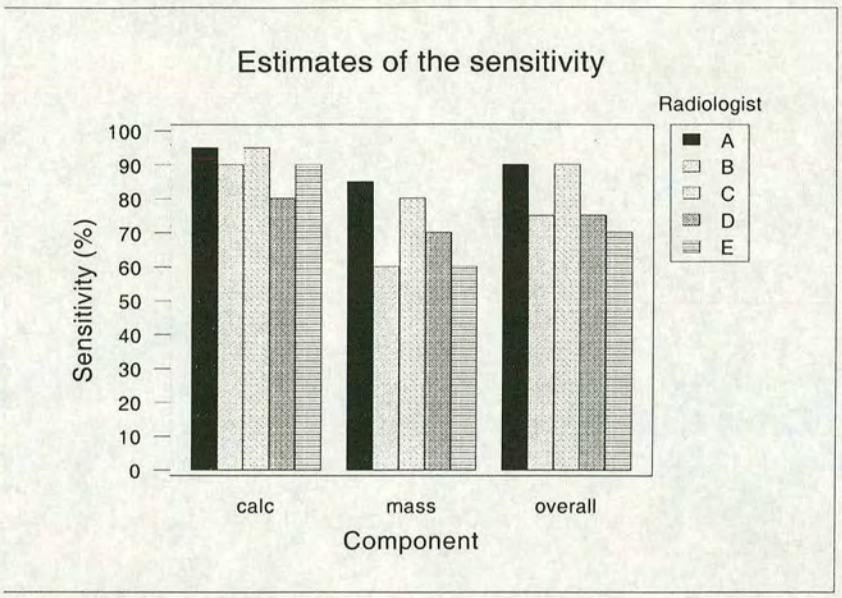


Figure 6.26: Shows subjects' post-trial estimates for the sensitivity of each algorithm.

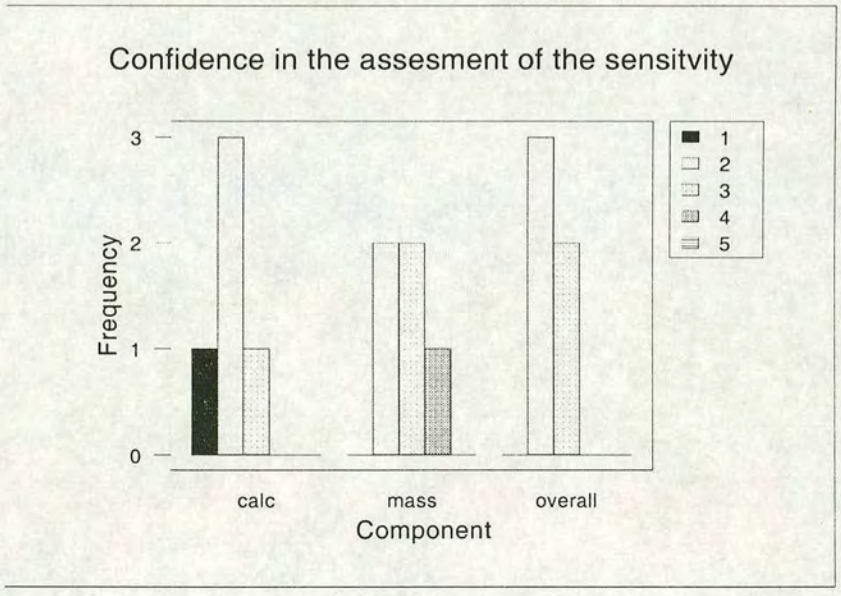


Figure 6.27: Subjects' confidence in their rating of the system's sensitivity on a five point scale from Most Confident (1) to Least Confident (5).

remarkably well with the actual sensitivity achieved by the system in practice (Table 6.21. Although the system designers' own estimates were made available to subjects before the trial, the ill-defined lesion algorithm failed to meet the ex-

Algorithm	Subjects' estimate of sensitivity			
	Mean	Median	St. Dev.	S.E. mean
Calcification	90.0	90.0	6.12	2.74
Ill-defined lesion	71.0	70.0	11.4	5.10
Overall	80.0	75.0	9.35	4.18

Table 6.20: Shows averages of subjects' sensitivity estimates

Sensitivity estimates given in training	Actual sensitivity
90%	93.8%[1]
81%	72.9%[1]
—	81.4%[2]

Table 6.21: Shows the system develops' estimates of the sensitivity of the system prior to the trial, and the actual sensitivity of the system for the cancers included in the trial. [1] These figures represent the percentage of correctly prompted cancers in either the ill-defined lesion only or micro-calcification only categories. It is not known which algorithm actually produced the correct prompt. [2] This figure indicates the percentage of correctly prompted cancers for all lesion types included in the experiment. Again it is not known which algorithm produced the correct prompt.

pected sensitivity of 81%, and subjects were able to down-grade their estimates accordingly. Furthermore, no details were given to subjects about the expected overall sensitivity of the system (that is, of the ill-defined lesion and microcalcification algorithms taken together), and yet they were able to give an accurate estimate.

Examination of Figure 6.27 indicates that subjects expressed greatest confidence in their estimate of sensitivity for the micro-calcification algorithm and least confidence in their estimates for the sensitivity of the ill-defined lesion algorithm. This pattern is reflected in the standard errors and standard deviations for these estimates.

6.7 Trial data

When reading the prompted conditions subjects were asked to indicate if a correct prompt was given for the significant feature in each case they recalled. Obviously, this information was not available for the 96 cases recalled by the unprompted reader alone, so a follow-up exercise was devised to determine which of the these recalls had actually been correctly prompted.

Prompt sheets for the unprompted single reader recalls were initially examined

by a member of the PROMAM team, and 43 cases that clearly had not been correctly prompted were eliminated. Eliminations included cases where there was no prompt on the breast for which the recall had been made, or where the prompt was quite obviously for a different feature or in a completely different region of the breast. The remaining 53 cases (that is, those where there was any ambiguity) were examined by a film reader to determine the veracity of the prompts.

Prompted reader recall	Unprompted reader recall	Correctly prompted?		Total
		Yes	No	
Yes	No	35	34	69
No	Yes	31	65	96
Yes	Yes	41	24	65

Table 6.22: Correctly prompted recalls made by the prompted and unprompted readers.

Table 6.22 shows the number of prompted single reader recalls, the number of unprompted single reader recalls and the number of cases recalled by both the prompted and unprompted reader. These are tabulated to show for each the number of recalled cases that were correctly prompted by the system.

Of the prompted single reader recalls, 50.7% were correctly prompted by the system, whereas only 32.3% of the unprompted single reader recalls were correctly prompted. A Chi-squared test indicates that this result would not be expected if exposure to the system and the proportion of correctly prompted recalls were independent (p=0.017). Thus there is a greater level of agreement between subjects and PROMAM when the subjects are exposed to prompting information — implying that the prompts have had an influence on subject's decision-making.

This influence could be due to the detection of a greater number of significant features that would have otherwise been overlooked, however, the result is also consistent with the evidence from the interview data that prompts are influencing classification decisions.

Of the cases recalled by both the prompted and unprompted reader, 63% were correctly prompted for. This greater consensus between subjects and prompting system where subjects also agree with each other is unsurprising, as the features precipitating recall are likely to be more obviously suspicious, and therefore more likely to be prompted.

6.8 Discussion

6.8.1 Models for effective prompting

The results of earlier trials suggest that film readers are able to use FP prompts as evidence to accurately assess the capabilities of a prompting system (Chapter 4). This led to the conclusion that FP prompts can actually be useful in that they provide an account of system behaviour over and above the typically sparse evidence available from TP prompts alone. Parallels were drawn with the practice evident in some screening clinics where film readers annotate benign but 'interesting' features to give a public account of their reasoning. From this it was inferred that prompts for 'candidate features' (features that a film reader might consider worth recalling or annotating) have a useful role in orienting users to the system's capabilities. According to this view, degrading effect of increasing FP rates (as reported by Hutt [70], for example) would be due to FP prompts that do not correspond to candidate features — i.e. those for obviously benign features, normal breast tissue and artifacts.

The evidence presented in this chapter provides additional support for the hypotheses that film readers use FP prompts to provide an account of system capabilities. Subjects were again shown to be able to make an accurate assessment of the sensitivity of each algorithm (Section 6.6.4.3), they also expressed concerns when the system omitted to prompt for significant features — even if they did not believe that such features were a particular cause for concern (Section 6.5.3.3). In addition, subjects were critical of FP prompts for obviously benign features or normal structures, and particularly of FP prompts which they could not account for (Section 6.5.2.2).

Although this model of how film readers use prompting information to inform their view of the system appears to be vindicated, it remains only a partial model of how a prompting system and film reader interact. This is because it says little about the way prompts influence a reader's decisions beyond how readers might establish a view of the credibility of the system.

It was assumed that at some level of system performance film readers would be inclined to examine each prompted region carefully and therefore benefit by detecting additional cancers that they might otherwise have overlooked. If a system performed less well, film readers would be inclined to take less notice of the prompts and therefore be more likely to overlook prompted abnormalities. There is some evidence for this effect — Section 6.5.2.1 describes how subjects make an a priori assessment of the significance of prompts based their on shape,

frequency and location. However, there is also an accumulation of evidence to suggest that prompts also influence film readers when they are deciding how to classify features, thus the above view that prompting merely creates the need for additional classification decisions to be made is overly simplistic.

It is unclear whether every subject participating in the trial was influenced by prompting information when making classification decisions. Only subjects B and E stated explicitly (in the free form response questions and during interviews) that they were using prompts to assist with classification. However, other subjects hinted that this might be the case in answer to specific questions in the post-session questionnaires (ie when asked if the presence/absence of a prompt would make them more likely/less likely to recommend a recall). Three subjects agreed with the statement 'The absence of a prompt makes you less likely to recommend recall', and two indicated that they were unsure. Therefore it is possible that the presence or absence of a prompt has a subtle effect on a film reader's confidence threshold when making a classification decision, and that film readers are not necessarily always aware of this influence. If film readers are being influenced involuntarily this would make the task of modifying their response more difficult — as demonstrated by this study, simply telling a film reader that they should not use prompts to assist with classification during training is not in itself sufficient.

The prevailing view is that systems that assist with detection are designed to address a different problem than those that assist with classification decisions. The goal of the former is to improve a detection performance while having minimal impact on specificity, and of the latter to decrease recall rates without having undue effect on sensitivity. Subjects participating in this trial, and those completing the 'clinic questionnaire' expressed the opinion that effective detection aids would be of greater value than effective classification aids (Section 6.6.1.2). This is not surprising since a higher priority was also given to improving detection performance over reducing recall rates (Section 6.6.3). However, this view is perhaps oriented to the goals of the screening programme and readers' professional concerns, rather than to the highly contingent demands made by the interpretation of individual mammograms [3]. The evidence from this trial suggests that it is difficult to draw a clear distinction between detection and classification aids — that when film readers are faced with a difficult classification decision, they can be influenced by, or they will appeal to, whatever evidence is available.

Some of the specific problems that might arise from using a detection aid to assist with classification decisions have already been discussed (Section 6.5.1.3), and

[3] See Chapter 5

it has been suggested that this mode of usage should be discouraged. However, if film readers persist in using a prompting system in this way, then it becomes worthwhile examining how minimum performance requirements might be affected. Interestingly, this model of system usage may go some way towards explaining the strict performance requirements reported by Hutt.

Hutt used ROC methodology to determine a maximum FP rate that could still support performance gains for film readers using a prompting system in an trial setting. What a ROC curve shows is how well an observer can discriminate between benign and malignant features over a range of confidence thresholds. However, there can be different reasons for improvement in a film reader's ability as measured by a ROC curve.

If a prompting system assists a film reader to detect features that they would have otherwise have overlooked, then consequentially more features will also have been classified correctly — giving an improved ROC curve [4]. It is also the case that if a system enables better discrimination between benign and malignant features that film readers have themselves detected, this will also result in an improved ROC curve.

If we imagine that the presence of a prompt increases the chances of a feature being recalled, and that the absence of a prompt reduces the chance of a feature being recalled, then at some operating point (TP and FP rate combination) sufficient information would be supplied by the system to assist film readers to make a better discrimination between the set of benign and malignant features that they themselves have detected. In this scenario any additional cancers 'detected' would be by virtue of correcting errors of classification.

Indeed, if the gains in performance reported by Hutt were due to improved discrimination rather than improved detection, then it is not surprising that the FP prompt rate demanded is particularly strict. Furthermore, the prevalence of detection errors may be reduced in small-scale ROC experiments if subjects respond to the experimental setting by being especially vigilant. If it is possible to reduce the reliance on the system for classification, then performance gains might still accrue due to correcting errors of detection in actual clinical practice. Some practical suggestions for reducing dependance on prompts for classification decisions are given below:

1. Make classification information available, even though the system's primary goal is to assist with detection. This might be done by exposing the clas-

[4]This assume that the majority of additional suspicious features discovered and recalled actually do turn out to be malignant.

sification mechanisms used in detection algorithms (for example, in the ill-defined lesion algorithm), or by combining the function of separately developed classification and detection aids.

2. Change reading practice so that decision to recall made before examining the prompts will automatically stand.

   This should effectively prevent the absence of a prompt from influencing a film reader's recall decision, thus mitigating the worst effects of using a detection aid to assist with classification (i.e., unwanted FN decisions.)

3. Systematic training to ensure that film readers develop the best strategy for interpreting the prompts.

   Since film readers may be involuntary users of prompting information for classification decisions, a systematic approach to training may be required. This could involve evaluated reading sessions so readers can be assisted in recognising the particular circumstances where influence is likely [5].

4. Depend upon readers adjusting to the best approach to using prompts over time and with feedback (from, for example, assessment clinics.) Readers might adapt their behaviour over time to maximise the performance benefits available from prompting by critically examining the outcome of their decisions.

6.8.2 Training

There was no formal evaluation of the training package, but there is evidence to suggest that subjects were able to use the training material to explain some of the prompts. Indeed one of the main goals of training should be to give film readers an account of how system behaviour relates to algorithm function — particularly for behaviour that appears counterintuitive to film readers. Subjects did discover prompts that they could not explain using the account given during training. It is inevitable that film readers get to see more cases than system developers, and therefore will become the principal experts concerning what the prompting system actually does. However, they will lack the technical expertise to relate this in a coherent fashion to algorithm function. Thus there have to be opportunities for

[5] This proposal, and the preceding one, mirror the strategies employed by readers (reported in Chapter 3) to manage bias in routine practice. These include the application of 'mental discipline' when a potential for bias is recognised, and the organisation of evidence to reduce the potential for bias

continuing interplay between users and developers of the system so that updated the account of behaviour might be provided.

Training should also encourage film readers to find the right balance between making an a priori assessment of the significance of prompts, and carefully examining each prompted region. Although the former may play a role in enabling readers to tolerate a higher FP rate, it may lead to degradation in the usefulness of the system as a detection aid.

One subject doubted her own opinion when faced with a mass prompt on dense breast tissue, and believed that because the prompt could not be explained in terms of the examples of FP prompts given during training, that perhaps the system had detected something that she could not herself see. This indicates that the scope of the system relative to film readers' own abilities should be made clear — specifically that the system does not have access to privileged information and is unlikely to perform better than a film reader in this circumstance.

The two subjects who reported that they were influenced by prompts when making classification decisions both expressed a concerned that this might be dangerous, particularly when the absence of a prompt gave reassurance for a decision not to recall. It is clear that they were unable to resolve these concerns given the information available to them. Training should also address this issue by providing the best evidence available in order to guide this type of decision. That is, an attempt should be made to minimise the circumstances in which a film reader might be uncertain about the best way of interpreting a prompt.

While it is believed that training can assist film readers develop strategies for coping better with weaknesses in prompting system performance, training should not be regarded as a panacea for solving all usability problems that arise from using a flawed system. The aim of training should be to minimise inappropriate use of a system, rather than to compensate for system weaknesses.

6.8.3 Desirable properties of a prompting system

This study has elucidated some of the strategies used by subjects to cope with the FP burden of PROMAM, enabling more specific conclusions to be drawn about the types of system behaviours that are desirable for effective prompting. A summary is given below:

Predictability If a system's behaviour is consistent, then this enables film readers to be able to predict which features will be prompted, thus reducing the amount of time spent re-examining the image. Predictability can be maximised if a system largely prompts for the types of features that film readers

consider significant, and if film readers are given a sufficient information so that they can account for prompts that are artifacts of algorithm function.

Robustness This is an adjunct to predictability. If a system is sensitive to small variations in the presentation of features that are considered by readers to belong to the same category, then this may make the system appear less consistent.

Few prompt types with regular characteristics Recurring FP types that have regular characteristics (for example, shape, location and frequency) should be avoided. Film readers will make an a priori assessment of significance based on regular characteristics and might not then examine the image thoroughly. This issue might also be addressed by a change in prompting strategy. If less information is given, perhaps by giving prompts a uniform shape, this may eliminate the cues that enable prediction.

Few multiple prompts In addition to the overall FP burden, subjects have expressed a concern about individual cases that have large numbers of prompts. The micro-calcification algorithm is susceptible to producing multiple prompts in particular circumstances, and it has been suggested that prompts might be combined to reduce the burden of examining a large number of them individually. One subject has suggested this approach for occasions where what has been considered by them to be a single lesion has been fragmented into a number of smaller prompted areas. The caveat to this approach is that it is an attempt at interpretation, and as such has particular dangers. For example, a genuine lesion might be included in a grouping of nearby vascular calcification prompts.

Chapter 7

Summary

This thesis shows how qualitative methods can be applied to the evaluation of decision-support systems, not only by detailing the context in which the system is to be deployed, but also by examining closely the interaction between system and user. In particular, it was possible to examine aspects of readers' use of a prompting system in ways that would have been difficult to achieve using quantitative methods alone. Using this approach it was shown how the traditional distinction between detection and classification aids in mammography is one that readers have difficulty sustaining in practice.

7.1 Evaluation of prompting systems

In Chapter 1 it was argued that a dependence on quantitative methods for evaluating prompting engenders a technical bias, making it difficult to frame the problems of prompting in terms other than system performance. Quantitative studies of prompting have been pessimistic about the level of performance required to raise reader performance, with one study suggesting that prompting systems might have to perform as well, or better, than experienced human observers before they have an appreciable effect. However, quantitative methods leave film readers' interpretation and use of prompts largely unexamined, and are thus unable to confirm whether prompts have actually been used appropriately or efficiently.

Recently there has been a growing interest in the application of qualitative methods to the evaluation of decision support systems in medicine, motivated by a recognition that many of the difficulties encountered in the deployment of such systems are social, rather than technical, in nature. This thesis shows how qualitative methods were successfully applied to the evaluation of PROMAM, a detection aid with a complex behaviour. Initially, ethnographic techniques were

used to examine conventional reading practice. In subsequent prompting experiments, interview, questionnaire and observation were the principle means of data collection, although quantitative measures were also used. This mixed approach enabled a shift of emphasis from technical to human factors issues. Rather than framing the problem of prompting as one of delivering sufficient system performance, it was possible to explore the difficulties readers faced when deciding what individual prompts mean and how they should be acted on. It was also possible to examine how readers learned about the PROMAM from its behaviour. Rather than being unchanging processors of images and prompts, readers were shown to adapt to PROMAM's weaknesses and to base their decision-making around an evolving understanding of the PROMAM's capabilities.

7.2 Reader error and prompting aids

The aim of prompting is to draw attention to what might be significant features within a mammogram and, by doing so, prevent them from being overlooked. In this way, prompting is supposed to support the detection of lesions, rather than their classification. It remains the responsibility of the film reader to decide whether a prompted feature merits additional investigation.

In the literature, a distinction is drawn between search, detection and classification errors, corresponding respectively to occasions where an abnormality does not enter the reader's useful field of view, is overlooked, or is mistakenly thought to be benign. The development of decision-aids for mammography mirrors this distinction, with detection aids and classification aids developed and implemented as separate systems. However, the presumption that detection can be supported independently of classification depends on these functions being distinct for a human observer. If this were not so, then use of a detection aid could bias a reader's judgement concerning significance, conceivably leading to a degradation of reader performance.

Two empirical methods have been proposed for distinguishing between detection and classification errors in human observers, one depending on gaze duration, and the other on whether an observer reports the presence of cancers that he or she does not recall. In Chapter 2 it was argued that, for some classes of observer error, the above methods disagree as to whether a detection or classification error had occured. There may therefore be some doubt about the processes of detection and classification are always distinct for human observers, and consequently, whether detection can easily be supported in an independent fashion. One mo-

tivation for examining how readers use PROMAM was to confirm that it was being used as anticipated, and as the rationale for prompting implies it should. One outcome of pursuing this question has been a contribution to a conceptual understanding of prompting, revealing readers interpretation and use of prompts to be more complex and sophisticated and than had previously been reported.

7.3 Reading practices in breast screening

The literature pertaining to radiological expertise suggested that the interpretation of medical images can be influenced by different types of prior information, including, knowledge of disease prevalence, clinical history and the likely locations of abnormalities. Readers are thus primed for the types of finding they are likely to make by the setting of the task and the circumstances of each individual case. The literature reviewed in Chapter 2 shows that the pattern of eye-movements employed by film readers reflects a complex series of responses to each image's unique visual terrain and to the unfolding goals of the inspection. The overall picture of expert visual search is that of an adaptive process — attention is metered according to the demands made by individual regions of an image. This psychological description of visual expertise concerns how individual film readers respond to images. A central claim of this work is that a more complete account should include how individuals' cognitive abilities are brought to bear in the workplace.

To examine the social dimensions of screening expertise, an ethnographic investigation into reading practices at six UK breast screening centres was undertaken. The investigation revealed that skilled image interpretation is not as straightforward as simply looking at an image and reporting what the reader's cognitive mechanisms uncover. Findings may be qualified in light of evidence gleaned from other sources, such as previous films, HRT status, and perceived flaws attributed to the process by which a mammogram is produced. Interpretation may be more or less difficult depending on the image content (for example, dense as opposed to lucent breasts) and on the 'depth' of evidence to hand (for example, whether previous films are available).

The psychological account of reading is not contradicted by the work practice study, rather, it is supported and contextualised. Readers were shown to have an insight into their cognitive abilities and consequently demonstrated a reflexive approach to reading. Perceptions about deficiencies in skill informed readers deployment of attentional resources and motivated informal monitoring of performance through collaborative practices. Despite appearing to be a solitary activity,

reading was actually found to have a significant collaborative dimension, often mediated through the practice of double reading. The notional goal of double reading is to improve readers' detection performance, however, its structure was also observed (with varying degrees of formality) to support training, to provide a standard against which performance can be judged and in conjunction with the practice of annotation, to provide a method of maintaining accountability within a community of practice. Again, a reflexive approach was in evidence. Often there was a tension between the benefits derived from informal use of particular sorts of information (for example, examining the first reader's decision as a check on one's own performance) and the concern that such usage might bias decision-making. Where such tensions exist, readers often managed the information potentially available to them so that perceived biases might be reduced. Psychological accounts of reading show that decision-making is highly sensitive to the circumstances in which it is undertaken. Readers demonstrated an awareness of this sensitivity by purposefully manipulating the setting in which films are read.

There are implications here for the deployment of prompting systems, especially where this would involve replacing double reading with a prompted single reader. The opportunities for collaboration that double reading enables would be undermined. It remains to be seen whether prompting can also substitute for the informal and nuanced exchanges evident in many double reading partnerships.

If a move from double reading to computer assisted single reading is contemplated then it would be advisable to closely monitor the effects on performance. In a trial situation, it would be prudent to move to blinded double reading prior to introducing prompting, to allow readers to 'acclimatise' and so that the effects of withdrawing double reading can be more readily ascertained.

7.4 Subjective responses to PROMAM

In the first of three reported investigations into the use of the PROMAM system, the subjective responses of four subjects to three system configurations were sought.

One aim of this exercise was to discover a FP rate that might be acceptable in clinical practice. Because representative test sets were used it was assumed that subjects would not be able to make reliable judgements about the system's sensitivity (laboratory benchmarking of system performance requires a large sample of pathology proven malignancies to be reliable). Since the test was primarily

aimed at assessing the acceptance of FP rates, subjects were only informed that the sensitivity in each condition would either be high, medium or low. However, subjects appeared able to judge the system's sensitivity accurately, and indeed, were more critical of its sensitivity than its specificity. It is likely that subjects were comparing their own decision with that of the system for features which were either recalled or considered for recall. Subjects tended to rate conditions more highly where the system's decisions had agreed with their own.

An analogy can be drawn here with the practice of annotation evident in clinical practice. By annotating benign but (implicitly) noteworthy features within a mammogram, readers' decision-making is made visible for a larger set of features than just those that result in a recall for assessment. As well as serving to demonstrate vigilance, the practice of annotation may help to establish and reinforce community norms for decision-making in a task where there is often ambiguity. Although prompts are provided primarily as an imperative (to examine a region of the image), they may also incidently betray aspects of the system's operation. This opens the possibility that certain types of FP prompt might be acceptable and useful to readers because they are informative of system behaviour.

Despite appearing able to accurately gauge the system's performance, readers sometimes misinterpreted the scope of the system's abilities, for example, after using the system some believed it capable, and indeed competent, at detecting asymmetry. At other times readers were confused by the prompts, describing the system's responses as 'inconsistent'. This suggests that exposure to the system is not in itself sufficient for radiologists to develop an accurate model of system behaviour.

Finally, subjects demonstrated a subjective tolerance for FP rates greater than had previously been suggested as necessary to produce an improvement in performance. Indeed, the observation data show that some minimum prompt rate is needed to maintain subject's engagement with the system. Of course, positive subjective assessment may not necessarily coincide with objective performance effects, but it is possible that previous work underestimates the FP prompt upper limit.

In order to examine more closely how subjects were able to infer (sometimes erroneously) details of PROMAM's scope and function from its behaviour, and their apparent tolerance to FP prompts, a followup exercise was devised. Subjects were asked to comment on their reasoning for a large number of prompted cases in a 'think aloud' protocol (Chapter 5). Again, subjects mistook the operational scope of the system and found apparent inconsistencies in the system's behaviour.

These phenomena appear to arise out of the strategies employed by subjects to make sense of the prompts.

7.5 Making sense of prompts

Reading is a constrained activity. One example of this is the physical arrangement of mammograms on a viewer. Different arrangements can serve to accentuate particular relationships between mammograms (for example, temporal or geometric), but a film viewer can only support one such arrangement at a given time. Another concerns the attentional resources available — neither readers' time nor stamina are unlimited. One consequence is that a detailed examination of each region of the mammogram is infeasible. Indeed, studies of eye movements rarely reveal any systematic pattern, and show that even for expert observers, not all of the image is foveated. This approach to examining images implies that readers are accountable for the content of mammograms in particular ways. For example, all of the features within an image are not equally accountable. Features that are common and conspicuously benign may merit only cursory examination, whereas features that have suspicious visual properties warrant closer scrutiny. Data from the work practice study suggests that the degree to which a feature is accountable depends not only on the suspiciousness of its appearance, but also to the context of its presentation. For example, subtly suspicious lesions occuring in the danger areas, or in dense breasts, or where they indicate change, or when co-present with another lesion, may be more accountable than a similar lesion appearing in other circumstances.

This helps explain why certain FP prompts are deemed acceptable — because they attend to what in the image the reader is accountable for. The think aloud protocol revealed that sometimes prompts for patently benign features are judged acceptable because there was some potential contextual significance (e.g. prompting for a screen artefact in a 'danger area'). In contrast, prompts for features within the image that do not demand an account are often deemed unacceptable, even disturbing. One explanation is that readers are disorientated if forced to examine closely regions that might otherwise be cursorily dismissed.

In the think aloud protocol, one feature of subjects' interpretation of prompts was an assumption that the system had responded in a purposeful or meaningful way with respect to the contents of each image. Subjects often falsely analogised the system's reasoning with that of a human film reader. They were inclined to believe the system had a greater operational scope than is in fact the case, for

example, believing it capable of detecting asymmetry. Because subjects did not possess even a cursory account of the system's operation (for example, the clustering rule employed by the microcalcification detection algorithm), sometimes its behaviour appeared to be inconsistent (for example, 'selective' prompting of vascular calcification).

The evident discomfort shown when subjects were unable to account for prompts may have arisen from the presumed possibility that the system had detected something of significance that they themselves could not see. Knowing that some cancers can be difficult for human observers to perceive, and not being sure of the operational scope of the system, creates a dilemma — is the prompt a merely an unaccountable FP, or a very 'astute' TP? If readers' default assumption is that prompts are purposeful, then this dilemma becomes even more acute. This suggests that not only are readers accountable for what they notice in a mammogram, but also the system is accountable for what it detects, and readers are accountable for their interpretation of prompts.

In summary, findings from the think aloud protocol suggest that, in practice, prompting information is not merely used to ensure that detected features are not overlooked. It is also used to provide reassurance that the mammogram has been examined thoroughly, as contributory evidence for readers' own interpretations, to resolve ambiguity and to address situational difficulties.

7.6 Usage in simulated practice

The think aloud protocol reproduced some of the findings made by the work-practice investigation and the subjective responses to prompting experiment, for example, subjects' belief that the system is capable of detecting asymmetry. However, other findings of the 'think aloud' protocol had little corroboration in observations of either naturalistic system use or of conventional film reading. Examples of these included the suggestions that use of PROMAM might influence readers' classification decisions, and that readers develop strategies to cope with the system's shortcomings by, for example, learning to predict what is prompted by the appearance and location of the prompts. Pre-clinical trials of PROMAM provided an opportunity to see if these phenomena are evident in readers' use of the system under more realistic conditions. A mix of qualitative and quantitative methods of data collection were again used, including observation, interviews, use of questionnaires, and analysis of subjects' recall decisions.

Prior to the trial a training package was delivered to each of the subjects. The

content of the package was informed by the findings of previous investigations, and consisted of a functional description of each algorithm along with selected examples of prompts. The aim was to illustrate the limits of the system's operational scope and its interpretive ability. Some examples were specifically chosen to demonstrate aspects of the system's operation which readers had previously found perplexing. Overall, it was anticipated that the training package would equip readers with a better means of accounting for system responses. Guidance was also given to subjects concerning use of the system, particular with respect to PROMAM's role as a detection aid.

The results confirmed what was hinted at previously. Subjects stated that they often judged the significance of a prompt based upon its size, shape and location, rather than on the content of the image that was prompted for. Subjects also reported an ability to sometimes predict what in the mammogram the system will prompt. This had the perceived benefit of reducing the overhead of re-examining the mammogram when prompts have been successfully anticipated. In the 'think aloud' protocol, it was suggested that readers use PROMAM to address the contingent difficulties associated with individual cases. In the pre-clinical trial, the interview and questionnaire data suggested that subjects sometimes use PROMAM to inform their classification decisions. A statistical comparison of the patterns of recalls between prompted and unprompted conditions added weight to this conclusion.

The quantitative analysis demonstrates a biasing effect, but does not show whether readers' overall performance is adversely affected, or whether the effect would be sustained over longer periods of routine use. These are matters for further investigation.

7.7 Qualitative versus quantitative approaches

Qualitative methods should not be seen as a substitute for quantitative modes of evaluation. Benchmarking performance is an important activity, but provides only a single view of the potential effectiveness of a decision-aid. Qualitative methods can be applied both to understanding conventional working practices and to exploring in detail users interaction with decision-aids. The former addresses the requirements for successful deployment of a technology into the work place and the latter provides the conceptual basis for system design and the interpretation of benchmarks. Quantitative investigations might reveal that a system is flawed, but it is through the application of qualitative methods that flaws are most easily

identified and remedies suggested.

To illustrate, consider the statistical analysis presented in Chapter 6 suggesting that subjects' decision-making may be biased by exposure to PROMAM. It is the qualitative data that provide a rich conceptual framework for explaining this bias. Both the work practice investigation and the think aloud protocol suggest that individual mammograms can present differing challenges to readers. In the think aloud protocol it was shown that subjects looked to the prompts to address the problem at hand, rather than the generic problem of accidentally overlooking a cancer. The qualitative data would further suggest that an absence of a prompt would make a reader less inclined to recall a patient for assessment and that the biasing effect is likely to occur where there is uncertainty or ambiguity, rather than uniformly over each case. Finally, studies show that readers' handling of conventional forms of evidence attends to the possibility of bias. By learning from these practices it is possible to suggest how delivery of prompts might be ordered to reduce their biasing effects.

In the above example, findings from the qualitative investigations prompted the attempt to quantify how exposure to the system had influenced decision-making. There are several points during this thesis where it would be appropriate to use quantitative methods to confirm the findings of qualitative investigations. These are discussed further below. In contrast, the qualitative investigations reported here were motivated by the results of previous quantitative studies into prompting and reader expertise.

7.8 Conclusions

Qualitative methods were used to investigate both reading practice and readers' use of PROMAM. The former show how, in practice, radiological expertise is at once constrained by the physical limitations of screening artifacts, supported by informal collaborative practices and conditioned by the setting of the task. This understanding of reading practice provided a conceptual basis for interpreting the findings of subsequent prompting investigations. It suggested that readers routinely encounter many sorts of interpretative problems when reading, whereas detection aids are only designed to address one specific problem. Thus there may often be a gap between a system's capabilities and a reader's immediate difficulty, and a temptation to use the prompts to address the problem at hand.

The use of qualitative methods to investigate prompting made it possible to conceive of prompting system design and implementation in terms other than

setting goals for system performance. Readers were found not to be mere passive receivers of prompting information — they do not respond mechanically to each and every prompt in an identical fashion. Instead, readers use prompts to learn about the system's capabilities. They then use this understanding to subsequently modify the way they interpret and act upon the prompts. However, prompts by themselves often do not indicate the most appropriate interpretation or action — additional accounts of the system's role, scope and operation are also required. The nature of these accounts, and the most appropriate mode of delivery are important issues to address as prompting technologies mature and the prospect of routine use (for example, in clinical trials) becomes likely.

7.9 Further work

The studies reported in this thesis suggest several further avenues of investigation.

7.9.1 Possible quantitative investigations

The literature review in Chapter 2 raised the possibility that Gale's and Kundel's methods for determining whether an error of detection or interpretation has occured might give contradictory results under certain circumstances. It should be possible to apply these methods simultaneously and examine the data for contradiction. A positive result would confirm that the processes of detection and classification are not always distinct, casting doubt on the ability of decision-aids to support detection independently.

The work presented in this thesis is concerned primarily with the application of qualitative methods in pursuit of a conceptual understanding of prompting. Several of the findings merit further quantitative investigation to ascertain the extent and importance of their effect.

In Chapter 4 it was suggested that there may be a qualitative difference between different types of FP prompt in their effect on observer performance. It would be useful to confirm this by quantitative investigation. An experiment could be performed to determine if systems with different types of FP prompts have a differential effect on performance. For example, one could compare the effect on reader performance of system FPs, randomly chosen FPs and film reader chosen FPs.

Subjects' accounts of managing FP burdens by either prediction or prior judgements of significance could be verified quantitatively. For example, an experiment could be devised to test how effective readers actually are at predicting prompts,

and to quantify the befits gained from employing this strategy. Accuracy at prediction may actually be a useful metric for determining readers' understanding of the operational scope and functioning of the system. Eye movement studies could be used to examine the extent to which the character of a prompt determines the level of scrutiny given to prompted region of the image. This would be informative of the trade-off between the chance of overlooking a prompted cancer, and the benefits of improved efficiency. If warranted, such an approach could also test strategies for encouraging readers to examine each prompt with equal care, for example, by giving prompts a uniform size and shape to make judgements about what they are for more difficult.

A training package was delivered to subjects prior to their involvement in pre-clinical trials of PROMAM. Although informed by earlier work, the training material has not been formally evaluated. This could be done in several ways. Firstly, readers could be tested on their understanding of the packages contents to ensure that delivery is effective. Secondly, trained readers' judgements concerning a test set of prompted cases could be examined. Finally, the performance effects of training could be measured. Of course, subjective metrics should also be used, such as readers' opinion concerning the clarity of prompts and confidence in their interpretation.

7.9.2 Accounting for system behaviour

Providing an account of system behaviour through training implies that readers have to interpret prompts on the basis of a generic description of the system's operation and a few specific examples. In routine use readers will inevitably come across prompts that are difficult to account for from training alone. One possible solution is to investigate ways of documenting the production of each individual prompt. This could be done by revealing some of the 'hidden' processing steps involved in a prompts production. In the case of the microcalcification algorithm, all particles of calcification identified in image could be shown, and for the ill-defined lesion algorithm, the regions identified as 'candidate' lesions. Much work would be required to discover what information of this sort readers can effectively made use of and how it can be unobtrusively delivered.

The main advantage of a hard copy prompting interface is that prompt sheets can be easily and economically introduced into existing reading procedures, however, in the future it is likely that digital capture of X-Rays will replace film based systems [137]. This would enable mammogram production and display to be more closely integrated with decision-aids such as prompting systems. The use of a soft

copy display opens up opportunities for improving PROMAM's accountability by overcoming the physical limitations of a hard copy prompt sheet. It would be possible, for example, to interrogate the system for accounts of particular prompts. This has the dual advantages of avoiding the presentation accountability information when it is not perceived to be necessary, and of providing a log of requests detailing readers' perceived difficulties using the system.

A further possibility is that instead of developing monolithic tools (for example, detection and classification aids), a fine-grained approach might be more effective. Tools could be developed to address specific problems that readers encounter in routine practice, for example, 'undressing' lesions, assessing coverage between screening rounds, assisting with the interpretation of dense breasts, etc. Rather than having a blanket application, such tools could be brought to bear interactively as the particular need arises.

7.9.3 Full scale clinical trials

The next logical step in the evaluation of the PROMAM system is full scale multi-centre trials designed to test its clinical effectiveness. Clinical trials would also provide a further opportunity for continued qualitative investigations. Of particular interest is readers use of prompts in the clinical setting where decisions are consequential, what readers learn using the system for protracted periods and how use of the system fits in with and effects clinical practice.

Bibliography

[1] E D C Anderson, B B Muir, and A E Kirkpatric. The efficacy of double reading mammograms in breast screening. *Clinical Radiology*, 49(248-251), 1994.

[2] R J Anderson. Representations and requirements: The value of ethnography in system design. *Human-Computer Interaction*, 7:151–182, 1994.

[3] I Anttinen, M Pamilo, M Soiva, and M Roiha. Double rading of mammography screening films — one radiologist or two? *Clinical Radiology*, 48:414–421, 1993.

[4] S Astley, R Zwiggelaar, C Wolstenholme, K Davies, T Parr, and C Taylor. Prompting in mammography: How good must prompt generators be? In N Karssemeijer, M Thijssen, J Hendricks, and L van Erning, editors, *The Forth International Workshop on Digital Mammography*, pages 347–354, Nijmegen, June 1998. Kluwer Academic Publishers.

[5] C J Babcock, G R Norman, and C L Coblentz. Effect of clinical history on the interpretation of chest radiographs in childhood bronchiolitis. *Investigative Radiology5*, 28(3):213–217, 1993.

[6] J A Bargh. Attention and automaticity in the processing of self-relevant information. *Journal of Personality and Social Psychology*, 43(3):425–436, 1982.

[7] J Beck. Similarity grouping and peripheral discriminablity under uncertainty. *American Journal of Psychology*, 85(1):1–19, 1972.

[8] K S Berbaum and E A Franken Jr. Letter. *Investigative Radiology*, 27(7):573, 1992.

[9] K S Berbaum, E A Franken Jr, K L Anderson, D D Dorfman, W E Erkonen, G P Farrar, J J Geraghty, T J Gleason, M E MacNaughton, M E Philips,

D L Renfrew, C W Walker, C G Whitten, and D C Young. The influence of clinical history on visual search with single and multiple abnormalities. *Investigative Radiology*, 28(3):191–201, 1993.

[10] K S Berbaum, E A Franken Jr, D D Dorfman, E M Miller, R T Caldwell, D M Kuehn, and M L Berbaum. Role of faulty visual search in the satisfaction of search effect in chest radiography. *Academic Radiology*, 5(9-19), 1998.

[11] K S Berbaum, E A Franken Jr, D D Dorfman, S A Rooholamini, C E Coffman, S H Cornell, A H Cragg, J R Galvin, H Honda, S C S Kao, D A Kimball, T J Ryals, W J Sickels, and T P Smith. Time course of satisfaction of search. *Investigative Radiology*, 26(7):640–648, 1991.

[12] K S Berbaum, E A Franken Jr, D D Dorfman, S A Rooholamini, M H Kathol, T J Barloon, F M Behlke, Y Sato, C H Lu, G Y El-Kohoury, F W Flickinger, and W J Montgomery. Satisfaction of search in diagnostic radiology. *Investigative Radiology*, 25(2):133–140, 1990.

[13] M Berg. The construction of medical disposals. Medical sociology and medical problem solving in clinical practice. *Sociology of Health and Illness*, 14(2):151–180, 1992.

[14] D C Berry and A E Hart. Evaluating expert systems. *Expert Systems*, 7(4):199–208, 1990.

[15] R E Bird, T W Wallace, and B C Yankaskas. Analysis of cancers missed at mammography. *Radiology*, 184:613–617, 1992.

[16] M S Blois. Clinical judgement and computers. *The New England Journal of Medicine*, 303:192–197, 1980.

[17] C Bosk and J Frader. The impact of place of decision making on medical decisions. *Proceedings of the 5th Annual Symposium on Computer Applications in Medical Care*, 2:1326–1329, 1980.

[18] D E Broadbent. The hidden preattentive process. *American Psychology*, 32:109–118, 1977.

[19] S Brown, R Li, L Brandt, L Wilson, G Kossoff, and M Kossoff. Development of a multi-feature CAD system for mammography. In N Karssemeijer, M Thijssen, J Hendriks, and L van Erning, editors, *Procedings of the Fourth*

International Workshop on Digital Mammography, pages 189–196, Nijmegen, June 1998.

[20] D P Carmody, C F Nodine, and H L Kundel. An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*, 9:339–344, 1980.

[21] D P Carmody, C F Nodine, and H L Kundel. Finding lung nodules with and without comparative visual scanning. *Perception and Psychophysics*, 26(6):594–598, 1981.

[22] J Chamberlain, S M Moss, A E Kirkpatric, M Michell, and L Johns. National health service breast screening programme results for 1991-2. *British Medical Journal*, 307:353–356, 1993.

[23] H Chan, K Doi, C J Vyborny, R A Schmidt, C E Metz, K L Lam, T Ogura, Y Wu, and H MacMahon. Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis. *Investigative Radiology*, 25:1102–1110, 1990.

[24] E E Christenson, R C Murry, K Holland, J Reynolds, M J Landay, and J G Moore. The effect of search time on perception. *Radiology*, 138:361–365, 1981.

[25] 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.

[26] A V Cicourel. The integration of distributed knowledge in collaborative medical diagnosis. In J Galegher, R E Kraut, and C Egido, editors, *Intellectual Teamwork: Social and Technical Foundations of Cooperative Work*, pages 221–242. Cambridge University Press, 1990.

[27] P D Clayton and G Hripcsak. Decision support in healthcare. *International Journal of Bio-Medical Computing*, 39:59–66, 1995.

[28] W A Cormack. The PROMAM project. System Development Note 3.1 (Unpublished PROMAM project internal document), February 1997.

[29] H Cowley, A Gale, and A R M Wilson. Mammographic training sets for improving breast cancer detection. In *Proceedings of SPIE: Medical imaging: Image perception*, volume 1712, pages 102–112, 1996.

[30] H C Cowley and A Gale. Time of day effects on mammographic film reading performance. In *Proceedings of SPIE: Medical imaging: Image perception*, volume 3036, pages 212–221, 1997.

[31] B de Lafontan, J P Daures, B Salicru, F Eynius, J Mihura, P Rouanet, J L Lamarque, A Naja, and H Pujol. Isolated clustered microcalcifications: Diagnostic value of mammography — series of 400 cases with surgical verification. *Radiology*, 190:479–483, 1994.

[32] E S de Paredes. Radiographic breast anatomy: Radiologic signs of breast cancer. In A G Haus and M J Yaffe, editors, *RSNA Categorical Course in Physics. Technical Aspects of Breast Imaging*, pages 35–46. RSNA, 2nd edition, 1993.

[33] 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.

[34] E R E Denton and S Field. Just how valuable is double reporting in screening mammography? *Clinical Radiology*, 52:466–468, 1997.

[35] P Doubilet and G Herman. Interpretation of radiographs: Effect of clinical history. *American Journal of Roentgenology*, 137:1055–1058, 1981.

[36] C J Downing and S Pinker. Spatial structure of attention. In M I Posner and O S M Martin, editors, *Proceedings of the 11th International Symposium on Attention and Performance*, volume 2, pages 171–187, Hillsdale, New Jersey, 1985. Lawrence Erlbaum Associates.

[37] A A Duncan and M G Wallis. Classifying interval cancers. *Clinical Radiology*, 50:774–777, 1995.

[38] J Duncan. Selective attention and the organisation of visual information. *Journal of Experimental Psychology: General*, 113(4):501–517, 1984.

[39] H E Egeth and S Yantis. Visual attention: Control, representation, and time course. *Annual Review of Psychology*, 48:269–297, 1997.

[40] T K P Egglin and A R Feinstein. Context bias - a problem in diagnostic radiology. *Journal of the American Medical Association*, 276(21):1752–1755, 1996.

[41] J G Elmore, C K Wells, D H Howard, and A R Feinstein. The impact of clinical history on mammographic interpretations. *Journal of the American Medical Association*, 277(1):49–52, 1997.

[42] J G Elmore, C K Wells, C H Lee, D H Howard, and A R Feinstein. Variability in radiologists' interpretation of mammograms. *The New England Journal of Medicine*, 331(22):1493–1499, 1994.

[43] Y Engeström. Objects, contradictions and collaboration in medical cognition: An activity-theoretical perspective. *Artificial Intelligence in Medicine*, 7:395–412, 1995.

[44] D Fafchamps, C Y Young, and P C Tang. Modelling work practices: Input to the design of a physicians workstation. *Proceedings of the 15th Syposium on Computer Applications in Medical Care*, pages 788–792, 1991.

[45] C L Folk, R W Remmington, and J C Johnston. Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, 18(4):1030–1044, 1992.

[46] P Forrest. *Breast Cancer Screening. Report to the Health Ministers of England, Wales and Northern Ireland by a working group chaired by Professor Sir Patrick Forrest.* HMSO, London, 1986.

[47] D E Forsythe. Blaming the user in medical inforamtics: The cultural nature of scientific practice. *Knowledge and Society*, 9:95–111, 1992.

[48] D E Forsythe. Using ethnography to build a working system: Rethinking basic design assumptions. *Proceedings of the 16th Symposium on Computer Applications in Medical Care*, pages 505–509, 1992.

[49] D E Forsythe. "It's just a matter of common sense": Ethnography as invisible work. *Computer Supported Cooperative work*, 8:127–145, 1999.

[50] D E Forsythe and B G Buchanan. Knowledge aquisition for expect systems: Some pitfalls and suggestions. *IEEE Transactions on Systems, Man and Cybernetics*, 19(3):435–442, 1989.

[51] D E Forsythe and B G Buchanan. Broadening our approach to evaluating medical systems. *Proceedings of the 15th Symposium on Computer Applications in Medical Care*, pages 8–12, 1991.

[52] C P Friedman and J C Wyatt. Subjectivist approaches to evaluation. In *Evaluation Methods in Medical Informatics*, chapter 8, pages 205–221. Springer-Verlag, 1997.

[53] A G Gale. Human responses to visual stimuli. In W R Hendee and P T N Wells, editors, *The Perception of Visual Information*, pages 115–129. Springer-Verlag, 1995.

[54] A G Gale, F Johnson, and B S Worthington. Psychology and radiology. In D J Oborne, M M Gruneburgh, and J R Eiser, editors, *Research in Psychology and Medicine*, volume 1, pages 453–460. Academic, 1979.

[55] A G Gale, C J Savage, E F Pawley, A R M Wilson, and E J Roebuck. Breast screening: Visual search and observer performance. In *Proceedings of SPIE: Medical imaging: Image perception*, volume 2166, pages 66–75, 1994.

[56] H Garfinkel. *Studies in Ethnomethodology*, chapter 1. Prentice-Hall, 1967.

[57] D J Getty, R M Pickett, C J D'Orsi, and J A Swetts. Enhanced interpretation of diagnostic images. *Investigative Radiology*, 23:240–252, 1988.

[58] M L Giger. Computer-aided diagnosis. In A G Haus and M J Yaffe, editors, *RSNA Categorical Course in Physics. Technical Aspects of Breast Imaging*, pages 283–298. RSNA, 2nd edition, 1993.

[59] R N Haber. Nature of the effect of set on perception. *Psychological Review*, 73(4):335–351, 1966.

[60] M Hammersley and P Atkinson. *Ethnography, Principles in Practice*. Tavistock Publications Ltd, 1983.

[61] J Hartland. The use of 'intelligent' machines for electrocardiograph interpretation. In G Button, editor, *Technology in Working Order: Studies of Work Interaction and Technology*, chapter 3, pages 55–80. Routledge, 1993.

[62] H Heathfield, D Pitty, and R Hanka. Evaluating information technology in health care: Barriers and challenges. *British Medical Journal*, 316:1959–1961, 1998.

[63] H A Heathfield and J Wyatt. Philosophies for the design and development of clinical decision support systems. *Methods of Information in Medicine*, 32(1):1–8, 1993.

[64] S Heddle, A C Hume, and A E Kirkpatrick. Evaluation of a prompting system using interval cancers. In N Karssemeijer, M Thijssen, J Hendriks, and L van Erning, editors, *Procedings of the Fourth International Workshop on Digital Mammography*, pages 355–358, Nijmegen, June 1998.

[65] J E Hoffman and B Nelson. Spatial selectivity in visual search. *Perception and Psychophysics*, 30(3):283–290, 1981.

[66] E Hollnagel. Responsibility issues in intelligent decision support systems. In D Berry and A Hart, editors, *Expert Systems: Human Issues*, pages 237–249. Chapman and Hall, London, 1990.

[67] J Hughes, V King, T Rodden, and H Anderson. Moving out of the control room: Ethnography in system design. In *Transcending Boundaries: CSCW'94*. ACM Press, 1994.

[68] A Hume and P Thanisch. Microcalcification detection for mass screening programmes. In *IEEE Colloquium on Digital Mammography*, London, 1996.

[69] A Hume, P Thanisch, M Hartswood, and R Procter. On the evaluation of microcalcification detection algorithms. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Proceedings of the Third International Workshop on Digital Mammography*, pages 273–276, Chicago, June 1996.

[70] I Hutt. *The Computer-Aided Detection of Abnormalities in Digital Mammograms*. PhD thesis, University of Manchester, Faculty of Medicine, 1996.

[71] Y Jiang, R N Nishikawa, R A Schmidt, C E Metz, M L Giger, and K Doi. Benefits of computer-aided diagnosis (CAD) in mammographic diagnosis of malignant and benign clustered microcalcification. In N Karssemeijer, M Thijssen, J Hendricks, and L van Erning, editors, *The Forth International Workshop on Digital Mammography*, pages 215–220, Nijmegen, June 1998. Kluwer Academic Publishers.

[72] W A Johnston and V J Dark. Selective attention. *Annual Review of Psychology*, 37:43–75, 1986.

[73] J Jonides and S Yantis. Uniqueness of abrupt visual onset in capturing attention. *Perception and Psychophysics*, 43(4):346–354, 1998.

[74] D Kahneman. *Attention and Effort*. Prentice Hall, New Jersey, 1973.

[75] B Kaplan. The influence of medical values and practices on medical computer applications. *Use and Impact of Computers in Clinical Medicine*, pages 39–50, 1987.

[76] B Kaplan. Development and acceptance of medical information systems: An historical overview. *Journal of Health and Human Resources Administation*, 11(1):9–29, 1988.

[77] B Kaplan. Objectification and negociation in interpretting clinical images: Implications for computer-based patient records. *Artificial Intelligence in Medicine*, 7:439–495, 1995.

[78] N Karssemeijer. Stochastic model for automated detection of calcifications in digital mammograms. *Image Vision and Computing*, 10(6):368–375, 1992.

[79] N Karssemeijer. Adaptive noise equalisation and recognition of microcalcification clusters in mammograms. *International Journal of Pattern Recognition and Artificial Intelligence*, 1993.

[80] P G W Keen. Information systems and organizational change. *Communications of the ACM*, 24(1):24–33, 1981.

[81] C Kimme-Smith, L W Bassett, and R H Gold. *Workbook For Quality Mammography*. Williams and Wilkins, 1992.

[82] E Krupinski. Visual scanning patterns of radiologists searching mammograms. *Academic Radiology*, 3(2):137–144, 1996.

[83] E A Krupinski. Influence of experience on scanning strategies in mammography. In *Proceedings of SPIE: Medical imaging: Image perception*, volume 1712, pages 95–101, 1996.

[84] H L Kundel and P S La Follette. Visual search patterns and experience with radiological images. *Radiology*, 103:523–528, 1972.

[85] H L Kundel and C F Nodine. Interpreting chest radiographs without visual search. *Radiology*, 116:527–532, 1975.

[86] H L Kundel, C F Nodine, and D Carmody. The influence of prior knowledge on visual search strategies during the viewing of chest radiographs. *Radiology*, 92:315–320, 1969.

[87] H L Kundel, C F Nodine, and D Carmody. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, 38:175–181, 1978.

[88] H L Kundel, C F Nodine, and L Toto. Searching for lung nodules: The guidance of visual scanning. *Investigative Radiology*, 26(9):777–781, 1991.

[89] A Lesgold, R Glaser, H Rubinson, D Klopfer, P Feltovich, and Y Wang. Expertise in a complex skill: Diagnosing x-ray pictures. In M T H Chi, R Glaser, and M J Farr, editors, *The Nature of Expertise*, chapter 11, pages 311–342. Lawrence Earlbaum Associates, Hillsdale, New Jersey, 1998.

[90] N H Mackworth. Stimulus density limits the useful field of view. In R A Monty and J W Senders, editors, *Eye Movement and Psychological Process*, pages 307–321. Erlbaum, Hillsdale NJ, 1976.

[91] B J McNeal and J A Hanley. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making*, 4(2):137–150, 1984.

[92] C E Metz. ROC methodology in radiological imaging. *Investigative Radiology*, 21(9):720–733, 1986.

[93] C E Metz. Evaluation of digital mammography by ROC analysis. In K doi, M L Giger, R N Nishikawa, and R A Schmidt, editors, *Procedings of the Third Internation Workshop on Digital Mammography*, pages 61–68, Chicago, June 1996. Elsivier.

[94] 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.

[95] A B Miller. Principles of screening and of the evaluation of screening programmes. In A B Miller, editor, *Screening for Cancer*, chapter 1, pages 3–24. Academic Press, London, 1985.

[96] L Miller and N Ramsay. The detection of malignant masses by non-linear multiscale analysis. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Proceedings of the Third International Workshop on Digital Mammography*, pages 335–340, Chicago, June 1996.

[97] R A Miller. Medical diagnostic decision support systems — past present and future. *Journal of the American Medical Informatics Association*, 1(1):8–27, 1994.

[98] M D Mugglestone, A G Gale, H C Cowley, and A R M Wilson. Diagnostic performance on briefly presented mammographic images. In *Proceedings of SPIE: Medical imaging: Image perception*, volume 2436, pages 106–115, 1995.

[99] M D Mugglestone, R Lomax, A G Gale, and A R M Wilson. The effect of prompting mammographic abnormalities on the human observer. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *The 3rd International Conference on Digital Mammography*, pages 87–92, Chicago, June 1996. Elsevier.

[100] B M Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man Machine Studies*, 27:527–539, 1987.

[101] U Neisser. *Cognition and Reality. Principles and Implications of Cognitive Psychology.* W H Freeman, San Francisco, 1976.

[102] R M Nishikawa, D E Wolverton, R A Schmidt, and J Papaioannou. Radiologists' ability to discriminate computer-detected true and false positives from an automated scheme for detection of clustered microcalcifications on digital mammography. In *Procedings of SPIE. Medical Imaging: Image Perception*, volume 3036, pages 198–204, 1997.

[103] C F Nodine, E A Krupinski, H L Kundel, L Toto, and G T Herman. Letter. *Investigative Radiology*, 27(7):571–573, 1992.

[104] C F Nodine and H L Kundel. The cognitive side of visual search in radiology. In J K O'Regan and A Lévi-Shoen, editors, *Eye Movements: From Physiology to Cognition*, pages 573–582. Elsevier Science Publishers, 1987.

[105] C F Nodine, H L Kundel, S C Lauver, and Lawreence C Toto. Nature of expertise in searching mammograms for breast masses. *Academic Radiology*, 3(12):1000–1006, 1996.

[106] C F Nodine, H L Kundel, S C Lauver, and Lawreence C Toto. The nature of expertise in searching mammograms for breast masses. In *Proceedings of SPIE: Medical imaging: Image perception*, volume 1712, pages 89–94, 1996.

[107] J Ost and P Antweiler. The social impact of high cost medical technology: Issues and conflicts surrounding the decision to adopt CAT scanners. In J A Roth and S B Ruzek, editors, *Research in the Sociology of Healthcare*, volume 4, pages 33–92. Jai Press, 1986.

[108] M I Posner, C R R Snyder, and B J Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109(2):160–174, 1980.

[109] R Procter, P Thanisch, A Hume, A Kirkpatric, S Astley, I Hutt, and P Hauke. User interface design and data management for digital mass mammography. In A G Gale, S Astley, D R Dance, and A Y Cairns, editors, *Proceedings of the 2nd International Workshop on Digital Mammography*, pages 415–422, York, July 1994.

[110] G I Rochlin. "High-reliability" organisations and technical change: Some ethical problems and dilemmas. *IEEE Technology and Society Magazine*, pages 3–9, September 1986.

[111] I Rock and D Gutman. The effect of inattention on form perception. *Journal of Experimental Psychology*, 7(2):275–285, 1981.

[112] S Samuel, H L Kundel, C F Nodine, and L C Toto. Mechanism of satisfaction of search: Eye position recordings in the reading of chest radiographs. *Radiology*, 194:895–902, 1995.

[113] W B Schwartz. Medicine and the computer. *The New England Journal of Medicine*, 283:1257–1264, 1970.

[114] W Simpson, F Neilson, and J R Young. The identification of false negatives in a population of interval cancers: A method for audit of screening mammography. *The Breast*, 4:183–188, 1995.

[115] I Sommerville, T Rodden, P Sawyer, and R Bently. Sociologists can be suprisingly useful in interactive system design. Technical Report CSCW/1/92, Computing Department, Lancaster University, 1992.

[116] L Stark, I Yamashita, G Tharp, and H X Ngo. Keynote lecture: Search patterns and search paths in human visual search. In D Brogan, A Gale, and K Carr, editors, *Visual Search 2*, pages 37–58. Taylor and Francis, London, 1993.

[117] P Strax. Mammography. In A B Miller, editor, *Screening for Cancer*, pages 141–161. Academic Press, London, 1985.

[118] G C Sutton. Computer-aided diagnosis: A review. *British Journal of Surgery*, 76:82–85, 1989.

[119] B Swanson, I B McIntosh, K G Power, and H Dobson. The psychological effects of breast screening in terms of patients perceived health anxieties. *British Journal of Clinical Practice*, 50(3):129–135, 1996.

[120] P Taylor, J Fox, and A Todd-Pokropek. Evaluation of a decision aid for the classification of microcalcifications. In N Karssemeijer, M Thijssen, J Hendricks, and L van Erning, editors, *The Forth International Workshop on Digital Mammography*, pages 237–244, Nijmegen, June 1998. Kluwer Academic Publishers.

[121] Randy L Teach and Edward H Shortliffe. An analysis of physician attitudes regarding computer-based consultation systems. *Computer and Biomedical Research*, 14:542–558, 1981.

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

[123] A Treisman. Preattentive processing in vision. *Computer vision, graphics and image processing*, 31(2):156–177, 1985.

[124] W J Tuddenham. Visual search, image organisation and reader error in roentgen diagnosis. *Radiology*, 78:694–704, 1962.

[125] E C M van de Weijgert. Foveal load and peripheral task performance: Tunnel vision or general interference. In D Brogan, A Gale, and K Carr, editors, *Visual Search 2*, pages 341–348. Taylor and Francis, London, 1993.

[126] S P Vecera and M J Farah. Is visual image segmentation a bottom-up or an interactive process? *Perception and Psychophysics*, 59(8):1280–1296, 1997.

[127] 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*, 311:1189–1192, November 1995.

[128] 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.

[129] J C Wells and J Cooke. Film reading practice of the UK breast screening units. *The Breast*, 5:404–409, 1996.

[130] L Williams, M Hartswood, and R Prescott. Methodological issues in mammography double reading studies. *Journal of Medical Screening*, 5(202-206), 1998.

[131] L J Williams, R J Prescott, and M Hartswood. Computer-aided cancer detection in the UK breast screening programme. In N Karssemeijer, M Thijssen, J Hendriks, and L van Erning, editors, *Procedings of the Fourth International Workshop on Digital Mammography*, pages 359–362, Nijmegen, June 1998.

[132] M D Mugglestone A G Gale A R M Wilson. Perceptual processes involved in mammographic film interpretation. In *Proceedings of SPIE: Medical imaging: Image perception*, volume 3036, pages 188–197, 1997.

[133] J M Wolfe. Guided search 2.0. a revised model of visual search. *Psychonomic Bulletin and Review*, 1(2):202–238, 1994.

[134] J M Wolfe. Visual search in continuous natural stimuli. *Vision Research*, 34(9):1187–1195, 1994.

[135] J M Wolfe. The pertinence of research on visual search to radiological practice. *Academic Radiology*, 2:74–78, 1995.

[136] J Wyatt and D Spiegelhalter. Evaluating medical expert systems: What to test and how? *Medical Informatics*, 15(3):205–217, 1990.

[137] M J Yaffe. Digital mammography. In A G Haus and M J Yaffe, editors, *RSNA Categorical Course in Physics. Technical Aspects of Breast Imaging*, pages 271–282. RSNA, 2nd edition, 1993.

[138] D W Young. What makes doctors use computers?: Discussion paper. *Journal of the Royal Society of Medicine*, 77:663–667, August 1984.

Appendix A

Likert Test

Likert test attempts to measure a subject's attitude towards a particular event, object or topic by asking them to rate their agreement to a series of pertinent statements. Usually a mix of statements is used, representing both positive and negative views.

In the subjective responses to prompting experiment and in the pre-clinical trial, a Likert test was used to determine subjects' disposition towards the PROMAM system after each exposure. Subjects were asked to indicate if they strongly agreed, agreed, were uncertain, disagreed or strongly disagreed with each statement in the list shown below.

For each statement expressing a positive view, a score of five was given for strong agreement, a score of four for agreement, and so on. Statements representing a negative view were scored in the opposite sense: five if they strongly disagreed, four if they disagreed, three if they were uncertain, and so on. These scores were then summed to give an overall measure of attitude where a higher score indicates a more favourable the assessment.

Statements used in the Likert test:

1. The system will be time consuming to use.
2. The information supplied by the system was distracting.
3. The prompts were confusing.
4. Many changes would be required before the system would be useful.
5. The system is inaccurate.
6. The system gives useful information.

7. Prompts were no better than random.
8. The system will be of no use to me as an aid for reporting.
9. The system makes me more confident that I will find any cancers.
10. It was clear what features the prompts referred to.
11. Too many false positive prompts were produced.
12. I would be keen to use the system.
13. I would recommend the system to my colleagues.
14. I would be happy using the system as it currently operates.
15. The system performed better than I had expected.
16. It is obvious what the system was prompting for.
17. Using the system was satisfying.
18. The effort needed to use the system was not justified by the benefits of using the system.

Appendix B

Answers to free-form response questions

Responses to the free form response questions from the post-session questionnaires for the pre-clinical trials are given below. Each statement is prefixed by the subject's identifier (A to E) and the session number.

What do you think the system's strengths are?

A S3 Good for calcs. Quite good for masses.

C S4 It doesn't miss much.

B S6 Draws attention to potential abnormalities.

B S7 Does not tire or get distracted. Does not get too concentrated on one abnormality to the exclusion of others.

E S9 Easy to use. Does not seem to interfere with prompting style.

E S11 Generally easy to use.

E S18 Generally easy to use. Help to become certain of 'no lesions'.

A S20 Detects nearly all micro-calcs.

D S25 Prompting calcifications which might be over-looked.

A S30 Successfully prompts for calcifications in most cases.

C S32 Occasionally spotting a "mass" or a small cluster of calcs I hadn't noticed (or dismissed — makes me think again).

E S36 Appears good at detecting significant micro-calcs.

B S38 Could consolidate suspicion of a particular area. Could draw attention to the 2nd lesion.

E S40 Good at picking up micro-calcs.

What do you think the system's weaknesses are?

A S3 Lack of distortion & asymm.

C S4 It can't look at previous films for comparison.

B S6 Could over do it or give a false sense of security — ie by ignoring prompts if there are too many.

B S7 Slows process down.

E S9 Rather too sensitive to normal structures eg vessels. Potential to falsely reassure.

E S11 Multiple, "normal" prompts.

C S13 (down arrow) specific. (up arrow) sensitive.

E S18 Its problems with sensitivities to normal structures / artifacts.

A S20 Misses some masses.

C S21 Too many prompts for the same area of micro-calcification (one would do).

D S25 High prompt rate, particularly for calcification which I can't identify.

A S30 (1) Some calc prompts where no calcs exist. (2) Prompts for vascular calcs. (3) Misses some masses which should be prompted.

C S32 Over-calling.

B S38 Danger of over-prompting therefore making each less valuable. Slows the process down significantly.

E S40 Too many prompts on normal structures which can lead to irritation when lot of prompts present. Mass detection seems very non specific.

What irritated you most about the system?

A S3 False prompts for composite shadows.

C S4 The "pectoral intersection" problem — prompting for things that just didn't exist.

B S6 Probably slowing me down.

B S7 Trying to justify some of the prompts.

E S9 Multiple prompts, when present, and normal structures.

E S11 Prompts on normal structures, particularly when there are many of them — delayed reporting a little.

C S13 Lots of prompts for m/c that I couldn't see.

E S18 Too many "normal" prompts.

C S32 Vascular calcification prompts +++++

E S36 Too many false positive prompts.

B S38 Trying to rationalise some of the prompts. Slowing effect. Still prompts for vascular calcs.

E S40 As above.

What aspects of the system did you find most useful?

A S3 Prompts for calcs.

C S4 It found one small cluster of m/c that I hadn't spotted.

B S6 If nothing prompted it could be quite reassuring.

B S7 Drawing attention to other abnormalities besides the most obvious one.

E S11 High lighting of areas to review, affected my decision in a few cases, to recall.

C S13 Small mass prompts.

A S20 None stands out in particular. Overall good.

A S30 Calc prompts.

B S38 Firming up on some questionable areas. Negative prompts could be reassuring (? over-reassuring).

Can you suggest how the system might be improved?

A S3 Develop algorithms for distortion + asymm.

C S4 It needs to get more specific for both categories.

B S7 (Correlation?) between oblique and CC's, also between current and previous mammos. Detection of architectural distortion/stellate lesions.

C S13 Use of previous films. Somehow exclude the vascular m/c prompts.

A S20 (up arrow) mass sensitivity.

(C S21] (up arrow) specific + somehow get rid of the multiple vascular micro-calcification prompts.

A S30 Detection of D of A.

C S32 As usual, by (down arrow) sensitivity, (up arrow) specificity + using previous films.

B S38 Detection of stellate distortions. Removal of vascular calcs. Correlation of past/present, Obl/CC. Overlay — might be feasible on video screen.

Appendix C

Materials used in experimental work

C.1 Clinic questionnaire

This questionnaire was administered to 16 film readers in 5 of the 6 screening centres studied as part of the work practice investigation reported in Chapter 3.

Questionnaire for radiologists

Introduction

All radiologists participating in the *PROMAM* clinical trials are being asked to complete this questionnaire. The aim of which is to explore attitudes towards current double reading practices, and the potential usefulness of a prompting system such as *PROMAM*. Hopefully the results of the questionnaire will assist out development work in the following ways:

- The design of clinical trials.
- To help determine the requirements for a commercial system.
- Ongoing research into radiologists' best working practices.

Confidentiality

The answers you provide in this questionnaire will be treated as confidential. In any subsequent reporting of the results of this questionnaire, the anonymity of the radiologists and trial centres will be maintained.

Profile

How many years of screening experience do you have? ☐

A number of types of reading practice are listed below. Please state which of these you have been involved with, and for how long (including those at any previous positions in other clinics).

1. Single reading
2. Double reading
   - 2a Worst opinion recalls
   - 2b Flagged cases are discussed by both readers before recalls are decided
   - 2c Recall of flagged cases is decided by a senior radiologist
   - 2d A third opinion is sought where there is a disagreement over recalls between the first two radiologists.
   - 2e Other

If 'Other', please specify:

	Clinic	Duration	Reading Practice
eg	*Edinburgh*	*2yrs*	*2a*

Attitudes towards double reading

Each of the following statements give an opinion regarding double reading. Please rate your agreement with each statement as: Strongly agree, Agree, Uncertain, Disagree, or Strongly disagree, by ticking one box only.

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
When reading first, I am likely to be more vigilant - because I wouldn't want to miss anything that the second reader might then pick up.	☐	☐	☐	☐	☐
A reader's vigilance can be strongly dependent upon whether they are the first or second reader.	☐	☐	☐	☐	☐
If I am reading second, and the first reader is relatively inexperienced - I will be more vigilant than usual.	☐	☐	☐	☐	☐
As a second reader I can trust the first reader to have identified the majority of abnormalities.	☐	☐	☐	☐	☐
I am typically more vigilant as a second than as a first reader.	☐	☐	☐	☐	☐
If I were predominantly involved with double reading I will be extra vigilant if I am the reader for a series of films which will only be single read (eg because of, for example, illness or absentia).	☐	☐	☐	☐	☐
As a first reader will not be so vigilant because I know that the films will be examined again.	☐	☐	☐	☐	☐
A reader will always be equally vigilant - irrespective of whether they read first or second.	☐	☐	☐	☐	☐
As a second reader I am the 'last chance' for detecting an abnormality.	☐	☐	☐	☐	☐
If the radiologist who reads first is a trusted and experienced colleague, then as a second reader I would not be so vigilant.	☐	☐	☐	☐	☐
If I was predominantly involved in a single reading clinic, I would be more vigilant than if I were predominantly involved in a double reading clinic.	☐	☐	☐	☐	☐
As a first reader I am typically more vigilant than when reading second.	☐	☐	☐	☐	☐

Please rate the following tasks according to how difficult or how easy you find them:

	Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
Detection of micro-calcification clusters	☐	☐	☐	☐	☐
Detection of ill defined lesions	☐	☐	☐	☐	☐
Detection of architectural distortions	☐	☐	☐	☐	☐
Detection of asymmetries	☐	☐	☐	☐	☐
Classification of micro-calcifications	☐	☐	☐	☐	☐
Classification of ill defined lesions	☐	☐	☐	☐	☐
Others					
...	☐	☐	☐	☐	☐
...	☐	☐	☐	☐	☐

(Please tick one box per statement. Feel free to add additional features, and rate them accordingly)

When double reading: do you believe that an inexperienced reader should:

Always read first ☐

Always read second ☐

It is unimportant whether they read first or second ☐

(Please tick one box only)

Why?

Which system of reading do you believe to be the most effective:

Single reading ☐

Worst opinion recalls. ☐

Flagged cases are discussed by both readers before recalls are decided. ☐

Recall of flagged cases is decided by a senior radiologist. ☐

A third opinion is sought where there is a disagreement over recalls between the first two radiologists. ☐

(Please tick one box only)

Why?

If a system of double reading is used where there is some further scrutiny of flagged cases before recalls are decided - who do you think should have the final say?

The radiologist(s) who originally recommended the recall	☐
A senior radiologist	☐
Any of the clinics' radiologists	☐

(Please tick one box only)

Why?

In a double reading system, do you think it is important to:

	Yes	No
Randomly pair radiologists?	☐	☐
Pair radiologists so that strengths and weaknesses are appropriately matched?	☐	☐
To adhere consistently to specified pairings?	☐	☐
Change pairings frequently?	☐	☐

(Please tick either yes or no for each statement)

Could you explain your reasoning?

Do you think there is any advantage to be gained from:

	Yes	No
Blinding the second reader to the first reader's decision?	☐	☐
Blinding readers so that they are unaware as to whether they are the first or second reader?	☐	☐

(Please tick either yes or no for each statement)

Could you explain your reasoning?

In your opinion, what performance gain (in cancer detection rates) do you think double reading gives over single reading (approximately):

- No better ☐
- 5% ☐
- 10% ☐
- 15% ☐
- 20% ☐

(Please tick one box only)

What has influenced you most in forming your opinion as to the benefit of double reading?

- Published double reading studies ☐
- Your own personal experience ☐
- Experiences of colleagues reported to you personally ☐
- Other ☐

(Please tick one box only)

If 'Other', please specify:

Prompting systems

Below are listed hypothetical properties of a prompting system, rate each in terms of how useful you perceive they might be in a screening practice:

	Essential	Useful	Doubtful	Of no use
Prompting for micro-calcification clusters	☐	☐	☐	☐
Prompting for ill defined lesions	☐	☐	☐	☐
Prompting for architectural distortions	☐	☐	☐	☐
Prompting for asymmetries	☐	☐	☐	☐
Classification of prompted micro-calcifications from benign to malignant	☐	☐	☐	☐
Classification of prompted masses from benign to malignant	☐	☐	☐	☐
Others				
...	☐	☐	☐	☐
...	☐	☐	☐	☐

(Please tick one box per statement. Feel free to add additional properties, and rate them accordingly)

Given that the capabilities of a prompting system are likely to evolve with time, prioritise following: (1 indicates the function should be developed first, 2 second etc. Use each number only once)

Prompting for micro-calcifications	☐
Prompting for ill defined lesions	☐
Prompting for architectural distortions	☐
Prompting for asymmetries	☐
Classification of prompted micro-calcifications as benign to malignant.	☐
Classification of prompted masses as benign or malignant	☐
Other	
...	☐
...	☐

(Feel free to add additional properties, and number them accordingly)

In a screening practice, what problems do you see a prompting system addressing? (Please rate the following in importance: 1 = most important, 2 = second in importance etc. Please use each number only once)

Reducing the number of interval cancers (false negatives) ☐

Improving the detection performance of a single reader. ☐

Improving the consistency of reading (eg compensating for fatigue) ☐

Supporting inexperienced radiologists? ☐

Addressing resourcing limitations (eg availability of radiologists) ☐

Reducing recalls (if classification available). ☐

Other

... ☐

... ☐

(Feel free to add any additional roles, and number them accordingly)

How do you see a prompting system being used in your screening clinic:

Replacing double reading with single reading and a prompting system ☐

Using the prompting system to enhance double reading. ☐

Other ☐

(Please tick one box only)

If other, please specify:

Thank-you for the time you have taken to complete this questionnaire.

C.2 Materials used in the 'subjective responses to prompting' experiment

This section contains materials used in the 'subjective responses to prompting' experiment, including:

1. Subjects' instruction sheets.
2. The pre-experiment questionnaire.
3. The post-session questionnaire.
4. The post-experiment questionnaire.

C.2.1 Instructions

ProMam prompting experiment

Introduction

You will be asked to report four sets of films on different days. Three of the sets will be prompted at different rates by the ProMam system - corresponding to system sensitivities of high, medium and low. The fourth set will be unprompted. Each of the conditions have been created by randomly selecting from the output from Ardmillan House, and should by typical of what you might see during a normal reading session. Previous screening films and CC views for first time screeners will be unavailable to you during the experiment.

Before reporting each condition proper you will be asked to report five cases to ensure familiarity with the prompting system and the reporting regime.

All the films you will see have been previously digitised and analysed by the ProMam system in order to detect potential abnormalities. The result of this process is a prompt sheet, an A4 piece of paper with a low resolution image of the mammogram pair. If a potential abnormality has been detected by the system, then an outline drawing will be present on the prompt sheet depicting the size and location of the lesion. (Example prompt sheets are given over-leaf).

For the purposes of this experiment, the system will attempt to detect and prompt for micro-calcification clusters, and masses. Prompts for potential masses will always appear as circles or ellipses, prompts for micro-calcifications will appear as irregular curved shapes that trace out the region containing the calcification.

Example prompt from the mass algorithm

Example prompt from the calcification algorithm

One prompt sheet will be produced for every case (excepting the unprompted condition) whether or not the system has detected an abnormality (ie whether or not there are any prompts drawn).

The system is not 100% sensitive, nor is it 100% specific - the majority of the prompts will be 'false positives'. You will be told the approximate sensitivity of the system (high, medium or low) and the prompt rate for the condition you are reading.

You will be asked to complete a questionnaire at the beginning of the experiment, and after you have reported all four conditions. You will also be asked to complete a questionnaire at the end of each condition. The time taken to report each condition will be recorded.

Please feel free to ask any questions.

Reading protocol - Prompted

Please report each cases in the order in that they are supplied observing the following protocol:

1. Examine the films
2. Examine the prompt sheet (by lifting the the reporting sheet).
3. Mark your decision on the reporting form as:
   - Routine recall
   - Technical recall
   - Review
4. Move onto the next case

Please examine and report each case before moving on to the next. Do not examine a 'batch' of cases before writing down your decision.

When making your decision assume that you are either a first or second reader in a blinded double reading system. Assume also that recalls for assessment will be made on a 'worst decision recalls' basis.

For any cases that you recommend to be recalled for assessment:

1. Annotate the reporting form as you normally would (eg by marking the position and type of lesion on the breast schematic).
2. Complete the 'Abnormality prompted for?' box on the reporting form - enter 'Y' if there is a prompt for that abnormality, and 'N' otherwise.

Approximate average prompt rates for this set

The **mass** detection algorithm: **1 prompt in every 2 cases**

The **calcification** detection algorithm: **1 prompt in every 3 cases**

The **sensitivity** of the system producing these prompts is: **High**

Reading protocol - Prompted

Please report each cases in the order in that they are supplied observing the following protocol:

1. Examine the films
2. Examine the prompt sheet (by lifting the the reporting sheet).
3. Mark your decision on the reporting form as:
   - Routine recall
   - Technical recall
   - Review
4. Move onto the next case

Please examine and report each case before moving on to the next. Do not examine a 'batch' of cases before writing down your decision.

When making your decision assume that you are either a first or second reader in a blinded double reading system. Assume also that recalls for assessment will be made on a 'worst decision recalls' basis.

For any cases that you recommend to be recalled for assessment:

1. Annotate the reporting form as you normally would (eg by marking the position and type of lesion on the breast schematic).
2. Complete the 'Abnormality prompted for?' box on the reporting form - enter 'Y' if there is a prompt for that abnormality, and 'N' otherwise.

Approximate average prompt rates for this set

The **mass** detection algorithm: **1 prompt in every 4 cases**

The **calcification** detection algorithm: **1 prompt in every 6 cases**

The **sensitivity** of the system producing these prompts is: **Medium**

Reading protocol - Prompted

Please report each cases in the order in that they are supplied observing the following protocol:

1. Examine the films
2. Examine the prompt sheet (by lifting the the reporting sheet).
3. Mark your decision on the reporting form as:
   - Routine recall
   - Technical recall
   - Review
4. Move onto the next case

Please examine and report each case before moving on to the next. Do not examine a 'batch' of cases before writing down your decision.

When making your decision assume that you are either a first or second reader in a blinded double reading system. Assume also that recalls for assessment will be made on a 'worst decision recalls' basis.

For any cases that you recommend to be recalled for assessment:

1. Annotate the reporting form as you normally would (eg by marking the position and type of lesion on the breast schematic).
2. Complete the 'Abnormality prompted for?' box on the reporting form - enter 'Y' if there is a prompt for that abnormality, and 'N' otherwise.

Approximate average prompt rates for this set

The **mass** detection algorithm: **1 prompt in every 8 cases**

The **calcification** detection algorithm: **1 prompt in every 12 cases**

The **sensitivity** of the system producing these prompts is: **Low**

Reading protocol - Unprompted

Please report the cases observing the following protocol for each case:

1. Examine the films
2. Mark your decision as:
    - Routine recall
    - Technical recall
    - Review
3. Move onto the next case

Please examine and report each case before moving on to the next. Do not examine a 'batch' of cases before writing down your decision.

When making your decision assume that you are either a first or second reader in a blinded double reading system. Assume also that recalls for assessment will be made on a 'worst decision recalls' basis.

For cases that you recommend to be recalled for assessment, annotate the reporting form as you normally would (eg marking the position and type of lesion on the breast schematic).

Prompted condition

You will be asked to report on three sets of films:

1. A short practice set.
2. Part one of the condition, after which you will be asked to take a fifteen minute break.
3. Part two of the condition.

Each set consists of a series of oblique view mammogram pairs. The protocol for reporting each case is described on a separate sheet.

1. The practice set

There are 5 cases in the practice set.

As this is a practice set, you are not being timed - also, please feel free to ask any questions.

2. Part one of the condition

There are 56 cases in part 1 of this condition.

Please spend as long as you feel is necessary over each case to reach your decision.

Please do not ask any questions once the experiment has been begun.

The time taken for you to report this set will be recorded. Please state when you have started reading, and when you have completed the condition, to assist with timing.

You will be requested to take a 15 minute break at the end of part 1, before beginning part 2.

3. Part two of the condition

There are 55 cases in part 2 of this condition.

Please spend as long as you feel is necessary over each case to reach your decision.

Please do not ask any questions once the experiment has been begun.

The time taken for you to report this set will be recorded. Please state when you have started reading, and when you have completed the condition, to assist with timing.

You will be asked to complete a questionnaire when you have completed part 2.

C.2.2 Pre-experiment questionnaire

Pre experiment questionnaire

Would you prefer a system which:

Has a high sensitivity but produces many false positives. ☐

A system which has a lower sensitivity but produces proportionally fewer false positives. ☐

Rate the following types of algorithm output on a scale of 1 to 5, where 1 means that being prompted for that feature would be useful to you, and 5 means that it would be distracting. (Please tick one box per feature)

	Useful 1	2	3	4	5 Distracting
Vascular calcification	☐	☐	☐	☐	☐
Benign clusters	☐	☐	☐	☐	☐
'Popcorn' calcification	☐	☐	☐	☐	☐
Film artifacts	☐	☐	☐	☐	☐
Lymph nodes	☐	☐	☐	☐	☐
Well defined masses	☐	☐	☐	☐	☐
Composite shadows	☐	☐	☐	☐	☐
Nodular glandular structure	☐	☐	☐	☐	☐
Cysts	☐	☐	☐	☐	☐
Other (Please state)					
...	☐	☐	☐	☐	☐
...	☐	☐	☐	☐	☐

1

Please rank the following categories of false positive as to the priority that should be given by algorithm developers to their removal (1 = the feature should be removed first, 2 = the feature should be removed 2nd, etc. Any number may be used more than once).

Vascular calcification	☐
Benign clusters	☐
'Popcorn' calcification	☐
Film artifacts	☐
Lymph nodes	☐
Well defined masses	☐
Composite shadows	☐
Nodular glandular structure	☐
Cysts	☐

Please rate your agreement with the following statements:

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
In cases where you are unsure, do you believe that					
The presence of a prompt will make you more inclined to recommend recall.	☐	☐	☐	☐	☐
The absence of a prompt makes you less likely to recommend recall	☐	☐	☐	☐	☐

Please give the following possible system configurations a rating on a scale of 1-5 as to how useful you believe each configuration to be (1 most useful, 5 least useful, tick one box only)

	Most useful 1	2	3	4	Least useful 5
High prompt rate, where most of the features prompted for are benign, but with a high probability that any malignancies will also be prompted for.	☐	☐	☐	☐	☐
Low prompt rate, where few of the prompts are for benign features, but with a high probability that some malignancies will be missed by the system.	☐	☐	☐	☐	☐
A system which is designed to prompt for micro-calcification clusters (whether malignant or benign) but not other types of calcification (eg vascular calcification, popcorn calcification).	☐	☐	☐	☐	☐
A system that will prompt for all types of calcification clusters, rather than one that tries to discard those with benign appearance.	☐	☐	☐	☐	☐
A system that will prompt for opacities that can usually be dismissed by radiologists with the aid of previous films or multiple views (eg composite shadows), as well opacities that are the result of a malignant process.	☐	☐	☐	☐	☐

Thank-you for completing this questionnaire

C.2.3 Post-experiment questionnaire

Post experiment questionnaire

Would you prefer a system which:

Has a high sensitivity but produces many false positives. ☐

A system which has a lower sensitivity but produces proportionally fewer false positives. ☐

Rate the following types of algorithm output on a scale of 1 to 5, where 1 means that being prompted for that feature would be useful to you, and 5 means that it would be distracting. (Please tick one box per feature)

	Useful				Distracting
	1	2	3	4	5
Vascular calcification	☐	☐	☐	☐	☐
Benign clusters	☐	☐	☐	☐	☐
'Popcorn' calcification	☐	☐	☐	☐	☐
Film artifacts	☐	☐	☐	☐	☐
Lymph nodes	☐	☐	☐	☐	☐
Well defined masses	☐	☐	☐	☐	☐
Composite shadows	☐	☐	☐	☐	☐
Nodular glandular structure	☐	☐	☐	☐	☐
Cysts	☐	☐	☐	☐	☐
Other (Please state)					
...	☐	☐	☐	☐	☐
...	☐	☐	☐	☐	☐

1

Please rank the following categories of false positive as to the priority that should be given by algorithm developers to their removal (1 = the feature should be removed first, 2 = the feature should be removed 2nd, etc. Any number may be used more than once).

Vascular calcification ☐

Benign clusters ☐

'Popcorn' calcification ☐

Film artifacts ☐

Lymph nodes ☐

Well defined masses ☐

Composite shadows ☐

Nodular glandular structure ☐

Cysts ☐

Please rate your agreement with the following statements:

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
In cases where you are unsure, do you believe that					
The presence of a prompt will make you more inclined to recommend recall.	☐	☐	☐	☐	☐
The absence of a prompt makes you less likely to recommend recall	☐	☐	☐	☐	☐

Please give the following possible system configurations a rating on a scale of 1-5 as to how useful you believe each configuration to be (1 most useful, 5 least useful, tick one box only)

	Most useful 1	2	3	4	5 Least useful
High prompt rate, where most of the features prompted for are benign, but with a high probability that any malignancies will also be prompted for.	☐	☐	☐	☐	☐
Low prompt rate, where few of the prompts are for benign features, but with a high probability that some malignancies will be missed by the system.	☐	☐	☐	☐	☐
A system which is designed to prompt for micro-calcification clusters (whether malignant or benign) but not other types of calcification (eg vascular calcification, popcorn calcification).	☐	☐	☐	☐	☐
A system that will prompt for all types of calcification clusters, rather than one that tries to discard those with benign appearance.	☐	☐	☐	☐	☐
A system that will prompt for opacities that can usually be dismissed by radiologists with the aid of previous films or multiple views (eg composite shadows), as well opacities that are the result of a malignant process.	☐	☐	☐	☐	☐

Of all the conditions you have completed, which do you believe would prove most useful to you in an actual screening context? (Tick one box only)

(Prompt rates: 1 case prompted in every x cases).

Mass Prompt Rate	Calcification prompt rate	Sensitivity	
1 in 2	1 in 3	High	☐
1 in 4	1 in 6	Medium	☐
1 in 8	1 in 12	Low	☐
No cases prompted	No cases prompted	-	☐

If you have any further comments with respect to any aspect of the experiment, please write them below:

Thank-you for completing this questionnaire

4

C.2.4 Post-session-questionnaire

Post condition questionnaire

All the questions in this questionnaire refer to the system configuration in the *condition you have just read*. Please answer all the questions with respect to this condition only.

Each of the following statements gives an opinion regarding the prompting system. Please state your agreement with respect to the condition you have just reported (Please tick one box per statement)

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
This system will be time consuming to use.	☐	☐	☐	☐	☐
The information supplied by this system was distracting.	☐	☐	☐	☐	☐
The prompts were confusing.	☐	☐	☐	☐	☐
Many changes would required before this system would be useful.	☐	☐	☐	☐	☐
This system is inaccurate.	☐	☐	☐	☐	☐
This system gives useful information.	☐	☐	☐	☐	☐
Prompts were no better than random.	☐	☐	☐	☐	☐
This system will be of no use to me as an aid for reporting.	☐	☐	☐	☐	☐
This system speeds up the reporting process.	☐	☐	☐	☐	☐
This system makes me more confident that I will find any cancers.	☐	☐	☐	☐	☐
It was clear what features the prompts referred to.	☐	☐	☐	☐	☐
Too many false positive prompts were produced.	☐	☐	☐	☐	☐

1

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
I would be keen to use this system.	☐	☐	☐	☐	☐
I would recommend this system to my colleagues.	☐	☐	☐	☐	☐
I would be happy using this system as it currently operates.	☐	☐	☐	☐	☐
This prompting system is effective.	☐	☐	☐	☐	☐
This system performed better than I had expected.	☐	☐	☐	☐	☐
It is obvious what this system was prompting for.	☐	☐	☐	☐	☐
Using this system was satisfying.	☐	☐	☐	☐	☐
The effort needed to use this system was not justified by the benefits of using the system.	☐	☐	☐	☐	☐

Overall rating:

What score would you give this system to indicate it's overall usefulness in a screening context? (Rate from 0-100, with 100 being the best possible score) ☐

Do you believe that: (Tick one box per question)

	Yes	No
Overall, this system would be useful to you in a screening context as it currently stands?	☐	☐
The mass detection component of this system would be useful to you as it currently stands?	☐	☐
The micro-calcification detection component of this system would be useful to you as it currently stands?	☐	☐

Please rate the system components: (Tick one box for each)

	Too sensitive	Just right	Not sensitive enough
Overall	☐	☐	☐
Masses	☐	☐	☐
Micro-calcification	☐	☐	☐

How would you rate the system you have just used if it had the following sensitivities? (Where, for example, 85% corresponds to 85% of malignant masses and malignant micro-calcification clusters being detected) Please tick one box per sensitivity setting.

	Very Useful	Useful	Doubtful	Of no use
95%	☐	☐	☐	☐
90%	☐	☐	☐	☐
85%	☐	☐	☐	☐
80%	☐	☐	☐	☐

General Impressions

What do you think the systems strengths are?

What do you think the systems weaknesses are?

What irritated you most about the system?

What aspects of the system did you find most useful?

Can you you suggest how the system might be improved?

Thank-you for completing this questionnaire

4

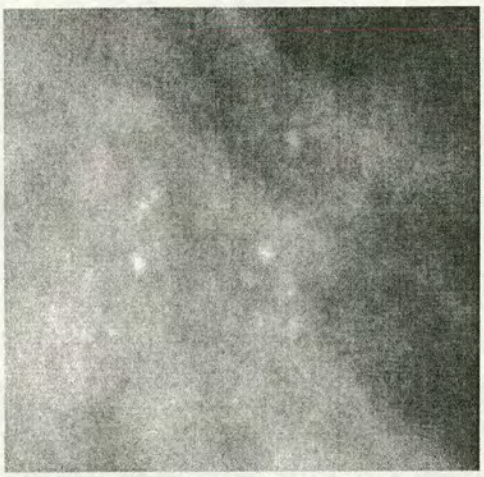
C.3 Materials used in the pre-clinical trials

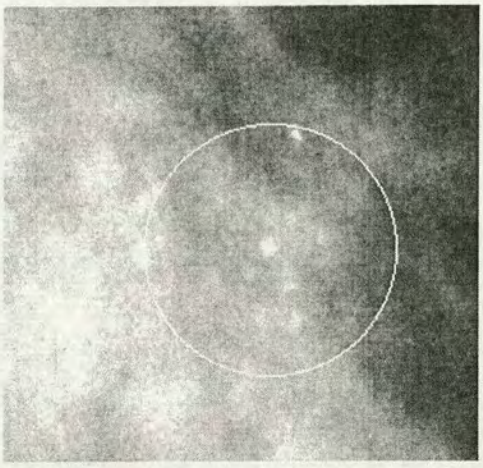
This section includes materials used in the pre-clinical trial, including:

1. Instructions.
2. Examples of training material (Produced by Ally Hume, the PROMAM team member responsible for developing the micro-calcification algorithm).
3. Training summary sheets.
4. Pre-trial questionnaire.
5. Post-session questionnaire.
6. Post-trial questionnaire.

C.3.1 Training examples

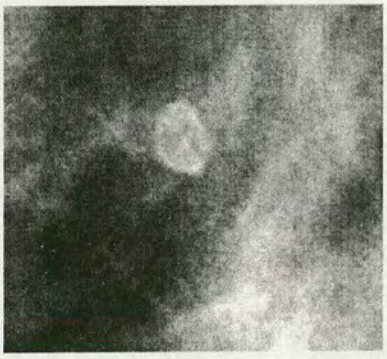
Microcalcification clusters not prompted

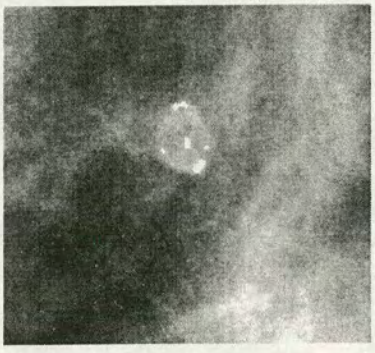




The computer only detects four of these diffuse calcifications.
These are grouped into two clusters of two and hence are not prompted
The clustering distance is only marginally too small to form a single cluster which would have been prompted.

False prompts for large benign calcifications and calcified cysts.





The computer looks for small subtle calcifications and can therefore occasionally detect the structure within a large calcification or calcified cyst.

In this example the computer has found five areas within the cyst which it thinks are microcalcifications and hence the calcified cyst is prompted as if it were a cluster of microcalcifications.

C.3.2 Training summary

Ill-defined lesion algorithm

What it detects

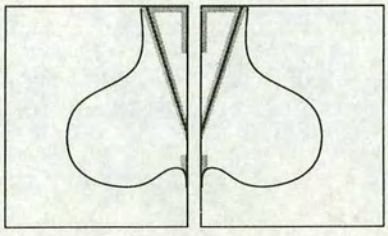
- Ill-defined, lobular, vague, fuzzy patches of density.
- Lesions between 5mm and 34mm (apparent size).

What it does not use as evidence

- Films from previous screening rounds
- The presence of spicules or tentacles
- Asymmetry
- Architectural distortion
- Skin thickening

Spatial limitations

The shaded areas in the diagram shows regions where the ill-defined lesion algorithm may fail to detect masses.



This only applies for oblique views where the pectoral muscle has been imaged (ie not for 'fronts' or CC views).

Features that may cause false positives

- Composite tissue
- Skin folds
- Blood vessel crossing the pectoral muscle
- Benign regions showing increased density may be miss-classified as suspicious

Microcalcification algorithm

What it detects

- Clusters of microcalcifications that consist of
  - five or more particles, which are
  - closer that 3.1mm together.

Features that may cause false positives

- Some vascular calcifications
- Some large benign calcifications
- Pectoral muscle edge – if not detected correctly
- Artifacts on the edge of the film
- Sometimes small overlapping linear structures may be mistaken for calcification particles.

C.3.3 Instructions

PROMAM - Prompting Experiment

- In total, there are 2000 cases to be double read over a period of two months (40 working days).
- The set includes oblique view mammograms, and CCs if it is a first time screen.
- This set contains normal cases (ie non-recalled cases), previously recalled cases and pathology proven cancers.
- We have included a higher proportion of cancers than you might expect to find in routine screening.
- All the cases in this set have been scanned, and subsequently analysed by the microcalcification and the ill-defined lesion detection algorithm.
- For reading, the set has been split into subsets — each containing 100 cases.
- The number of malignant cases per subset will vary. There may be subsets that contain no malignant cases.
- Each subset will be blind double read.
- One radiologist in each double reading pair will read with prompt sheet, and the other will read unaided. For some subsets the first reader will be prompted, for others the second reader will be prompted.
- CC films and previous films will be available for reading when appropriate.
- Previous films have not been used by the computer system to assist in the detection of lesions, neither are prompts produced for previous films. Previous cases are presented solely to assist interpretation by radiologists.

Reading protocol - Prompted

Please report each case observing the following protocol:

1. Examine the films
2. Examine the prompt sheet (by lifting the the reporting sheet).
3. Tick the box labelled 'Examined' on the prompting form to indicate that the prompting information has been considered.
4. Mark your decision on the reporting form as one of:
   - Routine recall
   - Technical recall
   - Review
5. Move onto the next case

When making your decision assume that you are either a first or second reader in a blinded double reading system. Assume also that recalls for assessment will be made on a 'worst decision recalls' basis.

For any cases that you recommend to be recalled for assessment:

1. Annotate the reporting form by marking the position and type of lesion on the breast schematic.
2. Complete the 'Abnormality prompted for?' box on the reporting form - enter 'Y' if there is a prompt for that abnormality, and 'N' otherwise.

Approximate prompt rates and sensitivities for the prompting system:

	Approx. prompting rate	**Sensitivity**
Masses	1 case in 2 prompted	81%
Microcalcifications	1 case in 4 prompted	90%

Reading protocol - Unprompted

Please report each case observing the following protocol:

1. Examine the films
2. Mark your decision as:
    - Routine recall
    - Technical recall
    - Review
3. Move onto the next case

When making your decision assume that you are either a first or second reader in a blinded double reading system. Assume also that recalls for assessment will be made on a 'worst decision recalls' basis.

For cases that you recommend to be recalled for assessment, annotate the reporting form by marking the position and type of lesion on the breast schematic.

C.3.4 Pre-experiment questionnaire

Pre experiment questionnaire

Rate the following types of prompted feature on a scale of 1 to 5, where 1 means that being prompted for that feature would be useful to you, and 5 means that it would be distracting. (Please tick one box per feature)

	Useful				Distracting
	1	2	3	4	5
Vascular calcification	☐	☐	☐	☐	☐
Benign clusters	☐	☐	☐	☐	☐
'Popcorn' calcification	☐	☐	☐	☐	☐
Film artifacts	☐	☐	☐	☐	☐
Lymph nodes	☐	☐	☐	☐	☐
Well defined masses	☐	☐	☐	☐	☐
Composite shadows	☐	☐	☐	☐	☐
Nodular glandular structure	☐	☐	☐	☐	☐
Cysts	☐	☐	☐	☐	☐
Other (Please state)					
...	☐	☐	☐	☐	☐
...	☐	☐	☐	☐	☐

1

Please rank the following categories of false positive as to the priority that should be given to their removal (1 = the feature should be removed first, 2 = the feature should be removed 2nd, etc. Please use each number only once.

Vascular calcification	☐
Benign clusters	☐
'Popcorn' calcification	☐
Film artifacts	☐
Lymph nodes	☐
Well defined masses	☐
Composite shadows	☐
Nodular glandular structure	☐
Cysts	☐

Please rate the following tasks according to how difficult or how easy you find them:

	Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
Detection of micro-calcification clusters	☐	☐	☐	☐	☐
Detection of ill defined lesions	☐	☐	☐	☐	☐
Detection of architectural distortions	☐	☐	☐	☐	☐
Detection of asymmetries	☐	☐	☐	☐	☐
Classification of micro-calcifications	☐	☐	☐	☐	☐
Classification of ill defined lesions	☐	☐	☐	☐	☐
Others					
...	☐	☐	☐	☐	☐
...	☐	☐	☐	☐	☐

(Please tick one box per statement. Feel free to add additional features, and rate them accordingly)

Below are listed hypothetical properties of a prompting system, rate each in terms of how useful you perceive they might be in a screening practice:

	Essential	Useful	Doubtful	Of no use
Prompting for micro-calcification clusters	☐	☐	☐	☐
Prompting for ill defined lesions	☐	☐	☐	☐
Prompting for architectural distortions	☐	☐	☐	☐
Prompting for asymmetries	☐	☐	☐	☐
Classification of prompted micro-calcifications from benign to malignant	☐	☐	☐	☐
Classification of prompted masses from benign to malignant	☐	☐	☐	☐
Others				
...	☐	☐	☐	☐
...	☐	☐	☐	☐

(Please tick one box per statement. Feel free to add additional properties, and rate them accordingly)

Given that the capabilities of a prompting system are likely to evolve with time, prioritise following: (1 indicates the function should be developed first, 2 second etc. Use each number only once)

Prompting for micro-calcifications	☐
Prompting for ill defined lesions	☐
Prompting for architectural distortions	☐
Prompting for asymmetries	☐
Classification of prompted micro-calcifications as benign to malignant.	☐
Classification of prompted masses as benign or malignant	☐
Other	
...	☐
...	☐

(Feel free to add additional properties. and number them accordingly)

In a screening practice, what problems do you see a prompting system addressing? (Please rate the following in importance: 1 = most important, 2 = second in importance etc. Please use each number only once)

Reducing the number of interval cancers (false negatives) ☐

Improving the detection performance of a single reader. ☐

Improving the consistency of reading (eg compensating for fatigue) ☐

Supporting inexperienced radiologists? ☐

Addressing resourcing limitations (eg availability of radiologists) ☐

Reducing recalls (if classification available). ☐

Other

... ☐

... ☐

(Feel free to add any additional roles, and number them accordingly)

How do you see a prompting system being used in your screening clinic:

Replacing double reading with single reading and a prompting system ☐

Using the prompting system to enhance double reading. ☐

Other ☐

(Please tick one box only)

If other, please specify:

Please rate your agreement with the following statements:

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
In cases where you are unsure, do you believe that					
The presence of a prompt will make you more inclined to recommend recall.	☐	☐	☐	☐	☐
The absence of a prompt makes you less likely to recommend recall	☐	☐	☐	☐	☐

Please give the following possible system configurations a rating on a scale of 1-5 as to how useful you believe each configuration to be (1 most useful, 5 least useful, tick one box only)

	Most useful 1	2	3	4	5 Least useful
High prompt rate, where most of the features prompted for are benign, but with a high probability that any malignancies will also be prompted for.	☐	☐	☐	☐	☐
Low prompt rate, where few of the prompts are for benign features, but with a high probability that some malignancies will be missed by the system.	☐	☐	☐	☐	☐
A system which is designed to prompt for micro-calcification clusters (whether malignant or benign) but not other types of calcification (eg vascular calcification, popcorn calcification).	☐	☐	☐	☐	☐
A system that will prompt for all types of calcification clusters, rather than one that tries to discard those with benign appearance.	☐	☐	☐	☐	☐
A system that will prompt for opacities that can usually be dismissed by radiologists with the aid of previous films or multiple views (eg composite shadows), as well opacities that are the result of a malignant process.	☐	☐	☐	☐	☐

Thankyou for completing this questionnaire

8

C.3.5 Post-experiment questionnaire

Post experiment questionnaire

Rate the following types of prompted feature on a scale of 1 to 5, where 1 means that being prompted for that feature would be useful to you, and 5 means that it would be distracting. (Please tick one box per feature)

	Useful				Distracting
	1	2	3	4	5
Vascular calcification	☐	☐	☐	☐	☐
Benign clusters	☐	☐	☐	☐	☐
'Popcorn' calcification	☐	☐	☐	☐	☐
Film artifacts	☐	☐	☐	☐	☐
Lymph nodes	☐	☐	☐	☐	☐
Well defined masses	☐	☐	☐	☐	☐
Composite shadows	☐	☐	☐	☐	☐
Nodular glandular structure	☐	☐	☐	☐	☐
Cysts	☐	☐	☐	☐	☐
Other (Please state)					
...	☐	☐	☐	☐	☐
...	☐	☐	☐	☐	☐

1

Please rank the following categories of false positive as to the priority that should be given to their removal (1 = the feature should be removed first, 2 = the feature should be removed 2nd, etc. Please use each number only once.

Vascular calcification	☐
Benign clusters	☐
'Popcorn' calcification	☐
Film artifacts	☐
Lymph nodes	☐
Well defined masses	☐
Composite shadows	☐
Nodular glandular structure	☐
Cysts	☐

Please rate the following tasks according to how difficult or how easy you find them:

	Very easy	Easy	Neither easy nor difficult	Difficult	Very difficult
Detection of micro-calcification clusters	☐	☐	☐	☐	☐
Detection of ill defined lesions	☐	☐	☐	☐	☐
Detection of architectural distortions	☐	☐	☐	☐	☐
Detection of asymmetries	☐	☐	☐	☐	☐
Classification of micro-calcifications	☐	☐	☐	☐	☐
Classification of ill defined lesions	☐	☐	☐	☐	☐
Others					
...	☐	☐	☐	☐	☐
...	☐	☐	☐	☐	☐

(Please tick one box per statement. Feel free to add additional features, and rate them accordingly)

Below are listed hypothetical properties of a prompting system, rate each in terms of how useful you perceive they might be in a screening practice:

	Essential	Useful	Doubtful	Of no use
Prompting for micro-calcification clusters	☐	☐	☐	☐
Prompting for ill defined lesions	☐	☐	☐	☐
Prompting for architectural distortions	☐	☐	☐	☐
Prompting for asymmetries	☐	☐	☐	☐
Classification of prompted micro-calcifications from benign to malignant	☐	☐	☐	☐
Classification of prompted masses from benign to malignant	☐	☐	☐	☐
Others				
...	☐	☐	☐	☐
...	☐	☐	☐	☐

(Please tick one box per statement. Feel free to add additional properties, and rate them accordingly)

Given that the capabilities of a prompting system are likely to evolve with time, prioritise following: (1 indicates the function should be developed first, 2 second etc. Use each number only once)

Prompting for micro-calcifications	☐
Prompting for ill defined lesions	☐
Prompting for architectural distortions	☐
Prompting for asymmetries	☐
Classification of prompted micro-calcifications as benign to malignant.	☐
Classification of prompted masses as benign or malignant	☐
Other	
...	☐
...	☐

(Feel free to add additional properties, and number them accordingly)

In a screening practice, what problems do you see a prompting system addressing? (Please rate the following in importance: 1 = most important, 2 = second in importance etc. Please use each number only once)

Reducing the number of interval cancers (false negatives)	☐
Improving the detection performance of a single reader.	☐
Improving the consistency of reading (eg compensating for fatigue)	☐
Supporting inexperienced radiologists?	☐
Addressing resourcing limitations (eg availability of radiologists)	☐
Reducing recalls (if classification available).	☐
Other	
...	☐
...	☐

(Feel free to add any additional roles, and number them accordingly)

How do you see a prompting system being used in your screening clinic:

Replacing double reading with single reading and a prompting system ☐

Using the prompting system to enhance double reading. ☐

Other ☐

(Please tick one box only)

If other, please specify:

Please rate your agreement with the following statements:

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
In cases where you are unsure, do you believe that					
The presence of a prompt will make you more inclined to recommend recall.	☐	☐	☐	☐	☐
The absence of a prompt makes you less likely to recommend recall	☐	☐	☐	☐	☐

Please give the following possible system configurations a rating on a scale of 1-5 as to how useful you believe each configuration to be (1 most useful, 5 least useful, tick one box only)

	Most useful 1	2	3	4	5 Least useful
High prompt rate, where most of the features prompted for are benign, but with a high probability that any malignancies will also be prompted for.	☐	☐	☐	☐	☐
Low prompt rate, where few of the prompts are for benign features, but with a high probability that some malignancies will be missed by the system.	☐	☐	☐	☐	☐
A system which is designed to prompt for micro-calcification clusters (whether malignant or benign) but not other types of calcification (eg vascular calcification, popcorn calcification).	☐	☐	☐	☐	☐
A system that will prompt for all types of calcification clusters, rather than one that tries to discard those with benign appearance.	☐	☐	☐	☐	☐
A system that will prompt for opacities that can usually be dismissed by radiologists with the aid of previous films or multiple views (eg composite shadows), as well opacities that are the result of a malignant process.	☐	☐	☐	☐	☐

At the outset of this experiment we gave you an estimate of of the sensitivity of the ill-defined lesion and microcalcification algorithms. Based on your experience of using the system, what would be your estimate of the sensitivity of these components be?

Microcalcifications	.....%
Ill-defined lesions	.....%
Overall (sensitivity for detecting all feature types)	.....%

Please rate your confidence in your assessment of the sensitivity of the system components given in the answer to the above question. (On a scale of 1 to 5, where 1=Most confident, 5=Least confident).

	Most confident				Least confident
	1	2	3	4	5
Microcalcifications	☐	☐	☐	☐	☐
Ill-defined lesions	☐	☐	☐	☐	☐
Overall (sensitivity for detecting all feature types)	☐	☐	☐	☐	☐

Do you believe your sensitivity in the prompted sessions has been better, the same, or worse, compared with your sensitivity in the unprompted sessions, for the following types of lesion:

	Better	Same	Worse
Microcalcifications	☐	☐	☐
Ill-defined lesions	☐	☐	☐
Overall (all lesion types)	☐	☐	☐

Do you believe your specificity in the prompted sessions has been better, the same, or worse, compared with your specificity in the unprompted sessions, for the following types of lesion:

	Better	Same	Worse
Microcalcifications	☐	☐	☐
Ill-defined lesions	☐	☐	☐
Overall (all lesion types)	☐	☐	☐

Thank you for completing this questionnaire

C.3.6 Post-session questionnaire

Prompted session questionnaire

Please answer these questions with respect to your opinion of prompting system taking into consideration all the prompted conditions that you have so far read.

Each of the following statements gives an opinion regarding the prompting system. Please state your agreement by ticking one box per statement.

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
The system will be time consuming to use.	☐	☐	☐	☐	☐
The information supplied by the system was distracting.	☐	☐	☐	☐	☐
The prompts were confusing.	☐	☐	☐	☐	☐
Many changes would be required before the system would be useful.	☐	☐	☐	☐	☐
The system is inaccurate.	☐	☐	☐	☐	☐
The system gives useful information.	☐	☐	☐	☐	☐
Prompts were no better than random.	☐	☐	☐	☐	☐
The system will be of no use to me as an aid for reporting.	☐	☐	☐	☐	☐
The system makes me more confident that I will find any cancers.	☐	☐	☐	☐	☐
It was clear what features the prompts referred to.	☐	☐	☐	☐	☐
Too many false positive prompts were produced.	☐	☐	☐	☐	☐

1

	Strongly Agree	Agree	Uncertain	Disagree	Strongly disagree
I would be keen to use the system.	☐	☐	☐	☐	☐
I would recommend the system to my colleagues.	☐	☐	☐	☐	☐
I would be happy using the system as it currently operates.	☐	☐	☐	☐	☐
The system performed better than I had expected.	☐	☐	☐	☐	☐
It is obvious what the system was prompting for.	☐	☐	☐	☐	☐
Using the system was satisfying.	☐	☐	☐	☐	☐
The effort needed to use the system was not justified by the benefits of using the system.	☐	☐	☐	☐	☐

Overall rating:

What score would you give the system to indicate its overall usefulness in a screening context? (Rate from 0-100, with 100 being the best possible score) ☐

Do you believe that: (Tick one box per question)

	Yes	No
Overall, the system would be useful to you in a screening context as it currently stands?	☐	☐
The mass detection component of the system would be useful to you as it currently stands?	☐	☐
The microcalcification detection component of the system would be useful to you as it currently stands?	☐	☐

Please rate the system's sensitivity: (Tick one box for each)

	Too sensitive	Just right	Not sensitive enough
Overall	☐	☐	☐
Masses	☐	☐	☐
Microcalcification	☐	☐	☐

Please rate the system's specificity: (Tick one box for each)

	Too specific	Just right	Not specific enough
Overall	☐	☐	☐
Masses	☐	☐	☐
Microcalcification	☐	☐	☐

This question concerns how easy it is to interpret the prompting information. Roughly, for what percentage of prompts have you had difficulty in being able to:

	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
locate the prompted region on the mammogram?	☐	☐	☐	☐	☐
understand why the system has prompted for a particular area?	☐	☐	☐	☐	☐

If there are any instances or categories of prompts that you have found particularly difficult to interpret then please give details below:

General Impressions

What do you think the system's strengths are?

What do you think the system's weaknesses are?

What irritated you most about the system?

What aspects of the system did you find most useful?

Can you you suggest how the system might be improved?

Thank you for completing this questionnaire

5

Appendix D

Publications arising from this thesis

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

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

3. Hartswood, M., Procter, R., Williams, L. (1998) Prompting in mammography: Computer-aided Detection or Computer-aided Diagnosis? Proceedings of Medical Image Analysis and Understanding. Leeds, July.

4. Hartswood, M., Procter, R. and Williams, L. (1998) Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography? To be published in Karssemeijer, N. (Ed.) Proceedings of the Fourth International Workshop on Digital Mammography. Nijmegen, June.

5. Hartswood, M., Procter, R. (2000) Computer-aided mammography: a case study of coping with fallibility in a skilled decision-making task. To be published in: Special Edition of Topics in Health Information Management on Human Error in Clinical Systems. 20:4

Subjective Responses to Prompting in Screening Mammography

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

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

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

1 Introduction

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

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

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

2 The Experiment

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

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

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

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

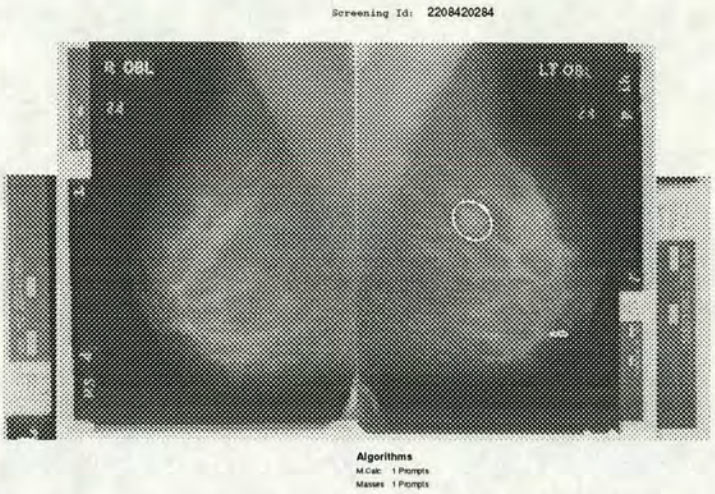


Fig. 1. Example prompt sheet

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

Table 1. Average prompt rates for prompted conditions

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

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

3 Results

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

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

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

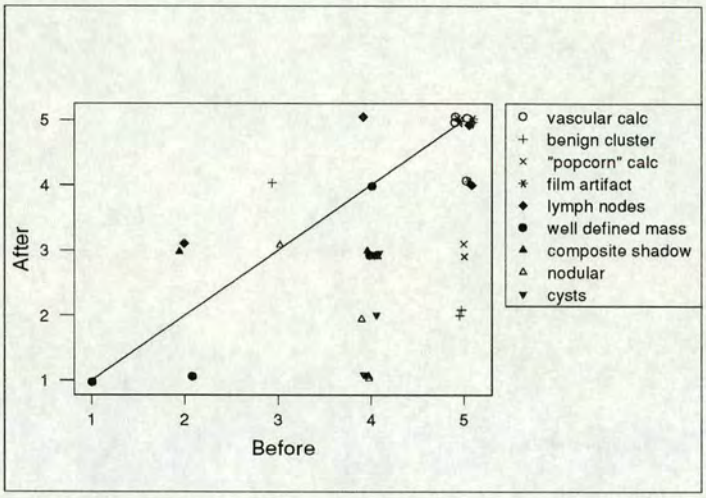


Fig. 2. Subjects' assessment of value of particular types of FP prompt (1 = useful and 5 = distracting) before and after the experiment.

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

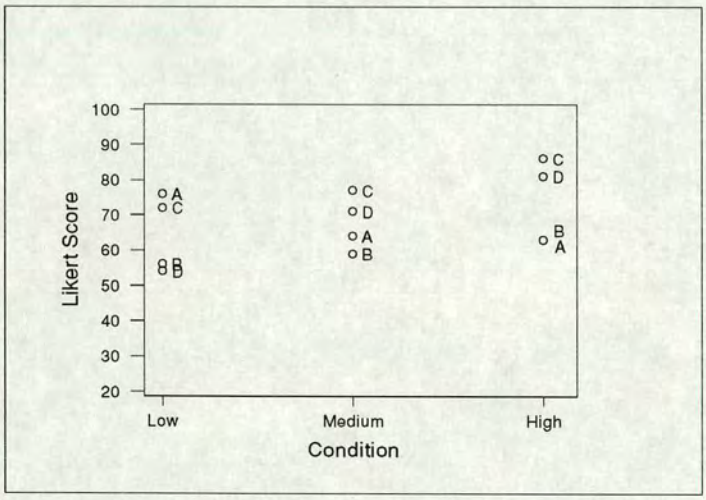


Fig. 3. Likert score against condition for subjects A to D.

4 Discussion

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

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

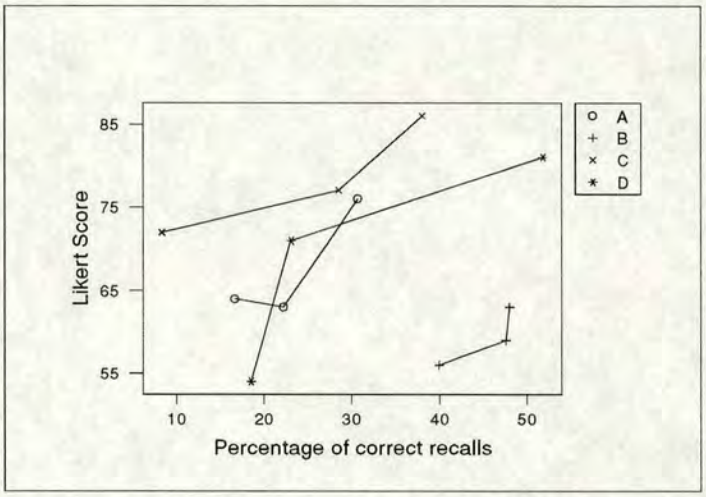


Fig. 4. Percentage of correctly prompted recalls against Likert score for each subject.

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

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

5 Conclusions and Further Work

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

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

References

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

rawing the line between perception and inter-etation in computer-aided mammography

Hartswood[a], R. Procter[a], L. Williams[b], R. Prescott[b], P. Dixon[a]

partment of Computer Science, Edinburgh University, Edinburgh, EH9 3JZ, Scotland.
partment of Public Health Sciences, Edinburgh University, Edinburgh, EH8 9AG,
land.

STRACT

eening mammography calls for a combination of perceptual skills to find what may be t and small features in a complex visual environment, and interpretive skills to rate r diagnostic significance. Evidence suggests radiologists' performance of this task can mproved by computer-aided prompting of target features.

The introduction of computer-aided mammography provides an interesting case study llocation of function' issues. One is where to 'draw the line' between perception and rpretation when determining the system's functional role. Our investigations indicate radiologists find a system which is 'perceptually acute', but 'interpretatively naive', e acceptable than predicted by earlier work. We present evidence that this is because wing the line in this way helps radiologists to understand, and to monitor, the system's aviour.

A second issue concerns the impact of computer-aided mammography on existing prac-s. Our studies reveal informal, but important, collaborative practices which help to ke radiologists' work routinely available to each other. We argue that such practices st be properly understood and accommodated within computer-aided mammography s benefits are to be fully realised.

Introduction

ast cancer is the commonest form of cancer in the UK. Each year there are about 00 new cases and 15,000 deaths from the disease, accounting for one-fifth of deaths ng women from all forms of cancer. Mammography (radiological imaging of the ast) remains the only method of detecting early stages of breast cancer, and prevent-e screening mammography programmes operate in many countries.

n the UK, women between the ages of 50 and 64 are invited to attend a clinic for ening every three years. In the UK, the rate of detection of abnormalities through ening is about 6% for women undergoing their first screening, falling to 3% for second subsequent screenings. Currently about 0.6% of those screened are found to have ignancies. The radiologists' task is a difficult one, not least because the small number ancers is hidden amongst a large number of normal cases. It is a task which demands gh level of perceptual and interpretative skill: under certain circumstances normal ue can have an abnormal appearance — and vice versa.

The goal of screening is to achieve a reliable and controlled cancer detection rate. Two formance parameters are particularly important: specificity and sensitivity. A high cificity (high true positive rate) means that few women will be recalled for further

s unnecessarily; a high sensitivity (low false negative rate) means that few cancers not be found. Achieving high specificity *and* high sensitivity is difficult.

The UK screening programme is continually investigating ways of improving detec- rates: current practice involves each mammogram being 'double read' (examined ependently by two radiologists) which has been shown to improve true positive rates npared with single reading (examination by one radiologist only). In the past five rs, interest has grown in the possibility of developing computer-based image analysis ls which will enable a single radiologist to achieve performance equal to that achieved double reading.

Computer-aided mammography (CAM) raises some important questions regarding the cation of function between human and computer-based agents. We begin by review- allocation of function issues within the general medical application context, and then efly outline the UK breast screening programme and the nature of reading work. We n present evidence from our investigations, and finally we discuss its implications for nputer-aided mammography.

Allocation of Function Issues in Medical Work

e early promise that expert systems would master the intellectual aspects of med- practice (Schwartz, 1970) remain largely unfulfilled. Of the many medical decision port systems (MDSSs) implemented, few have found routine use (Forsythe, 1992a; thfield and Wyatt, 1993). Explanations for the failure of MDSSs fall broadly into ee categories.

1. Expert system technologies have not met performance expectations (Sutton, 1989): MDSS developers have been unable to deliver systems that meet promised operational specifications.

2. Design and development methodologies have been inadequate (Forsythe, 1992b): MDSS developers have misunderstood how human and MDSS performance may be best combined.

3. There have been broader methodological failings (Kaplan, 1982): MDSS developers have been unable to grasp that the culture and values of practitioners may be such that they will be resistant to using MDSSs.

These problem categories can be equated with three specific allocation of function is- scope, role and work practice.

pe

e technical difficulties associated with meeting operational specifications are typically re severe for MDSSs that target general application domains. This is because the know- ge base for general domains is often less well defined: knowledge from many sources y be integrated under a variety of different reasoning strategies to reach a decision. nore specific application domains, the knowledge base is often better formalised, and reasoning process limited to a few well-defined strategies, thus both knowledge and soning become more amenable to computer representation (Blois, 1980). There has n a move away from systems that try to duplicate the general diagnostic capability of nysician towards systems that focus on more specific problem domains (Miller, 1994).

e

ıe MDSSs support decision-making by simply providing information that can assist sicians to reach their own conclusion, e.g., performing a literature search. At the other of the scale there are MDSSs which offer their own interpretation of the facts, i.e., ɔmated diagnosis. In general, the latter are more difficult to design, more difficult to loy in a working environment, and often are difficult to use.

rk practices

rk practice issues in MDSS applications are inevitably multi-faceted, and problematic designers. An issue of particular importance is control. For example, the physician have the power both to decide when to use the MDSS, and to decide how to act on advice. On the other hand, MDSS use may be compulsory. In general, the latter tends ɔe resisted by physicians (Kaplan, 1988), whereas MDSSs that give useful reminders lerts have been well received (Clayton and Hripcsak, 1995).

Allocation of function issues in computer-aided mammography

liologists' expertise in reading mammograms is a combination of the perceptual skills ded to find what may be very faint and small features in a complex visual environ-ıt, with the interpretative skills required to rate their diagnostic significance (Tabar Dean, 1985). False negatives can be attributed to a number of factors:

. incomplete visual search, e.g. because of fatigue, attention diversion,

. missing of features e.g. because they are very faint, and

. mis-classification of features e.g. deciding that a feature is benign when it is actually malignant.

The first two of these represent errors of perception as the feature is never actually ı. The third represents an error of interpretation.

We are involved in a project to develop a CAM system to analyse mammograms for ıs of features known to be associated with the early stages of breast cancer. For each ure found, the system generates a prompt on a paper copy of the mammogram (see ure 2). The approach is based upon experimental evidence that shows prompting can rove radiologists' performance by reducing errors of perception (Hutt, 1996).

A CAM system poses a challenge with respect to each of the MDSS allocation of ction issues outlined earlier. In the case of scope, problems may occur for two reasons. st, current image analysis techniques are not able to find all the various types of nmographic feature in which radiologists are interested. Second, some kinds of feature be hard to distinguish from one another, and features may also overlap, with the result radiologists may misattribute a prompt to a feature which the system is not actually able of detecting. Together, these two factors raise the possibility that radiologists fail to understand the precise limits of the system's feature detection scope.

We have attempted to address some aspects of the control issue by allowing for dis-ionary and flexible use to be made of the prompting information: the radiologist will free to determine when to consult the prompts and may choose to ignore them. It vident, however, that changing from double reading to computer-aided single reading ld present significant problems, and should not be attempted without a much better erstanding of current clinic practices.

The issue of role concerns the question of where to draw the line between perception and erpretation when determining the functional role of the CAM system. The project's l is to increase radiologists' sensitivity by reducing the number of false negatives at- utable to errors of perception. The system is not intended to address the issue of false atives attributable to errors of interpretation. In principle, the system's functional may therefore be defined as perceptual, and not interpretative. However, in practice, question of where to draw the line in CAM between perception and interpretation is blematic.

Drawing the line so as to limit the system's interpretative function has the virtue of ieving a complementary synthesis of system and radiologists' strengths: the former is e consistent in its visual search performance and the latter has interpretative skills ch the system cannot match (Claridge, 1997). However, given the nature of the mam- gram image, drawing the line in this way may lead to many 'low value' false positive mpts, i.e., prompts for features that radiologists can see are obviously benign. The ger is that radiologists may find such prompts distracting and ignore them, including ie true positive prompts. In contrast, drawing the line so as to increase the system's erpretative function, and so reduce false positive prompts, is likely to cause its false ative prompt rate to increase.

n practice, some interpretative function (even if relatively simplistic) is essential in a M system. As with the human observer, perception and interpretation are operation- closely linked. For example, a CAM system which was unable to distinguish between ndom distribution of microcalcifications and microcalcification clusters would be use- . The problem is to find the correct balance between perception and interpretation: little interpretation and the system will fail in its objective of reducing false negative es; too much and it could conceivably cause them to increase.

To explore issues of scope, role and control further, we carried out a programme of estigation of screening practices at a number of clinics in the UK. This included ex- iments, semi-formal interviews with radiologists and radiographers, and ethnographic- ed observation of work practices.

Breast screening in the UK

e UK Breast Screening Program (UKBSP) is a national service with a regional organ- ion. Each region is served by a number of screening clinics, each with two or more iologists. The initial screening test is by mammography, where one or more X-ray s (mammograms) are taken of each breast by a radiographer. Each mammogram is mined for evidence of abnormality by two experienced radiologists. Types of feature t are indicators of malignancy include:

crocalcifications are small deposits of calcium visible on a mammogram as tiny bright specks. They can be due to benign processes: for example, 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.

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: for example, they

may be cystic. Lesions that do not have a well-defined edge are considered suspicious.

llate 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.

chitectural 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.

ymmetry 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.

When reading, radiologists may consult information provided by the radiographer that ld have a bearing on mammogram interpretation: for example, the location of scars, ether the woman is taking HRT, etc. In this way, information from several sources is nbined in the reading process. However, screening largely relies on radiologists' pertual and interpretative skills. Radiologists are highly trained and their work practices e evolved to reduce the likelihood of mistakes, especially false negatives.

ding practices

uble reading involves each mammogram being examined independently by two raogists. Various studies have indicated that double reading may give a 5% to 15% provement in cancer detection (Anderson et al., 1994; Warren and Duffy, 1995). There variations in the way that double reading is implemented. The most simple method o recall on a 'worst opinion recalls' basis, i.e., if either, or both, radiologists decide to all. Alternative methods include calling in a third radiologist to make the final decision en radiologists disagree.

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

However, the reading process is more complex than it appears at first sight. Our estigations indicate that categorisation of feature types is less clearly delineated than taxonomy described above suggests, particularly for ill-defined lesions. For example, appearance of some features may be ambiguous. Any linear structures associated with sion might be interpreted as evidence for spiculation. Such structures are examined sely. If they are perceived to pass through, rather than originate within the feature, n grounds for suspicion are diminished.

Radiologists may alter their recall threshold according to the type of tissue present a given mammogram. A feature in a mammogram that has a lot of asymmetrically ributed ('patchy') tissue might be treated with less suspicion than a similar feature pearing in a mammogram that has more evenly distributed tissue.

nitoring and articulation of work

aspects of screening work are closely monitored to reduce mistakes, particularly false atives. Clinic staff monitor their own, and each others' performance through formal cedures for quality assurance and work documentation. Clinic staff hold regular meet- and these may take several forms, for example:

- multi-disciplinary pathology meetings where radiological appearance and pathology data are compared;
- review of interval cancers, i.e., cancers appearing during the three year period between screening rounds, and which may be evidence of false negatives, and
- informal (and at some clinics, formal) discussion about differences in recall opinions.

Such meetings provide an opportunity for radiologists to articulate — i.e., make pub- — aspects of their work which they perform as individuals, such as their reasons for ng a 'recall' or 'no recall' opinion. This emphasises the fact that despite its appar- individualised character, reading work is performed within a specific "community of ctice" (Jordan, 1996). Review meetings, for example, serve to establish, reinforce and ew where radiologists should be setting the recall threshold. It is important, for ex- ple, that differences between radiologists' recalls are maintained within a manageable ge: the 'virtuous' difference which accounts for the improved detection rates observed louble reading. If the difference is too large, however, clinic specificity targets may be oardised, and changes in procedure may follow, like changing from a 'worst case' to a rd reader arbitration' recall decision-making policy.

n many workplaces, a more informal kind of work articulation is achieved through the lic character of documents (Hughes et al., 1996). In the screening clinic, the reporting n provides a particularly noteworthy example of this. Its design, together with the work ctices in which it is embedded, mean that second readers see the first reader's opinion en they record their own. This provides second readers with the opportunity to compare r performance with that of their colleagues, *within* the context of their own reading k. We found no evidence that the availability of the first reader's opinion directly uences the second reader's opinion: on the contrary, we believe that radiologists do ch their decisions independently. Instead, we suggest that it serves to maintain a more eral awareness of each others work.

We observed that in some clinics this informal articulation of work has evolved further: readers sometimes annotate the reporting form. In a significant number of instances found that these annotations related to features that the first reader had interpreted e in category C2 (benign), i.e., cases which the first reader had decided not to recall. ure 1 (1) shows one example of such an annotation. The first reader has marked on breast schematic printed on the reporting form the site of a feature with an "X" written "NRC" (no real change) beside it. In Figure 1 (2), the first radiologist has rked the site of a feature with "?" and written "BT" (breast tissue). Discussions with iologists revealed that these annotations serve several purposes. First, in the event he second reader deciding to recall, the first reader's annotations will provide useful rmation should the case go to third reader arbitration. Of particular interest, however, that the radiologists emphasised how this practice of annotating benign features plays ss overt, but important role of keeping each other informed about their work. One iologist remarked:

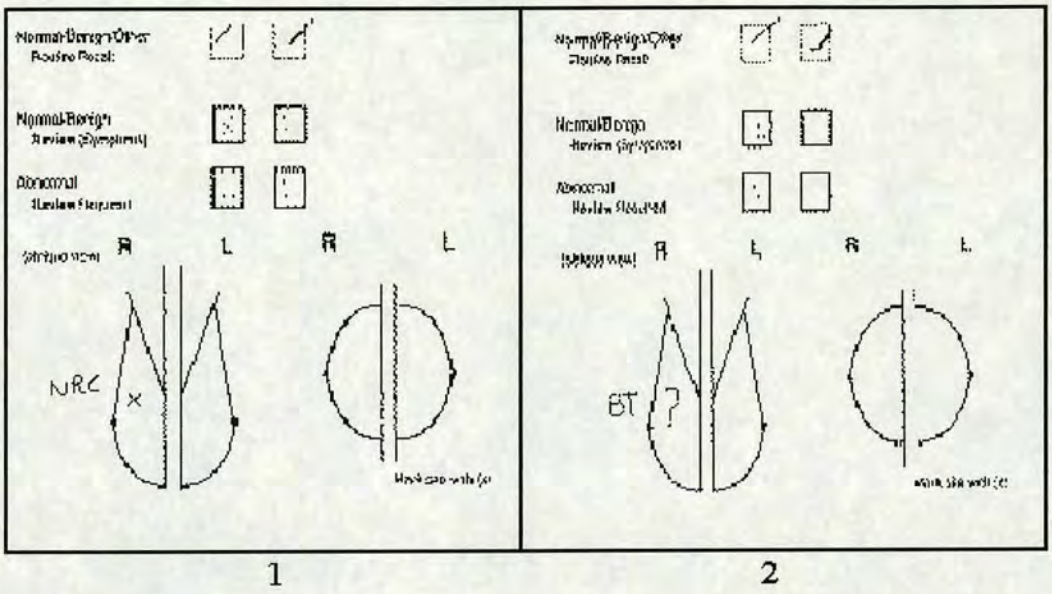


Figure 1: Examples of first readers' benign feature annotations.

"It's good to know that someone else is seeing the same thing (...) for example, that something hasn't changed (...) the second reader gets confirmation that they are thinking along the same lines."

The annotations in Figure 1 show the first reader making available to the second the soning behind her 'no recall' opinion. The first example (1) suggests little doubt in the reader's mind that her opinion is correct: the annotation seems intended merely to force it. In contrast, the second example (2) seems, through the use of "?" ("I think also appears quite frequently in this category of annotation), to express — and to w attention to — the first reader's uncertainty.

The fact that these informal work articulation practices should focus on features that on the benign side of the recall threshold may seem surprising. However, the region und the recall threshold is where most false positive and false negative decisions are ly to occur, and where the impact of differences in radiologists' opinions will be most nificant. In choosing to document this aspect of their work, radiologists display an ntation to the collective monitoring and management of their recall decision-making community of practice.

Previous investigations of prompting

perimental evidence suggests that prompting can improve radiologists' performance by ecting their attention towards suspicious features, but it was also found that if the false itive (FP) prompt rate is more than 1.5 times the True Positive (TP) prompt rate, n prompting ceased to be effective (Hutt, 1996). Since, in screening mammography, underlying cancer rate is approximately 0.5%, then for a 90% sensitivity target, we conclude that a prompting system may only be allowed 0.68 FP prompts per 100 es. This represents a combination of specificity and sensitivity which is far superior hat achieved by any existing image analysis techniques, and, indeed, to that of any iologist.

Though this conclusion is pessimistic of the value of current CAM techniques, it is

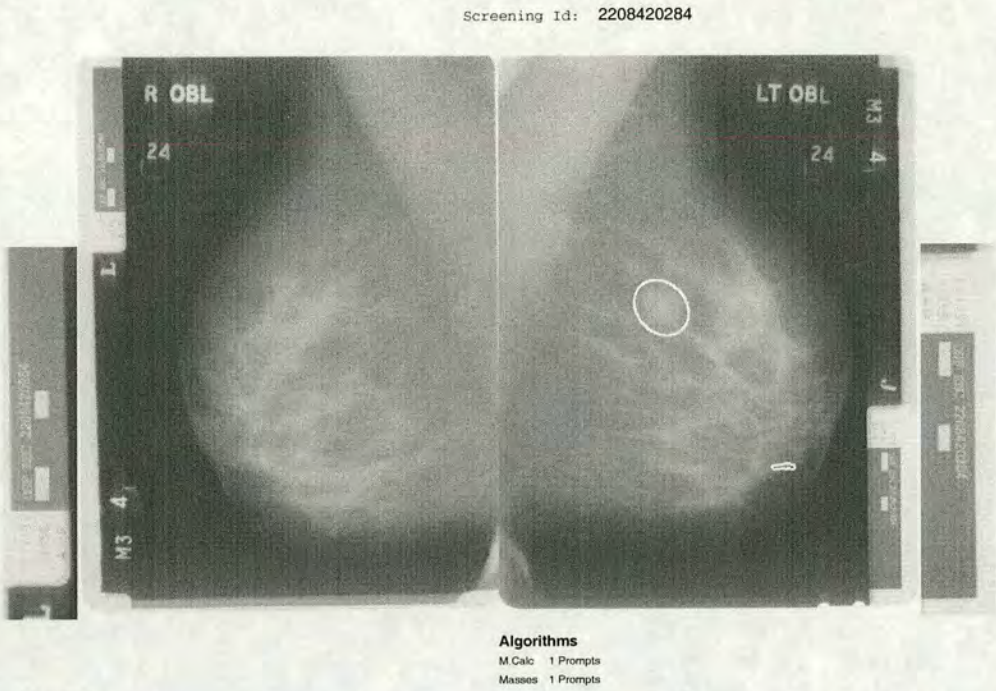


Figure 2: Example prompt sheet.

n to question. The studies employed heavily biased test sets in order to obtain a istically significant measure of sensitivity improvement, and so the results may not directly applicable to the circumstances in which reading is performed in the clinic. investigate this further, we decided to explore how radiologists assessed the value of mpts under conditions more typical of reading in the clinic.

Investigating radiologists' assessment of prompting

a series of experimental sessions, realistic reading conditions were simulated, including of standard reporting forms and attaching reporting forms and prompt sheets to a bag (Hartswood et al., 1997). Outputs from two of the CAM system's feature detion algorithms were used to generate prompts for microcalcification clusters (Hume al., 1996) and ill-defined lesions (Miller and Ramsay, 1996). Representative film sets e selected at random from four average days' screening at one clinic and balanced with pect to number of recalled cases, density of breast tissue and nodularity. There were pathology-proven malignancies in the set.

Prompt sheets consisted of a hard-copy, low resolution image of the mammogram pair h prompt information superimposed (Procter et al., 1994). Prompts for ill-defined ons consisted of an ellipse surrounding the suspect region, and for microcalcifications, irregular outline of the potential cluster (see Figure 2). Prompt sheets were attached reporting forms via a paper clip in such a way that a subject would have to lift the orting form to examine the prompt sheet. A prompt sheet was produced for each case spective of whether that case was actually prompted or not. Before the experiment, jects were given an overview of how the CAM system worked, including the types of ture it was capable of detecting.

The experiment consisted of four conditions. In three, subjects were prompted at dif- ent rates (High, Medium or Low; see Table 1) and one condition was an unprompted

Table 1: Average number of prompted cases in the prompted conditions.

Sensitivity Condition	Ill-defined lesions		Microcalcifications	
	Prompt rate	Sensitivity	Prompt rate	Sensitivity
High	55.7	62 %	35.5	94%
Medium	28.25	37 %	18.75	86%
Low	14	22 %	9.25	76%

trol. The data recorded included recall rate and time taken to read each condition. jects were recorded on video, and their actions subsequently transcribed. Question- es were administered before and after the experiment, and after each condition. Sub- s' attitudes to the system were assessed after each condition using a 20 point Likert , with the higher the total score, the more favourable the assessment.

ults and discussion

jects were asked to rate each prompt on a scale of one (useful) to five (distracting). -test of the results showed that subjects were significantly more likely to believe that mpting for benign features would be less distracting after the experiment than they e before it ($p<0.05$) (Hartswood, et al., 1997).

For the cases they recalled, subjects were asked to indicate whether the relevant feature been correctly prompted. For the majority of subjects, a monotonically increasing ert score was apparent as the number of prompts they judged to be correct increased. s suggests that subjects were more favourably disposed towards the system when the mpts corresponded with their expectations: i.e., when the 'opinion' of the system and t of the subject broadly coincided.

The protocol for the experiment instructed subjects to examine the films, examine prompt sheets, and then record their opinion. The video transcripts revealed that jects sometimes failed to follow instructions correctly. Table 2 shows the number of asions when the subjects either failed completely to examine the prompt sheet (Type ror), and when they recorded their opinion before examining the prompt sheet (Type ror). In the latter case, subjects may have turned the reporting form over after ording their opinion, and then gone back to it realising that they had forgotten to mine the prompt sheet. Taking radiologists' differences into account, there remained a istically significant variation in the frequency of errors between conditions ($p < 0.0001$ $p < 0.0111$), with a marked trend for subjects to make an error at the Low, rather n at the High, prompt rate. These results suggest that at lower prompting rates re was insufficient information to hold the subjects' attention, either because of the uency, or quality, (or both) of the prompts.

n eliciting post-condition comments, we sought to explore how use of the CAM system tributed to subjects' understanding of its behaviour. The results were mixed: for mple, since there were so few pathology-proven cancers in the test set, we had expected t subjects would not be able to assess the system's sensitivity accurately. In fact, their nimous opinion that the sensitivity of the system for ill-defined lesions (62% maximum) too low showed their grasp of this aspect of system behaviour was good. In contrast, eral subjects expressed the belief that the system was detecting asymmetries, even ugh it could not.

Overall, the results of this experiment indicated that under more realistic conditions,

ole 2: Number of occasions subjects did not examine prompt *at all* (Type 1 error), or y examined prompt sheet *after* making a decision (Type 2 error).

	Number of errors	
Prompt rate	Type 1	Type 2
Low	10	26
Medium	2	18
High	1	8

Table 3: Comparison of radiologists' recall opinions.

		Recall by Radiologist A	
		No	Yes
Recall by Radiologist B	No	109	9
	Yes	25	12
		Recall by Radiologist C	
		No	Yes
Recall by Radiologist B	No	107	11
	Yes	14	23
		Recall by Radiologist A	
		No	Yes
Recall by Radiologist C	No	113	8
	Yes	21	13

iologists' tolerance level for FP prompts was appreciably higher than the upper limit viously established for improved radiologist performance. Of course, positive assess- nt by radiologists may not necessarily coincide with improvement in performance, but se results raised the possibility that, perhaps because of its artificiality, earlier work underestimated the FP prompt upper limit. One explanation is that subjects' toler- d FP prompts in the new experiment because they provided useful information about CAM system's behaviour. To test this, a follow-up study was devised to examine in re detail how radiologists might use prompts to construct and confirm a model of the tem's behaviour.

Classification of prompts

ree experienced radiologists were asked to examine the prompts produced at the highest sitivity in the earlier experiment, and to decide whether they would recommend a recall the basis of the prompted feature. They were also asked to asked to classify each of the tures prompted according to their own confidence scale: C1 (normal), C2 (benign), C3 uivocal), C4 (suspicious) and C5 (malignant), and whether they thought prompting these features would be acceptable in routine screening. In addition, radiologists were ouraged to vocalise their thoughts, and these were recorded and transcribed.

luating the system

ole 3 compares how different pairs of radiologists classified the same set of prompted tures as a 'recall' or 'no recall'. The interesting cases are those where the radiologists

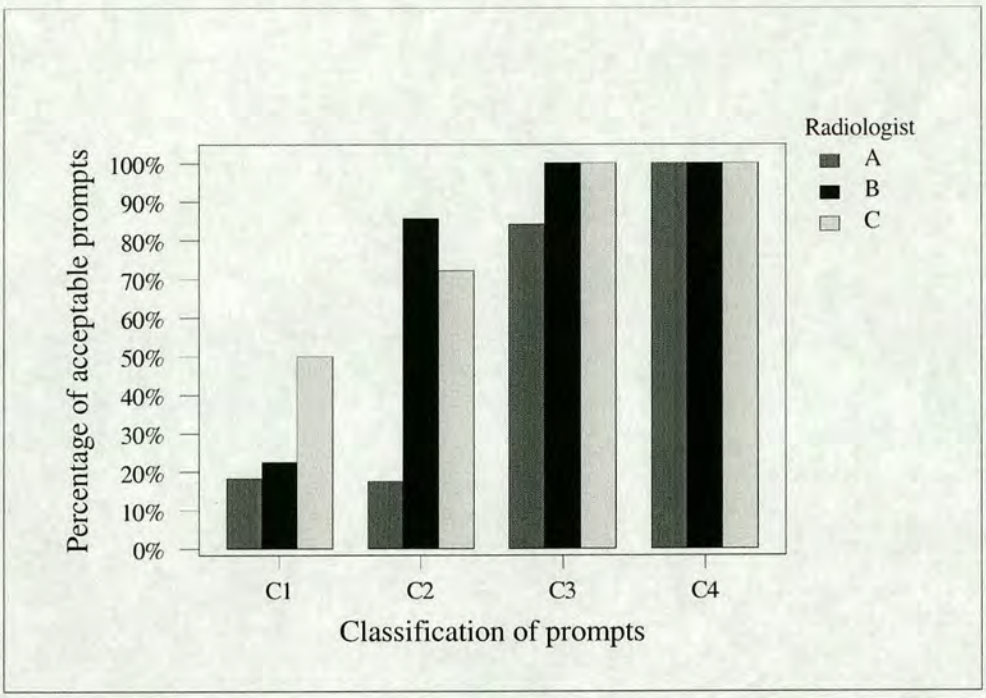


Figure 3: Percentage of acceptable prompts by prompt classification.

greed (the highlighted cells). As noted earlier, radiologists do not always agree on ch cases to recall. It is not surprising therefore that radiologists' classification of mpted features also shows some marked differences of opinion.

Figure 3 shows radiologists' ratings of prompt acceptability broken down by prompt sification (C1:C4). It is clear from these results that the boundary between 'not ac- table' and 'acceptable' lies within the C2 category, i.e., prompts for benign features. s leads us to suggest that the effect of FP prompts will depend on their classifica- . When reading unaided, radiologists perceive and interpret *candidate* features which ude members of the C2 (probably benign) category: i.e., features that have some perties in common with those they interpret as suspicious (i.e., C3:C5). We argue t prompts for *candidate* features may be acceptable to radiologists in the clinical set- , whereas prompts for *other* features (i.e., C1) would be distracting. This may explain results seen in earlier work. We conclude therefore that the appropriate place to w the line between perception and interpretation is so that the system can distinguish ween C1 and C2 features.

What is also interesting about these results is the parallel between radiologists' appar- tolerance of C2 prompts and their ad-hoc practice of annotating C2 features. There always be cases where the absence of a prompt may give ambiguous evidence of the em's performance. It could mean that the system found no feature (a possible 'error' erception), or that it found a feature and then determined it to be benign (a possible or' of interpretation). We suggest that radiologists may find this ambiguity a source confusion when attempting to understand where the system draws the line between ception and interpretation. In this critical region of performance, they prefer to have less unambiguous evidence of a prompt because of its capacity to document the CAM em's behaviour.

We argue that radiologists find some of the CAM system's FP prompts useful in much same way as they find each others annotations of benign features useful. To reiterate, iologists annotate these features as a way of documenting a particularly critical region heir reading performance. It is not so surprising therefore that radiologists should show an interest in this same region of the system's performance. We conclude that mpts for *candidate* features afford learning about — and confirmation of — the sys-'s behaviour.

king sense of the system

e following extracts of session verbal protocols illustrate how radiologists tried to make se of the CAM system's behaviour from the evidence of the features it had prompted. a number of instances, the transcripts show examples of misunderstanding of the ex- of the system's capabilities, and confusion because of apparent inconsistencies in its aviour.

n this first set of extracts, the comments suggest radiologists were unable to accurately ce the CAM system's operational scope: i.e., the types of feature it is capable of ecting:

> "Now what's been prompted for is the vascular calcification and this kind of asymmetry on the right."

> "I think that it's interesting that they've not prompted for this area of asymmetry, as I was saying earlier on there are certain review areas, the so-called milky way areas that Tabar teaches you of (...) and there is marked asymmetry there which has not been picked up there so I'll call that 1 (...) and 1, I think, should have been prompted."

n the first extract, the radiologist interpreted as an asymmetry a feature which the tem prompted as an ill-defined lesion. In fact, the system does not prompt asymmet-, but it was evident from the transcript as a whole that radiologists explained the aviour of the system by assuming that it was capable of detecting asymmetry. This to expectations that were difficult to fulfil. In the second extract, the radiologist ressed disappointment with the system precisely because it had failed to prompt an a of asymmetry.

The next set of extracts focuses on radiologists' problems with understanding how the tem interprets microcalcification clusters:

> "It's interesting that there's some clusters of calcification elsewhere that it has not picked up."

> "It's interesting it's prompted the vascular calcification on the one side and not on the other. So that gives me (...) I'm thinking the whole thing's inconsistent you know."

> "Again, extensive vascular calcification (...) There's actually some calcification associated with the breast parenchyma which I think is more obvious on the left side as probably benign lobular. Now let's go to the prompts. First thing I'm looking at when I look at that is what did they think about the lobular calcs or things I think are lobular it's not prompted. So I'm a bit disappointed."

n the first extract, the radiologist did not interpret the particles of microcalcification forming a series of discrete clusters: her interpretation was that there was simply a espread random distribution. The system prompts a region of microcalcifications if dentifies five or more particles in the neighbouring area. In this instance, by chance, ne of the randomly distributed particles met the system's criteria, and so a prompt produced. The system only examines the image locally to determine if the cluster erion is met. In contrast, the radiologist is able to make a global appraisal, and can over larger scale trends that are not apparent to the CAM system. In this case, the iologist concluded there were no microcalcification clusters, and was perplexed as to one part of the "random distribution" should be prompted over any other.

n the second extract, confusion arose because the radiologist automatically classified calcification present as being vascular, then posed the question: "why some vascular ification and not others?" Again, the system has a much simplified interpretation: t of the vascular calcification had fragmented into number of particles which were iciently close together for the system to interpret them as a cluster. The remainder hese vascular calcifications maintained their characteristic tram line appearance, and e not prompted.

The third extract is particularly interesting. Once more, the radiologist was perplexed ause the system's interpretation of a cluster was less sophisticated than her own. ially, the radiologist decided that there was a single cluster of lobular calcification, several clusters of vascular calcification present. The radiologist was more interested he former than the latter, and so was disappointed when only the vascular clusters were mpted. The lobular calcifications were very subtle, and so would have needed to form ght cluster in order to be prompted. On the other hand, some vascular calcifications lified as clusters according to the system's interpretation. The radiologist made a litative distinction between the vascular and lobular clusters, but the system has no n interpretative capacity, and so fell short of the radiologist's expectations.

n the final set of extracts, the radiologists indicated that they had not see anything ignificance in the areas prompted:

> "What's been prompted is (presumably)? a cluster of calcifications posterially (...) I'm struggling to see it (...) I think there might be a vessel in that area (...) I think that probably has been quite distracting. I wouldn't expect that to be prompted and I wouldn't recall. I think it's probably vascular calcification (...) there (...) a tiny cluster (...) if it's present at all."

> "There's some calcification on the right which I think is probably benign, and in fact she's got a cluster on the left as well. So we've picked that up (...) and we've picked up a third cluster which I obviously haven't (...) what's that? (...) struggling (...) I don't see it."

On closer examination of these cases, we found that there was a small number of very tle calcifications present. Radiologists do make a point of looking for subtle clusters, ever, very subtle clusters occur relatively frequently, are mostly benign, and present fficient information in terms of size, shape and distribution for a radiologist to identify ignant ones. Furthermore, if there is disease present, at this stage it is likely to develop tively slowly, and so there is a reduced risk in waiting until the subsequent screening nd when there might be more evidence. In these examples the system is too percep- ly acute, producing prompts for features that are difficult for radiologists to locate,

l that also have little diagnostic relevance.

Conclusions and future work

nputer-aided mammography raises allocation of function issues with regard to scope, e and work practice. Taking the issue of role first, our investigations indicate that a M system should have sufficient interpretative capability to distinguish between *can-ate* and *other* features. Our evidence suggests that from the radiologists' point of view, s would mean drawing the line between perception and interpretation at the C1 (nor-l) : C2 (benign) boundary. We acknowledge that these are subjective effects, and that far we have no evidence that radiologists' actual reading performance will be improved. ge scale trials are being planned to obtain statistically reliable measures of the latter. Our investigations also show that radiologists may have problems in understanding operational scope of CAM systems, particularly at their boundaries. This points to importance not only of determining where to draw the line between perception and erpretation, but also of radiologists *knowing* where it is. Our evidence suggests that mpts not only serve as a cue to examine particular features, but also as an aid to development of radiologists' understanding of how the system works, and what its abilities are.

Training sessions and manuals are useful resources for explanation, and we have used results of our investigation to inform the content of such materials. However, for tainable understanding, systems need to provide accounts of their behaviour which both relevant to, available, and understandable within the actual *doing* of the work y support. The problem is that CAM systems are complex, and that prompts only vide a very limited account of their behaviour. We are currently exploring ways in ch these accounts can be enriched. Our approach is informed not only by the way ividual radiologists' make decisions, but also by the ways in which they sustain their ader community of practice.

This brings us to the final issue of work practice. Through the public character of the orting form, double reading provides a means by which radiologists can make their k available to each other *as they do it*, i.e., where this information is most likely to relevant, and understandable. Double reading therefore contributes to the collective ntenance of clinics' screening performance in ways, which though they are informal sometimes ad-hoc, may be just as important as its nominal effect. Double reading evolved practices through which radiologists' work can simultaneously both be expli-y distributed *and* implicitly collective. The adoption of computer-aided single reading uld inevitably mean the disruption of these practices. We conclude, therefore, that the ective dimension of reading work must be better understood if the potential benefits computer-aided mammography are to be fully realised. Further investigation of this e is also planned.

ferences

derson, E. D. C., Muir, B. B., Walsh, J. S. and Kirkpatrick, A. E., 1994, The Efficacy of uble Reading Mammograms in Breast Screening, *Clinical Radiology*, **49**, pp. 248-251.

is, M. S., 1980, Clinical judgment and computers, *The New England Journal of Medi-e*, **303**, pp. 192-197.

ridge, E. 1997, Experts' assessment as a "gold standard" for characterisation of le-ns? In Proceedings of Medical Image Analysis and Understanding '97, Oxford.

yton, P. D., and Hripcsak, G., 1995, Decision support in healthcare. *International*

ırnal of Biomedical Computing, **39**, pp. 59-66.
sythe, D. E., 1992a, Blaming the user in medical informatics: The cultural nature of entific practice, *Knowledge and Society*, **9**(3), pp. 95-111.
sythe, D. E., 1992b, Using ethnography to build a working system: Rethinking basic ign assumptions. In Proceedings of the 16th Symposium on Computer Applications Medical Care, pp. 505-509.
athfield, H. A. and Wyatt, J., 1993, Philosophies for the Design and Development of nical Decision Support Systems, *Methods of Information in Medicine*, **32**(1), pp. 1-8.
rtswood, M., Procter, R., Williams, L., and Prescott, R., 1997, Subjective Reaction to ompting in Screening Mammography. In Proceedings of Medical Image Analysis and derstanding '97, Oxford.
ghes, J., King, V., Mariani, J., Rodden, T., and Twidale, M., 1996, Paperwork and its sons for Database Systems. In The Design of Computer Supported Cooperative Work l Groupware Systems by D. Shapiro, M. Trauber and R. Traunmuller (eds.) (North-land), pp. 43-62.
me, A., Thanisch, P., Hartswood, M., and Procter, R., 1996, On the evaluation of rocalcification detection algorithms. In Proceedings of the Third International Work-p on Digital Mammography, Chicago.
tt, I., 1996, The Computer-Aided Detection of Abnormalities in Digital Mammograms. ublished Ph.D. Thesis, Manchester University.
dan, B. Ethnographic Workplace Studies and CSCW. In The Design of Computer pported Cooperative Work and Groupware Systems by D. Shapiro, M. Trauber and R. unmuller (eds.) (North-Holland), pp. 17-42.
plan, B., 1982, The influence of medical values and practices on medical computer plications. In Proceedings of the First International Conference on Medical Computer ence/Computational Medicine, pp. 83-88.
plan, B., 1988, Development and acceptance of medical information systems: an his-ical overview, *Journal of Health and Human Resources Administration*, **1**, pp. 9-29.
ler, L., and Ramsay, N., 1996, The detection of malignant masses by non-linear ltiscale analysis. In Proceedings of the Third International Workshop on Digital Mam-graphy, Chicago.
ler, R. A., 1994, Medical Diagnosis and Decision Support Systems — Past, Present Future, *Journal of the American Medical Informatics Association*, **1**(1), pp. 8-27.
cter, R., Thanisch, P., Astley, S., and Hutt, I., 1994, User interface design and data nagement for digital mass mammography. In Proceedings of the Second International rkshop on Digital Mammography, York.
wartz, W. B., Medicine and the Computer, 1970, *New England Journal of Medicine*, **3**, pp. 1257-1264.
ton, G. C., 1989, Computer-aided diagnosis: a review, *British Journal of Surgery*, **76**, 82-85.
par, L. and Dean, P., 1985. Teaching Atlas of Mammography, Theime.
rren, R. M. L., and Duffy S. W., 1995, Comparison of single reading with double read-of mammograms, and changes in effectiveness with experience, *The British Journal Radiology*, **68**, pp. 958-962.

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

Mark Hartswood[a]*, Rob Procter[a] and Linda J. Williams[b]

[a]Department of Computer Science, Edinburgh University, Edinburgh, EH9 3JZ
[b]Department of Public Health Sciences, Edinburgh University, Edinburgh, EH8 9AG.

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

1 Introduction

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

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

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

2 Procedure

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

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

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

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

3 Results and Discussion

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

3.1 Aiding detection

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

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

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

> "Yeah, one, on micro-calcifications ... that I didn't see and then I brought back." (Subject E)

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

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

3.2 Aiding diagnosis

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

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

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

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

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

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

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

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

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

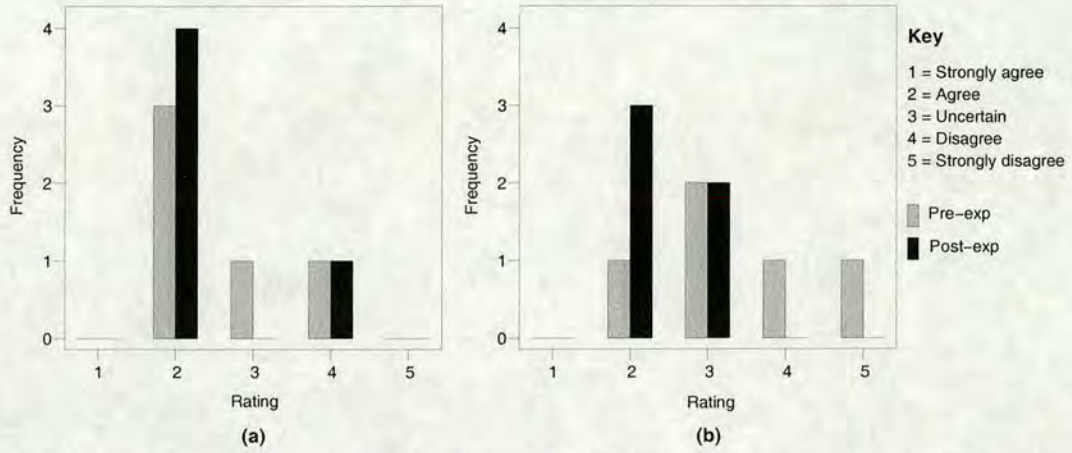


Figure 1. (a) the presence of a prompt will make me more inclined to recommend recall (b) the absence of a prompt will make me less likely to recommend recall.

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

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

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

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

Recalled by:	Correctly Prompted		Total
	Yes	No	
Prompted reader only (P^+U^-)	35	34	69
Unprompted reader only (P^-U^+)	31	65	96

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

Table 1 shows that for the case P^+U^- 50.7% of recalls were correctly prompted, whereas only 32.3% of recalls were correctly prompted for the case P^-U^+. A Chi-squared test indicates that this result would not be expected

if exposure to the system and the proportion of correctly prompted recalls were independent (p=0.017). Thus there is a greater level of agreement between subjects and PROMAM after the subjects were exposed to prompting information — implying that the prompts have had an influence on decision making. This influence could be due to the prompted condition leading to the detection of a greater number of significant features that would have otherwise been overlooked. However, it is also consistent with our earlier conclusion that radiologists' diagnostic decisions are being influenced by the presence or absence of prompts.

4 Summary and Conclusions

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

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

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

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

References

1. M. Hartswood, R. Procter, L. J. Williams et al. "Subjective reaction to prompting in screening mammography." In C. Taylor (editor), *Proceedings of Medical Image Analysis and Understanding*. Oxford, July 1997.
2. M. Hartswood, R. Procter, L. J. Williams et al. "Drawing the line between perception and interpretation in computer-aided mammography." In L. Bannon (editor), *Proceedings of the First International Conference on Allocation of Functions.*, pp. 275–291. Galway, IEA Press, October 1997.
3. H. Chan, K. Doi, C. J. Vyborny et al. "Improvement in radiologists' detection of clustered microcalcifications on mammograms." *Radiology* **25**, pp. 1102–1110, 1990.
4. M. Mugglestone, R. Lomax, A. G. Gale et al. "The effect of prompting mammographic abnormalities on the human observer." In *Proceedings of the Third International Workshop on Digital Mammography*. Chicago, June 1996.
5. L. Williams, R. Prescott & M. Hartswood. "Computer-aided cancer detection and the uk national breast screening programme." In N. Karssemeijer (editor), *Proceedings of the Fourth International Workshop on Digital Mammography*. Nijmegen, June 1998.
6. M. Hartswood, R. Procter & L. J. Williams. "Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography?" In N. Karssemeijer (editor), *Proceedings of the Fourth International Workshop on Digital Mammography*. Nijmegen, June 1998.
7. M. Giger. "Computer-aided diagnosis." In A. Haus & M. Yaffe (editors), *A categorical course in physics: Technical aspects of breast imaging*, pp. 283–298. RSNA, 1993.

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

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

1. Introduction

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

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

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

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

In a recent small scale clinical evaluation of PROMAM's performance we collected interview and questionnaire data to address these issues further [4]. The results suggest that radiologists use prompting information not only as attention cues, but also to inform their decision-making where there is uncertainty in

the interpretation of a lesion. Furthermore, we found that radiologists developed strategies to economise on the effort required to dismiss false positive prompts: (a) by anticipating where prompts were likely to appear, and (b) by making a judgement on the value of a prompt based on information in the prompt itself, rather than on the image content of the prompted region.

2. Methods

Five subjects were recruited from radiologists at a Scottish breast screening centre. Two thousand and two archive cases (including 102 pathology proven cancers) were digitised and analysed by the PROMAM system. The system performance was as follows: microcalcification sensitivity 93.8%, with 54% of cases falsely prompted; mass sensitivity 72.9%, with 66% of cases falsely prompted [6]. The films were then divided into twenty sets of approximately one hundred films and double read, once by a subject in a prompted condition and once by a subject unprompted. Constraints on subject availability meant that it was impossible to ensure that subjects read the same number of prompted as unprompted conditions. In the prompted conditions, subjects were asked to first examine the films, then examine the prompt sheet, and then to record their decision.

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

3. Training

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

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

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

4. Impact on decision-making

In each of the post-prompted session interviews, subjects were asked if the prompts had had some influence on their recall decision. Out of a total of sixteen interviews held after prompted sessions, subjects indicated that their recall decisions had been affected *one or more* times in a total of eleven of those sessions. Subjects reported a number of occasions where the prompts had drawn significant features to their attention which they had overlooked, sometimes resulting in a recall decision.

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

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

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

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

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

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

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

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

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

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

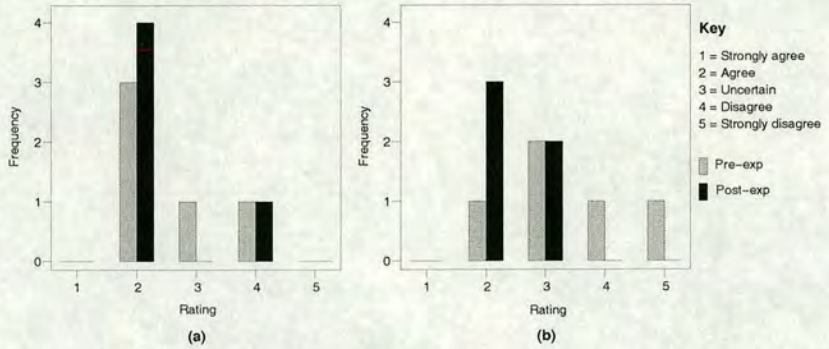


Figure 1. (a) the presence of a prompt will make me more inclined to recommend recall; (b) the absence of a prompt will make me less likely to recommend recall.

Data based upon self-reporting may be subject to various unconscious biases. By comparing unprompted and prompted recalls, it is possible to gain a more objective view of the influence of prompts on subjects' recalls. In the prompted conditions, subjects had been asked to record if a correct prompt was given for the significant feature in each case they recalled. This information was not available for cases recalled only by the unprompted reader, so a follow-up exercise was devised to determine which of these recalls had been correctly prompted.

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

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

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

Table 1 shows that 50.7% of recalls in the prompted condition were correctly prompted, system, where as only 32.3% of the unprompted recalls had correct prompts. A Chi-squared test indicates that this result would not be expected if exposure to the system and the proportion of correctly prompted recalls were independent (p=0.017). Thus there is a greater level of agreement between subjects and PROMAM when the subjects were exposed to prompting information, which implies that the prompts did have an influence on decision-making. This influence

could be due to the detection of a greater number of significant features that would have otherwise been overlooked, but it is also consistent with the interview data showing that prompts influence classification decisions.

5. Dismissing prompts

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

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

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

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

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

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

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

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

6. Anticipating prompts

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

> "I think that I'm beginning to get so that I can guess what's going to be prompted for." (Subject C)

> "I sometimes look at the films and say 'I bet it's going to prompt for that'..." (Subject B)

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

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

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

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

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

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

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

The success of anticipation is dependent upon consistency of the prompting system as *perceived* by the radiologist. Image analysis algorithms can be sensitive to variations in appearance which are too subtle for the radiologist to appreciate without close examination — if at all. Though system behaviour may be *strictly*

deterministic, it may not be *observably* deterministic if it doesn't respond in the same way to features that radiologists would classify as being similar.

7. Summary and conclusions

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

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

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

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

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

designed to encourage radiologists to recognise the circumstances in which this particular bias is likely to occur.

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

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

8. Acknowledgements

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

References

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

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

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

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

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

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

Computer-aided mammography: a case study of coping with fallibility in a skilled decision-making task

Mark Hartswood, M.Sc. and Rob Procter, Ph.D.
Institute for Communicating and Collaborating Systems
University of Edinburgh

Abstract/key words

Breast screening requires radiologists to exercise keen perceptual skills to find what may be faint and small features, and sophisticated interpretative skills to classify them correctly. Understandably, radiologists sometimes make errors, and evidence suggests that these can be reduced by employing a computer prompting aid. To investigate prompting aid requirements, we have studied both current reading practices and radiologists reading with prompts. These studies have enabled us to understand better how radiologists manage errors in current practice, and how they deal with prompting aid errors. They also show that such aids may get used in ways quite different from those originally envisaged.

Breast screening, computer-aided mammography, prompting, error management, ethnographic studies

Introduction

We have been working as members of a team which is developing PROMAM (PROmpting for MAMmography), a computer prompting aid intended for use in the UK breast screening programme.[1] PROMAM provides an interesting case study of error management issues raised by applying computer aids to a skilled decision-making task. Prompting aids are designed to improve observer performance by using image analysis techniques to cue areas that the observer should examine. In principle, errors arising from inattention, fatigue, etc. can be reduced. However, the practical realisation of this goal is not straightforward. Prompting aids are not infallible, so observers must be able to recognise prompting errors if their performance is not to be adversely affected by them. Further, the introduction of such aids may involve changes in work practices which actually reduce the effectiveness of existing error management procedures. In this paper we present the results of a series of investigations into these issues.

We begin with an overview of breast screening, followed by a summary of the role of prompting in improving observer performance. Next, we present a field study of breast screening work. The results show how, formally and informally, individually and collectively, radiologists seek to manage their performance so as to minimise errors. We then present studies of radiologists using PROMAM, focusing on how they deal with prompting errors. The results reveal the role that making sense of PROMAM's behaviour has in the management of its errors. They also suggest that prompting aids may actually be used in a way which may be quite different from that originally envisaged. We conclude with a discussion of the implications of our investigations for the successful adoption of prompting aids in breast screening.

Screening mammography

In breast screening the initial test is by mammography: one or more X-ray films (mammograms) are taken of each breast by a radiographer and examined for evidence of abnormality by at least one radiologist. Several types of mammogram feature are early indicators of breast cancer: micro-calcification clusters are small deposits of calcium visible

as tiny bright specks; ill-defined lesions are areas of radiographically-dense tissue appearing as a bright patch that might indicate a developing tumour; stellate lesions are visible as a radiating structure with ill-defined borders; architectural distortion may be visible when tissue around the site of a developing tumour contracts; asymmetry between left and right mammograms may be the only visible sign of some features.

Two performance parameters are particularly important in screening: specificity and sensitivity. A high specificity (low false positive (FP) rate) means that few women will be recalled for further tests unnecessarily; a high sensitivity (high true positive (TP) rate) means that few cancers are missed. Achieving high specificity with high sensitivity is difficult, not least because the small number of cancers is hidden amongst a large number of normal cases. Reading demands a high level of perceptual and interpretative skill: in some circumstances normal tissue can have an abnormal appearance, and vice versa.

Figures from the UK Breast Screening Program (UKBSP) show that in the prevalent (first screening) round 6.4% of women are recalled for further tests, falling to 3.0% in incident (later screening) rounds. More cancers are detected in the former (6.3 per thousand), than in the latter (3.4 per thousand).[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. In contrast, false negative (FN) errors, are life-threatening.

The UKBSP is continually investigating ways of reducing FP and FN errors. For example, current practice typically involves mammograms being 'double read' (i.e., examined independently by two radiologists). Interest has grown in the possibility of using computer prompting aids to enable a single reader (radiologist) to achieve performance equal to that of double reading.

Computer-aided mammography

The goal of computer-aided mammography is to reduce radiologists' errors. Kundel et al. have classified three types of error that can result in a FN decision for radiological search tasks.[3] 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 an empirically determined threshold; classification errors occur if visual dwell time exceeds it. Savage et al. define search and detection errors as failure to report a lesion's presence, and classification errors as a lesion reported, but inappropriately acted upon.[4]

Computer-aided detection aims to improve sensitivity by reducing search or detection errors. Promptimg aids such as PROMAM work by drawing attention to features that may otherwise be overlooked. It is not intended that radiologists should attach any clinical significance to the presence (or absence) of a prompt. In contrast, classification aids aim to improve specificity by providing an interpretation (either in terms of a probability value, or some explicit reasoning), and thus support a radiologist's judgement about the significance of a lesion already detected.[5]

Prompting aids are not 100% sensitive, but this is not important as long as the correlation between prompting aid and radiologists' FNs is low. The principle of prompting rests on achieving a complementary synthesis of computer and radiologists' strengths: the former has more consistent visual search performance; the latter have superior interpretative skills. If prompt aid specificity is too low, however, radiologists will have to attend to many

FP prompts, and the effort of using the prompting aid may be perceived as outweighing its benefits. Worse, the overall effect could be to lower radiologists' specificity. Finally, prompting assumes that radiologists' assessment of non-prompted features is unaffected by attending to prompts.

PROMAM. PROMAM detects micro-calcification clusters and ill-defined lesions.[6,7] Prompts for the former consist of an irregular outline of the cluster, prompts for the latter consist of an ellipse surrounding the suspect region. PROMAM's prompt user interface consists of a hard copy, low resolution image of the mammogram pair with prompt information superimposed.[8] This 'low tech' approach was chosen 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.

Methodology

The techniques we employed in our investigations ranged from ethnographically-styled studies of clinic work and reading practices to controlled studies of the effects of prompting on radiologists' performance. Ethnography is acknowledged as a valuable technique for requirements investigation because of its capacity to bring out the social organisation of activities in the workplace, the practical participation of individuals in the collaborative achievement of work.[9] Gaining an understanding of breast screening work in this way was important for two reasons. First, introducing a computer prompting aid into breast screening would not only affect the work of radiologists, but of other clinic staff.[10] Second, though it has been customary to study the impact of prompting aids at the level of individual cognition, there is an obvious collaborative aspect to reading practice at many UK breast screening clinics.

Data collection methods in our investigations included participant observation, formal and informal interviews and documentary sources. The findings were used to suggest issues that could be followed up in controlled studies of the effects of prompting on radiologists. More than this, however, these studies were designed with the same underlying goal as the ethnography: to observe experts at work and to listen to them talk about it.

Data analysis followed a common pattern throughout. A series of reviews of notes, transcripts, etc. were conducted in which findings and evidence were used to elaborate one another.[11] The presentation of our findings makes use of extracts of people's commentaries about their own and each other's behaviour. This does not mean, however, that we have taken what people say they do at face value. To do so would be to risk treating the anecdotal as fact. To avoid this, our findings were corroborated by systematically triangulating data collected from different sources, settings, and people.

An overview of breast screening work

Six UK breast screening centres were investigated over a six month period. The centres are referred to here by the letters A through to F. Both observational and interview data were collected during a two month period of investigation at centre F, and during a one week period in each of the other five. Observations of reading sessions were conducted by asking readers to indicate and explain their reasoning when they encountered something 'interesting'. Where data is presented, the mode of collection is indicated (e.g., interview, field notes). Where a quote is attributed to a reader (R), the reader is identified by a number and the screening centre by a letter (A-F). Thus R1-C refers to Reader 1 in centre C.

Reading practices. Each mammogram is examined by at least one qualified person, typically a trained radiologist. With appropriate training, other medical specialists may also assume this role, so here the generic term 'reader' is used. On average, between 50 and 150 mammograms are read in a single session. Readers work through the cases on the viewer and mark their opinions on the screening form. If the reader decides that the case is normal, the woman will be screened again in three years' time. This is know as 'routine recall'. If a suspicious feature is found the reader recommends follow up tests at an assessment clinic. This is referred to as 'recall for assessment' (or simply 'recall'). If the reader decides that an opinion is impossible because of imperfect mammography then repeat mammograms may be requested. This is a 'technical recall'.

The degree of certainty of malignancy can vary considerably. Some features are unequivocally malignant, whereas others might be only mildly suspicious. There are various natural processes in the breast that can give the appearance of malignancy to varying degrees, and some malignancies mammographically 'occult'. It is common practice to classify features found according to probability of malignancy. For instance, at one clinic readers use a five point scale: C1 (normal), C2 (benign), C3 (equivocal), C4 (suspicious), and C5 (malignant), and set the recall threshold at C3.

In double reading, two readers give their opinions separately. If there is a difference of opinion, a final decision must be reached. Centres B, D, E and F use a 'worst opinion' rule, centre A uses third reader arbitration. In centre C, disagreements are resolved by discussion between the readers. Double reading is often done without any formal blinding of second readers to first readers' opinions. This is because the paper form on which readers record their opinions is routinely used as a source of additional evidence, (e.g., HRT status). In centre E reading is formally blinded by using a computer to record opinions. In the absence of formal blinding procedures, readers may avoid bias by employing tactics that decrease the accessibility of the first reader's opinion. The simplest involves "trying not to look" at the first reader's opinion before making their own. Comments made by readers confirm awareness of the possibility of bias:

> "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 R1-D)

One double reading study found a strong relationship between readers' sensitivity and the percentage of time they read second.[12] One interpretation is that second readers are 'prompted' by first readers and so pick up cancers that they might otherwise overlook. Two readers in centre A (R1-6 and R4-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 (R4-A) reads the batch in reverse order. This is done under the assumption that readers are likely to be more fatigued, and so more likely to make errors, towards the end of the batch.

Readers remarked on the logistical difficulties of making recall decisions by discussion, and expressed concerns that this too could bias results. When asked if they ever discussed cases, a reader in centre E replied that this was only done at review sessions and interdisciplinary meetings, stating that they were "worried about the effects of dominant

personalities" (field notes R3-E). A reader in centre B expressed similar concerns. In clinic D, a system of discussing recall decisions had been in place, but was discontinued:

> "[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, micro-calcs 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 R1-D)

Studies of double reading are typically concerned with its effects on sensitivity and specificity (e.g., Warren and Duffy[13]). However, by facilitating informal collaboration, double reading as typically practiced at the centres studied appears to serve a number of objectives that are arguably as important as any direct performance effects. We will now discuss these in 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 novice reader's autonomy 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 R1-C)

In centre C, an experienced reader (R1-C) was observed 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 case the senior reader scribbled the request out, and it was returned to routine recall. Centres B and E were also involved in training at the time of study. Both centres employ the worst opinion rule and have a policy of incrementally introducing novices to reading. Having attended a recognised training course, trainees then spend a period of time reading in the screening centre and discussing their opinions with experienced readers. Trainees are introduced into reading proper as a first reader.

In summary, training is organised to take advantage of the structure of double reading to provide a safe, supportive environment where trainees can be encouraged to make independent decisions. At the same time, the second reader is able to monitor and manage trainee errors.

Monitoring and feedback. The work of clinic staff is formally monitored through procedures for quality assurance and work documentation. Clinic staff hold regular meetings, for example to: compare radiological appearance and pathology data; review of interval cancers which may be evidence of FNs; and to discuss informally (and at some clinics, formally) differences in recall opinions.

Such meetings provide an opportunity for sharing experiences, such as reasons for reaching an opinion. Readers are concerned, however, to ensure their performance is consistent on a day by day basis and our studies suggest that they use double reading as an informal mechanism for achieving this. After returning from maternity leave, a reader from

centre E asked for all her cases to be double read so that she could check her performance. Another reader at centre E uses a similar mechanism to monitor his day to day performance. When reading second, he compares his opinions with the first reader's to see if he has missed a lesion, or classified one differently. He estimated that in cases differences of opinion, 3/4 of the time he has seen the lesion and has dismissed it, and in the remaining 1/4 he has overlooked it. 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 R1-C)

First readers effectively provide a standard for second readers. Feedback gained in this way may fulfil a number of functions. Readers can monitor their performance session by session and gain reassurance that intra-observer variations are compensated for. It may also play a role in maintaining readers' FP errors within a manageable range by reinforcing normative interpretations. In two clinics it was evident that informal feedback has evolved further: first readers sometimes annotate the breast schematic on the reporting form in routine recall cases:

> "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 R2-A)

Annotating implies that some characteristic of a feature warrants particular attention so that suspicion it arouses may be discharged. It enables first readers to demonstrate accountability to the decision-making process and second readers to assess their specificity.[14] A common annotation is to label a feature by writing "Benign" or simply "B". No reason is offered, 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, but no reason given. Both types suggest little doubt in the reader's mind that his or her opinion is correct; the annotation seems intended to reinforce it and to show vigilance. On occasions, however, "I think" and "?" is added. More complex annotations are also used which make explicit a reader's reasoning by referring to evidence used to mitigate the initial suspicion. Examples include "NRC" (no real change) -- the feature has not changed over time, so is less suspicious.

Double reading's reported sensitivity gains are due its capacity to turn inter-observer differences into an advantage.[15] However, if inter-observer differences become too great, then specificity may fall and changes in procedure, such as replacing a worst case recall policy with third reader arbitration, may be necessary. Feedback is the tool used to maintain reader differences within acceptable limits. Assessment clinics and annotations both serve to communicate and establish norms about the significance of particular kinds of features and their presentation. In this light, it is not surprising that readers annotate features falling either side of the recall threshold: this is precisely the region where reader differences are most likely to occur.

Summary. Readers are conscious of the extent and limitations of their expertise and apply these insights routinely in their work, particularly in respect to the management of errors. Readers show awareness of their sensitivity and specificity in respect of particular feature types and circumstances, and 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 helps alert them to the biases and errors to which they may be subject.

The study also shows how, through an informal, and still evolving extension of reading procedures, readers may use each other to help maintain errors within acceptable limits. There is a tension between the decision-making and monitoring aspects of reading, however. Access to a first reader's opinion is recognised as useful as a barometer of performance, but also as potentially harmful if it serves then to bias decision-making.

Studies of the effects of prompting

The key test for prompting is its effect on readers' performance. Owing to the low incidence rate of breast cancer, it is impossible to measure this quantitatively without a large scale clinical trial.[1] However, our initial investigation raised several issues that could be addressed using small scale studies. These included whether there is an upper limit to FP prompt errors before a prompting aid becomes useless, and whether some kinds of FP prompt errors are more tolerable than others. In turn, this raised the question of how readers make sense of prompting aid behaviour and what bearing this has on how they use it. Finally, it was important to examine the possibility that prompts may bias decision-making and to explore how access and interpretation might be managed to reduce any such effects.

In contrast to a human observer, computer-based image analysis will typically make use of only a subset of the available evidence, and will be limited in the ways in which it can combine evidence from different sources. Consequently, a prompting aid 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.

We conducted three studies of readers using PROMAM under pseudo-clinical conditions. The first was designed to elicit readers' subjective responses to prompting. Earlier work had suggested that prompting can improve performance only if the FP prompt rate is no more than 1.5 times the TP rate.[16] However, there are problems with extrapolating from such studies to the clinical setting as 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.

The results of the first study indicated that, under realistic reading conditions, readers' subjective tolerance for FP prompts was significantly higher than suggested by earlier studies.[17] One explanation is that FPs are not equally distracting. Readers attend to, and account for, a larger set of features within an image than they actually recall. FP prompts for accountable, 'candidate' features may provide readers with reassurance that they have made a thorough inspection of the image. Second, earlier studies simulated a prompting aid by randomly placing FP prompts and it was unclear whether they were representative of the types of FP that a prompting aid might actually produce. In contrast, 'real' FP prompts should be more consistent and will reveal something about the prompting aid's behaviour. Thus attending to FP prompts may also afford an understanding of how a prompting aid works, and what its capabilities are.[14]

To examine in greater detail readers' use of prompts, we devised a second study. Three radiologists from a Scottish breast screening centre were asked to comment on prompted features in the following ways:

1. Indicate whether the prompt would be acceptable in a screening environment.
2. Rate the prompt as 'useful' to 'distracting' on a five point scale.
3. State whether they would recommend recall on the basis of the prompted feature.
4. Classify each prompted feature.
5. Rate the significance of each feature on a five point confidence scale.

Subjects were also asked to annotate and describe any additional features for which they felt prompting would be useful and to rate such prompts in the same way as the actual ones. Finally, they were encouraged to give a verbal commentary on their interpretation of the mammograms and of PROMAM's behaviour using a 'think aloud' protocol. Subjects' commentary was tape recorded and transcribed. 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. Transcript extracts are labelled according to the subject (H, J or R), the session (1 or 2) and the case. E.g., (H-1.23) identifies the extract as belonging to subject H reading case 23 in the first session.

Accountability. Given the reading workload, it would be impossible for readers to exhaustively examine and analyse each part of each mammogram. Moreover, attention is a limited resource, and readers are prone to fatigue. Studies show that visual search in radiology is often incomplete[18] and that experienced readers are able to quickly 'zero in' on significant features.[19] Mammograms can be more or less difficult to interpret for a number of reasons, including variations in tissue type, and tissue distribution. Similarly, there can be variation in the difficulty of interpreting individual features. Some may be obviously benign or malignant, others may be ambiguous because they are in the early stages of development, or because they are imperfectly imaged. It is not necessary to attend equally to every feature within the mammogram: some can be cursorily dismissed, others require more protracted 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, examine closely features that 'catch their eye' (perhaps with a magnifying glass or a bright light source), use other evidence such as additional views or previous mammograms, or apply particular techniques, such as 'undressing' lesions. They may also pay greater attention to regions of the breast known to be sites of missed cancers -- the so-called 'review areas'. In short, readers demonstrate an array of tactics to render the reading task tractable.

Readers do not have to account for every feature within an image, but they do try to account for those that satisfy generally accepted heuristics for significance and to do so in a particular way. Accountability 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 feature, and by the analytic tactics most appropriate 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 I think that is useful to prompt ... 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 this, as in other cases, the prompted feature was judged to be benign and had no influence on the final opinion. However, subjects believed the prompt had served a useful function by encouraging a closer inspection. In so doing, prompting may have a psychological benefit by reducing anxiety about the thoroughness of the visual search. It may also improve readers' self-awareness and capacity for reflection. It is natural for readers to reflect on whether they saw the prompted feature, how much attention they gave it, and the interpretation they reached.

Context. One of the ways readers orientate themselves to their task involves attending to recognised problems. Subject H was keen for PROMAM to prompt 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, it's also a relatively common site of cancers." (H-1.3)

Here subject R observes that features that present in a particular region may be missed:

"Well sometimes we see wee cancers down there ... 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 ... 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)

Subject R's reaction to the prompt is interesting because it is clear that there are no micro-calcifications present. He finds the prompt acceptable because of the effect it has in drawing him to a region of the breast that deserves attention. Dense breast tissue complicates the task of interpretation and readers are aware that their judgements may be less reliable:

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

Subjects commented that a prompting aid might be useful in addressing this problem.

In the next case, subject H attends to mammograms rich with features that have suspicious characteristics, and is thus faced with the problem of differentiating between many confusing, attention grabbing, benign presentations and any actual malignancies. In the case of multiple presentations, additional effort is required to organise how the lesions might be considered:

"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)

In using the prompt to focus her analysis, subject H makes a tacit assumption that prompted features are more likely to be significant than unprompted ones. In the following extract, the presence of suspicious clusters heightens subject J's alertness towards micro-calcifications more generally. She is happy to have her attention drawn to other instances of micro-calcification so that they might be accounted for:

"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 the situation where they've got other clusters, then of course this could be another cluster of the same." (J-2.43)

Accounting for prompts. Using a prompting aid is not a simple matter of examining prompted regions for signs of cancer: prompt themselves demand interpretation.

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

In this extract, subject J makes a point of re-examining prompted regions to confirm her initial analysis. She takes reasonable steps to ensure that PROMAM has not detected something that the she did not initially apprehend and finally identifies a characteristic of the mammogram as a possible reason for the prompts. Readers are aware that minimal signs of cancer can be overlooked or misinterpreted, so a plausible explanation for prompts, in terms of both image properties and prompting aid behaviour, is sought. If a reason for a prompt is not readily apparent, then this can pose problems. 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)

Here subject H has a dilemma: has PROMAM detected something significant that she cannot see? There is no obvious cause for the prompt that can be used to account for its presence; there is no good reason for discounting it other than that she cannot see what it is for. A prompt does not explain itself, other than in the broadest sense of being produced by either the micro-calcification or ill-defined lesion algorithm. It simply highlights a region for examination. 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 feature detection algorithms work if a rationale is not obvious from the examination of the prompted region alone.

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 ... 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)

The prompt makes subject J think about recalling 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)

Readers face a dilemma in making borderline recall decisions because they know that some cancers present minimal signs, but recalling all border features would overwhelm assessment clinics. By way of a resolution, subject R formulates a heuristic by drawing an analogy between prompts and the effect of seeing a first reader's opinion when reading second. In a similar vein, 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? ... 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 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 PROMAM has assessed a region and found it to be benign.

Making sense of PROMAM. In subjects' repeated attempts to understand what prompts 'mean' and how they should be 'interpreted', we find evidence of their active engagement in making sense of PROMAM'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 so explain its responses. Strategies for making sense of PROMAM included:

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

Subjects expected consistency and were puzzled when PROMAM didn't prompt for features that were to them (diagnostically) similar to those it did prompt:

"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)

Subject J expresses confusion because of inconsistent prompting of seemingly similar regions of benign vascular calcification. This is due to lack of familiarity with PROMAM's behaviour and the effect of variations in image properties that may seem insignificant, or are difficult for a reader to perceive. The criteria for prompting micro-calcifications is base on a simple clustering rule. In its early stages, vascular calcification can be discontinuous or fragmented, and it is this type of presentation that satisfies the rule.

In the next case there are two ill-defined lesion prompts, A and B: A circles a large region of dense tissue in the upper part of the right breast; B circles a smaller region wholly within A.

"... so A, an increase in density and it does merit ... 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 ... B is in the middle of this bit ... I don't quite know 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 J feels that prompt A is relevant, but has difficulty finding an adequate explanation for B because it both misses the focal region of significance, and occurs within the region of prompt A, which is seen as significant. This poses several questions simultaneously: if a small focal area such as B can be prompted, why is A not more focal? If the entire region prompted by A is significant, then why bother prompting smaller regions within A at all? Such questions can be settled by an understanding of the ill-defined lesion algorithm. Processing is done in two stages. First, features within the mammogram are extracted and segmented according to four different scale sizes. This is a sieving operation which allows features falling into broad categories of size to be treated separately. Features within each category are then classified according to known properties of malignant lesions. The regions highlighted by prompts A and B belong to different scale sizes and their significances are unrelated.

Perceived inconsistencies in prompting may sometimes stimulate a search for more sophisticated explanations of 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 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 this case subject J is able to identify subtle differences between prompted and unprompted features, i.e., improved contrast between calcifications and background tissue. In another case, 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 informed by the assumption that not only is there some reason for the presence of a prompt, but also that it is diagnostically relevant:

"Why has it prompted that lymph node, and not others, I wonder? ... 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 ... 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." (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 ... 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.

The form of the prompts, and their relationship to the prompted feature, were chosen to provide certain types of information. Prompts for the micro-calcifications and ill-defined lesions are easily distinguished. Those produced by the former delineate the shape of the cluster detected, while the latter consist of an ellipse that circumscribes the lesion. However, subjects suggested that they learnt to recover additional information from prompt characteristics:

"Multiple prompts ... 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)

Summary. The results of this study confirm readers' tolerance of FP prompts and provide interesting pointers to how they may use prompts as an error management tool in the clinical setting. Overall, subjects' views on 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 even where the feature prompted has little or no significance.

The results also show that readers began to develop quite a detailed understanding of PROMAM as evidence of its behaviour accumulated. Memory of previous prompts

contribute to a biography of behaviour that may be used to account for new prompts. This working understanding of PROMAM is subject to incremental -- or sometimes radical -- revision as evidence accumulates. In principle, the user of a prompting aid does not need to know how it works, merely to attend to those areas in the image that it prompts. However, these results suggest that readers benefit from investing effort in making sense of PROMAM's behaviour because this facilitates more efficient and effective use, especially in managing the impact of FP prompts.

PROMAM is designed for use as an attention cue. The presence of a prompt simply implies that attention is required (because PROMAM is sensitive), but not that a recall is appropriate (because PROMAM is not very specific). The responsibility for assessing the significance of a prompted feature, and for making a recall decision, should rest with the reader. However, the results show that readers sometimes used PROMAM in ways that are contrary to this principle.

Pilot clinical evaluation of PROMAM

The first two clinical studies were followed up by a pilot clinical trial. Although providing less detail about individual prompted cases, the results pertain more directly to the interpretation of prompts in a screening context. In particular, we were able to examine more closely how, in the course of prolonged exposure to PROMAM, readers learned to both manage its errors, and to use PROMAM to manage their own.

Five subjects were recruited from radiologists at a Scottish breast screening centre and trained to use PROMAM. They were given a functional overview of how the image analysis algorithms work, including the types of FN and FP errors they make, and instructed to use prompts simply as attention cues. Data collection methods included observation of all the experimental sessions. Subjects were interviewed and completed a questionnaire immediately following prompted sessions; the interviews were tape recorded and subsequently transcribed. Further questionnaires were administered pre- and post-trial. A full account of methodology and results can be found in Hartswood et al.[20]

Dealing with FP prompts. Ideally, readers should give all prompts equal consideration, and only dismiss them after careful examination of the prompted region. However, interview data suggests that subjects developed other ways of determining the significance of prompts. 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 ... 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 subject indicated that some prompts could be assessed from examination of the prompt sheet alone. Evidently, subjects learnt from experience to recognise patterns in shape, frequency and location that characterise FP prompts and used this knowledge to determine how much effort to invest in further scrutiny of the mammogram. In such cases, consideration of possible explanations is not deferred until all the evidence has been gathered.[21] Two subjects commented 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:

> "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 ... you're kind of expecting prompts on certain things so you sort of ... very quickly dismiss it as (harmless?) without looking again."

There is a cost associated with this approach as it requires that readers form a more accurate model of prompting aid behaviour. However, checking whether prompts meet with expectations appears to be intuitive, and is perhaps essential for maintaining trust in aid performance. Also, playing the "prediction game" may be valuable because it reinforces the practice of making an initial assessment on mammogram evidence alone. Its success is dependent upon prompting aid consistency as perceived by the reader. Image analysis algorithms can be sensitive to variations in appearance which are too subtle for the reader to appreciate without close examination, if at all. PROMAM's behaviour is strictly deterministic, but as noted earlier, it may not seem observably so if it doesn't respond in the same way to features that readers would classify as similar.

Impact on decision-making. Despite the instructions given in training, both questionnaire data and post-session interviews indicate that subjects sometimes used prompts as classification aids. Subjects referred to occasions where they had found the absence of a prompt reassuring. However, they also reported cases where 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."

Comments suggested that the presence or absence of a prompt was most likely to influence subjects when the evidence of the mammogram alone was ambiguous. In these situations, they seemed prepared 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."

A comparison was made of subjects' recalls in the prompted and unprompted conditions. The results show that significantly more recalled cases were correctly prompted by PROMAM in prompted sessions, despite there being no significant difference in overall recall rates between conditions.[20] This corroborates findings from the interview data, that readers use the prompts to inform their classification decisions.

Summary. The results of the pilot clinical trial reveal how readers adapted to using PROMAM in the clinical setting. They show that subjects spontaneously improvised their use of prompts in a number of ways which helped them economise on the effort required to deal with PROMAM 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 evidence was ambiguous. However, there was no evidence that the use of PROMAM increased readers' FP decisions. Readers' recall rate in when using PROMAM was no higher than in an unprompted control condition.[20]

Summary and conclusions

The use of prompting aids in breast screening is intended to improve performance by helping readers avoid errors of attention. The problem lies in the fact that prompting aids themselves are not infallible. In the extreme case, low prompting aid specificity might negate their value altogether. Equally problematic is the danger is that while they may reduce some types of reader error, prompting aids could induce others. Our studies set out to investigate the relationship between reader errors and prompting aid errors, focusing on two main issues: how readers orientate themselves in current practice to the problem of managing errors in their own performance, and how they learn to cope with prompting aid errors.

While reliable quantitative evidence of performance improvement must necessarily await the outcome of large scale clinical trials, our studies provide some support for the achievement of the prompting aid goal. They also show that readers can learn to manage the undesirable effects of prompting aid errors on their specificity, and can tolerate a higher FP rate than previous work had suggested. However, we also find that presumptions of prompts being used purely and simply as attention cues may be misplaced.

The design rationale for prompting aids assumes generic difficulties -- i.e., that readers sometimes have difficulty ensuring that the entire mammogram is examined. However, our findings show that the uses to which prompts are put are often highly contingent. The problems readers face are actually very specific and contextualised. For example, reading dense, or feature rich, breasts poses demands very different from those of lucent, or uncomplicated, breasts. Furthermore, although readers have general concerns that they might overlook a malignancy, they also have a more specific understanding of particular deficiencies in their expertise. For example, readers 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 a prompting aid in breast screening can be achieved if readers are expected to treat it as a mere 'black box', with no understanding of how it behaves. On the contrary, our studies show that efficient and accurate use of prompting aids depends on their behaviour being accountable to readers. For example, as they became more experienced, readers were able to develop heuristics that enabled them to predict FP prompts. On the other hand, readers often came up with ad-hoc and mistaken explanations for prompts, even occasionally falsely analogising PROMAM with reader behaviour. We have attempted to address this issue through reader training using worked examples of common TP, FP and FN prompts. In the pilot trial, however, we still found examples of prompts that readers were unable to account for, suggesting that the training materials need further work.

Training will always have its shortcomings, however, if no attention is paid to supporting learning in use. To this end, we plan to investigate ways in which the design of the prompt user interface can be enhanced so that how PROMAM works becomes more observable to its users. The basic approach would be to extend PROMAM's outputs (i.e., the prompts) by adding a causal account of the processing leading to each prompt. This would be

presented so that it was available for inspection, but without interfering with the essential simplicity of prompt use.

To address the issue of how prompting aids should be used, 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 pilot clinical trial whereby subjects examined each mammogram before examining the prompts, 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 recalls to be amended.

Finally, the goal of replacing double reading with a single, prompt assisted reader raises wider problems. Our studies show that there is an informal, collaborative dimension to double reading, and to readers' management of errors, which has so far been ignored. Though readers may evolve ways of compensating for the move to single, prompted reading, we argue that its implications deserve careful consideration. At the least, we suggest that provisions be made for a transitional period where readers double read with prompts. This would enable them both to gain experience of a prompting aid under more familiar reading conditions and, by exploiting double reading's collaborative affordances, to enhance their understanding of its behaviour by learning from one another as they use it.

Acknowledgements

We wish to thank the PROMAM team and the staff of various UK breast screening centres, especially those at the South East Scotland Breast Screening Centre, for their support and co-operation. This work was supported by Scottish Office Home and Health Department grant K/CSO/1/327.

References

1. L. Williams et al., "Computer-aided cancer detection and the UK National breast screening programme," in *Proceedings of the Fourth International Workshop on Digital Mammography*, ed. N. Karssemejer et al. (Netherlands: Kluwer Academic Publishers, 1998).
2. J. Chamberlain et al., "National Health Service breast screening programme results for 1991-2," *British Medical Journal*, no. 307 (1993): 353-356.
3. H. Kundel et al., "Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection," *Investigative Radiology*, no. 13 (1978): 175-181.
4. C. Savage et al., "To err is human to compute divine?," in *Proceedings of the Second International Workshop on Digital Mammography*, ed. A.G. Gale et al. (Amsterdam: Elsevier Science, 1994).
5. M. Giger, "Computer-aided diagnosis," in *A categorical course in physics: Technical aspects of breast imaging*, ed. A. Haus and M. Yaffe (RSNA, 1993).
6. A. Hume et al., "On the evaluation of micro-calcification detection algorithms," in *Proceedings of the Third International Workshop on Digital Mammography*, ed. K. Doi et al. (Amsterdam: Elsevier Science, 1996).
7. L. Miller and N. Ramsay, "The detection of malignant masses by non-linear multiscale analysis," in *Proceedings of the Third International Workshop on Digital Mammography*, ed. K. Doi et al. (Amsterdam: Elsevier Science, 1996).
8. R. Procter et al., "User interface design and data management for digital mass mammography," in *Proceedings of the Second International Workshop on Digital Mammography*, ed. A.G. Gale et al. (Amsterdam: Elsevier Science, 1994).
9. *Technology in working order: studies of work, interaction, and technology*, ed. G. Button (London, England: Routledge, 1993).
10. M. Hartswood and R. Procter, "Designing for Breakdowns and Repairs in Collaborative Work Settings," *International Journal of Human-Computer Studies* (Academic Press, forthcoming).
11. H. Garfinkel, *Studies in Ethnomethodology* (Englewood Cliffs: Prentice Hall, 1967).
12. L. Williams et al., "Methodological issues in mammography double reading studies," *Journal of Medical Screening*, no. 5 (1998): 202-206.
13. 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*, no. 68 (1995): 958-962.
14. M. Hartswood et al., "Drawing the line between perception and interpretation in computer-aided mammography," in *Proceedings of the First International Conference on Allocation of Functions*, ed. E. Fallon et al. (Dublin: IEA Press, 1997).
15. C.E. Metz and J. Shen, "Gains in accuracy from replicated readings of diagnostic images," *Medical Decision Making*, no. 12 (1992): 60-75.
16. I. Hutt, "The Computer-Aided Detection of Abnormalities in Digital Mammograms," (Ph.D. thesis, University of Manchester, 1996).
17. M. Hartswood et al., "Subjective reaction to prompting in screening mammography," in *Proceedings of Medical Image Analysis and Understanding'97*, ed. C. Taylor (Oxford, 1997).
18. C.F. Nodine and H.L. Kundel, "Eye Movements: From Physiology to Cognition," in *The cognitive side of visual search in radiology*, ed. J.K O'Regan and A. Levi-Shoen (Elsevier Science, 1987).
19. C.F. Nodine et al., "Nature of expertise in searching mammograms for breast masses," *Academic Radiology*, no. 3(12) (1996): 1000-1006.

20. M. Hartswood et al., "Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography?," in *Proceedings of the Fourth International Workshop on Digital Mammography*, ed. N. Karssemejer et al. (Netherlands: Kluwer Academic Publishers, 1998).
21. A.G. Gale, "Human response to visual stimuli," in *The perception of visual information* , ed. W. Hendee and P. Wells (Springer-Verlag, 1995).