Methodology for taking a computer-aided breast cancer screening system from the laboratory to the marketplace

Linda Jane Williams, B.Sc., M.Sc.

Doctor of Philosophy
University of Edinburgh

2001

To my husband, Rob. I couldn't have done it without you.

Abstract

Breast cancer is one of the most common causes of death in women, and yet is one of the more 'curable' cancers if caught early. Since its inception in 1987, the Breast Screening Programme has been the principal tool in the National Health Service's fight to reduce the number of cancer related deaths in the UK.

Breast screening using mammography is widely viewed as the most effective way of detecting early breast cancer, with the UK population of women over the age of 50 being invited to a screening session every three years. However, national shortages of clinical staff willing to enter and remain in this field mean that the NHS Breast Screening Programme is severely understaffed.

This thesis discusses one way in which technology can assist in the screening programme; specifically, the use of a computer-aided cancer detection system. Here, we will present the design and analysis of a sequence of experiments used to develop and evaluate such a system. PROMAM (PROmpting for MAMmography) involved the scanning and digitising of mammograms, and the subsequent analysis of the digital image by a series of algorithms.

Initial evaluation was done to ensure that the algorithms were performing satisfactorily at a technical level before being introduced into a clinical setting. Two large experiments with the algorithms were designed and evaluated:

1. offering radiologists three levels of algorithm prompting and, as a control, an unprompted level, on samples of mammographic films, with outcomes being their recall rate and subjective views at each prompting level,

2. a pre-clinical experiment, conducted under semi-clinical conditions, where two readers would see a batch of films seeded with higher than normal numbers of cancers, with readers allocated randomly to prompted and unprompted views of films.

The first experiment was designed using a Graeco-Latin Square, with three 'nuisance' variables and the treatment factor of prompting levels (no prompts, low level of prompting, medium and high). Four radiologists read at each level of prompting once, on different sets of films. One of the more interesting results was that the recall rate did not increase as the prompting rate rose - contrary to prior expectations. Most of the differences seen between the prompting rates could be explained as radiologist differences.

Once these were taken into account, the level of prompting had little effect. Additionally, although the time taken to read a set of films increased as the prompting rate increased (as would be expected), it was only an increase of 26% from the unprompted set to the set with the highest number of prompts. Observational data suggested that the lowest level of prompting was not maintaining the interest of the radiologist, thus leading them to neglect the prompts.

The following experiment moved the system a step closer to a true clinical demonstration of the efficacy of PROMAM, being conducted under semi-clinical conditions. Using a method of minimisation, the number of cancers each radiologist viewed as first reader, second reader, prompted or unprompted were balanced. Preliminary exploratory analysis indicated that the recall rate declined with the introduction of the prompting system, but more detailed, analysis indicated that much of this difference was due to a radiologist effect. Although cancer detection was slightly lower with the prompting system, examination of the 11 cancers missed by the prompted radiologist showed that six of these had been correctly prompted by the algorithms. This demonstrated scope to improve the cancer detection rate by nearly 5%.

These experiments determined the 'production' version of the prompting system. A design to evaluate the system in a sample of 100,000 women in six centres was produced, but due to circumstances beyond the project team's control, it was not possible to take this work to the stage of a full 'trial' of the system. The design concept can, however, apply to the evaluation of any similar prompting system. The recommended design is therefore presented, together with an analysis of data from a simulated application of this design.

This simulation has allowed recommendations to be made on the most appropriate ways to analyse the extensive and complicated dataset that will be obtained. In particular, it identified technical problems that can arise from the application on one candidate analytical method, and an explanation for the failure obtained

It is quite clear from the evidence presented in this thesis that there is much scope for improvement in the cancer detection rate by the use of a prompting system, without a corresponding loss in the specificity. With the shortage of radiologists and radiographers, and the increasing demand placed on the Breast Screening Programme, technology could play a beneficial role in screening for breast cancer in the coming years.

Acknowledgements

I would like to thank all the members of the PROMAM team, in particular Mr William Cormack, for making the project a stimulating and enjoyable experience. This thesis could not have been possible without co-operation between all aspects of the project. I would also like to thank the radiologists and staff at the Edinburgh and Glasgow Screening Centres for their patient co-operation with our often strange requests and protocols.

Thanks are also due to my supervisors, Dr. Robin Prescott and Professor Freda Alexander. Without their encouragement, this thesis may never have been completed.

Lastly, I would like to thank my husband, Rob, for teaching me LaTeX and for viciously pointing out on numerous occasions that he had a PhD and I didn't. Without such 'support' I might never have started this thesis at all.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Linda Jane Williams, B.Sc., M.Sc.*)

Table of Contents

List of Figures

List of Tables

Preface

This thesis is the culmination of work done for the PROMAM (PROmpting for MAMmography) project, a computer-aided detection system for screening mammography. PROMAM involved the scanning and digitising of mammograms, radiological images of the breast, and the subsequent analysis of the digital image by a series of algorithms. Each chapter shall consist of a short description of the chapter, followed by a more detailed introduction and will then explore the topic in greater depth.

The aim of this project was to transfer the technology developed by the Royal Observatory, Edinburgh, into other fields, namely mammography. The PROMAM team created and built algorithms, developed graphical user interfaces and trialed the system under semi-clinical conditions, in order to produce a computer-aided detection system capable of increasing the sensitivity of radiologists without a concurrent loss of specificity.

Chapter 1 introduces the concept of screening for breast cancer and the history behind it. Methods of improving cancer detection rates are discussed.

Much work has been undertaken in the field of digital mammography, including the PROMAM project. Chapter 2 looks at the present state of affairs and introduces the PROMAM project in greater detail.

Once the algorithms had developed to a minimum satisfactory level, they were tested in a clinical setting. Chapter 3 examines the work done and the radiologists' reactions to the prompts and prompting levels.

Chapter 4 details an experiment conducted under semi-clinical conditions, where two readers would see a batch of films seeded with higher than normal numbers of cancers. One reader would be prompted, the other unprompted. This was minimised to ensure that no one reader would see more cancers as first or second reader, or as the prompted or unprompted reader.

Due to circumstances beyond the project team's control, we were unable to take this work to the stage of a full 'trial' of the system. The concept will, however, apply to the evaluation of any similar prompting system. The recommended design is therefore presented, together with a simulation of data from such a design. This may be seen in Chapter 5.

Much of the work in this thesis will refer to work done by other members of the PROMAM team, principally Dr Mark Hartswood. Where the work was undertaken by others, this is explicitly noted in the text. Where decisions were made by all involved, these are noted as being made by 'the team'. Otherwise, the work was conducted by the author.

Mark Hartswood's PhD concerned the evaluation of PROMAM from a human factors perspective; in particular, his work is concerned with understanding how readers made sense of PROMAM in experimental settings informed by ethnographic studies of everyday screening practices.

Dr Hartswood presents five studies, only two of which are referred to in this thesis. These two studies were designed in collaboration with the PROMAM team (with the author of this thesis responsible for the 'statistical' design of the experiments and Dr Hartswood for protocols concerned with 'human factors' aspects of the evaluation). Both Dr Hartswood and the author were responsible for the day to day running of the experiments, and were supported in this by the rest of the PROMAM team (e.g. in withdrawing films from the archives, running algorithms to produce prompts etc). In addition to quantitative measures, as reported in this thesis, the experimental protocols also included interview data, observation and free-response questions in questionnaires.

Dr Hartswood's PhD made extensive use of the qualitative data in the reporting of these experiments, using quantitative results to provide a context in which to place the qualitative findings. Dr Hartswood's PhD credits Linda Williams for all statistical work.

PROMAM was funded by the Scottish Office, EPSRC and PPARC.

Chapter 1

Introduction

1.1 Aims and Objectives

This chapter will lay the foundations for the work discussed in later chapters. The inception and rationale behind the National Health Service Breast Screening Programme will be examined, along with a description of the screening process. Finally, the various methods of improving the cancer detection rates will be discussed.

1.2 The UK Breast Screening Programme

Breast cancer is one of the most common causes of death in women, with around 7% of women in Scotland developing the disease at some time in their life and about 4% dying from it. In Scotland alone there are approximately 3000 new cases every year [1]. Currently, mammography (radiological imaging of the breast) is the most effective method of early detection of breast cancer. Since the late 1980's, it has been the principal tool in the fight to reduce the number of cancer related deaths in the UK's National Health Service Breast Screening Programme (NHSBSP).

Women between the ages of 50 and 64 are routinely invited to screening clinics every three years, although women aged over 64 may also request regular screening. In the UK, with an uptake of roughly 70 - 75%, this means about 1.6 million women are screened every year. Each film is examined by at least one radiologist (double reading is the official standard in Scotland only, although many clinics in England and Wales also double read), and suspicious cases are recalled for further examination. Approximately 5% of the screened women are recalled due to a suspicious region on their mammograms, and about 10% of these have malignancies.

Radiological imaging of the breast using soft tissue X-rays is currently the only way to detect non-palpable lesions, although several alternative methods (e.g. changes in body temperature within the breast tissue [2], analysis of hair structure [3]) are in development. Early detection, particularly before the lesion can be detected by palpable examination of the breast, leads to a potentially higher chance of curability and a longer expected survival time [4].

1.2.1 How it began

In 1985 a working group was appointed under the chairmanship of Professor Sir Patrick Forrest to examine the available evidence on the efficacy of screening asymptomatic women for early signs of breast cancer [5], and to suggest what policies should be implemented in the light of their conclusions. Several large scale randomised controlled trials had been completed by this time, principally the Health Insurance Plan (HIP) of Greater New York [6], the Swedish Two County Trial [7] and the two Dutch studies (Nijmegen [8] and Utrecht [9]), although the latter two were not randomised controlled trials but rather case-control studies. Two UK trials had begun [10, 11], but neither were sufficiently advanced at that time to contribute mortality data.

In the same year, breast cancer accounted for 20% of all female deaths by cancer in the UK (4% of deaths from all causes), and 27% (12.5%) of deaths in women between the ages of 45 and 64 (the ages of the women participating in the UK trials). England and Wales had the highest mortality rates from breast cancer in the world, closely followed by Scotland and Northern Ireland, with no immediate hope of prevention of the disease and only a poor chance of survival. Two thirds of women who developed breast cancer were still likely to die from it. Some form of intervention was urgently required.

The Working Group's main conclusions were thus:

- The information from the overseas trials suggested that deaths from breast cancer in the 50-64 years age group could be reduced by at least a third.
- High quality mediolateral oblique (diagonally across the chest) view mammography was the preferred option for mass screening
- There was insufficient evidence on the optimum frequency for routine repeated screening, and more research was to be done. Three years was suggested as an initial interval, but should be periodically reviewed.

Thus, in 1987, the National Health Service Breast Screening Programme was officially launched nationwide, becoming the first truly national programme, independent of Health Authorities and Districts, basing their scope only on 'Forrest units', a theoretical measure indicating the number of women one radiologist and their support team could cope with in the period of one year. Its aim was "to reduce mortality from breast cancer by regularly screening women in order to identify and assess abnormalities and treat those which are diagnosed as cancer", with the target of reducing breast cancer deaths by 25% by the year 2000 compared to 1990 figures [12].

1.2.2 Changes since inception

Age range

One of the factors that makes the UK screening programme uncommon, although not unique, is the upper age limit on the screening invitations. Once a woman reaches 65 years of age, she stops being automatically invited to screening, although she can request an appointment. However, a large proportion of breast cancers in this country are in the over 65 age group. Cancer rates are higher in the older age groups, as can be seen in figure 1.1.

In fact, there is only one country in the IBSN/European Network group that has a lower upper age limit [13]. Finland has an age range of 50-59, with screening every

2 years, but has an age adjusted annual breast cancer death rate of 16.6 per 100,000. Many other countries have extended the upper limit to 69, with some having dispensed with a limit completely. A recent pilot study [14] suggested that women over the age of 65 would respond favourably to an invitation to screening, with similar response rates to those in the women 50-64 years of age. The NHS has recently completed the pilot stage for extending the upper invitation age range to 69 [15, 16], with the result that the upper age limit for invitation would be extended to women up to the age of 70 [17].

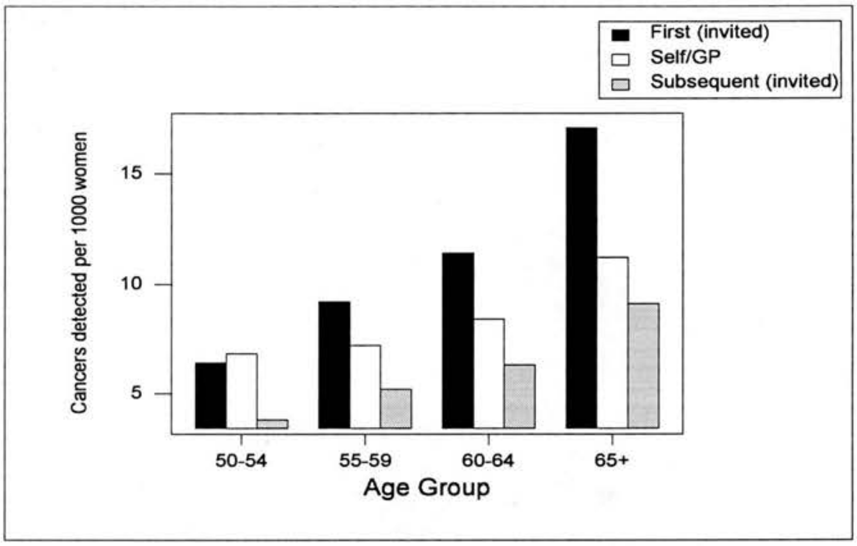


Figure 1.1: Cancer detection rates by age group [17]

A computer simulation by Boer *et al* [18] examined the effect of extending the age range to 69, as well as examining the effect of reducing the interval to two years (see section 1.3.5). It suggested that an increase in the age range would increase the reduction in mortality from 12.8% to 16.4%, leading to a relative improvement in mortality of 28%.

Extending the age range in the opposite direction, however, would seem to be contraindicated. Cancers in the under 50 age group are harder to detect, possibly due to the physiology of the pre-menopausal breast tissue. Sensitivity (the ability to detect cancers) in this age group is much lower than for the 50-69 and ≥ 70 groups (64%, 85%, 80% respectively in an analysis of the Nijmegen screening programme [19]), indicating that the benefit of screening is likely to be low in younger women. Kavanagh *et al* [20] agree, claiming that the sensitivity of screening mammography increased significantly with age, from 49.4% in the 40-49 age group to 85.2% in women aged 70-79. Meta-analysis of studies between January 1966 to October 31 1993 by Kerlikowske *et al* [21] concurs, with the relative risk for breast cancer mortality of screened women compared with women who had not been screened significantly lower than one for women aged 50 to 74 (0.74, 95% CI 0.66 to 0.83), while the relative risk for women aged 40 to 49 was not significantly different to one (0.93, 95% CI 0.76 to 1.13).

International Union Against Cancer [22] disagree, however, stating that screened women in this age group had a relative risk of less than one. Results from the Swedish screening programme suggest that the relative mortality associated with screening was 0.77 (95% CI 0.59 - 1.01). They also proposed that women under the age of 50 should be screened every 12 - 18 months, to take into account the more rapid progression of cancers in the pre-menopausal age groups.

Screening interval

The screening interval in the UK is still three years, but this is more due to pressures of turnover and cost than because this is the optimum frequency. Later work in this field has suggested that two years is more likely to be the optimum interval [23, 18]. This will be examined in more detail in section 1.3.5.

The results of longer term follow up

Later work, principally the ten year follow-up to the Edinburgh randomised trial [24], suggested that the expected decrease in mortality rate from breast cancer was unlikely to be as high as 30% (as stated by the HIP study [6]), putting it at about 18%. Other studies have reported similar effects at ten years [25].

Controversy has recently been stirred by the publication of an article by Gøtzsche and

Olsen [26], which claims that the evidence for breast cancer screening is flawed, stating that the evidence from several large scale trials is unreliable because of imbalances in the age and social class between the screened and unscreened populations. This has been comprehensively criticised in other articles [27], including the commentary from the same journal [28], and letters in the following edition of the Lancet (vol 355, February 26 2000).

1.2.3 Current state of play

In the accounting year 1998/9 (reported in 2000) [17], 1,699,727 women over 50 years old were invited to a screening session in the UK, of whom 1,290,126 (75.5%) accepted. A further 116,191 women attended screening as self or GP referrals. 68.2% of these self/GP referrals were over 64, when automatic routine recall (an invitation to screening every three years) ceases. The remainder were women who have perhaps previously declined an invitation and subsequently changed their minds, or possibly moved as the screening round moved into their area, or they themselves believe they have a problem in their breasts.

Table 1.1 summarises the data that the breast screening programme considers the most important indicators of performance. One of the most encouraging indicators is the number of small cancers (those less than 15mm). Small cancers such as these have a greater probability of complete excision without recurrence, improving the woman's chances of survival dramatically.

From 1992/3 the breast screening service has been inviting a full third of the population for screening each year, resulting in a three year rotation between visits. Since then, attendance (by invitation) has risen from 71.3% to 75.5%, with a peak of 76.7% in 1994/5 (see figure 1.2). 1992/3 also saw most of the programme enter its second screening round, where the women who were prevalent (first screening) three years previously were now invited for a second screen. By 1999, some screening centres were entering their fourth screening round.

	Accounting year 1998/9
Invited	1,699,727
Screened (invited)	1,290,126
Uptake (invited)	75.5%
Screened (self/GP referrals)	116,191
Total screened	1,406,317
Recalled for assessment	76,114
Recall rate	5.4%
Benign biopsies	2,033
Cancers	8,771
Cancer detection rate	6.24 (per 1000)
In situ cancers	1,733
Invasive cancers <15mm	3,722

Table 1.1: Data summarised from NHS Breast Screening Programme Review 2000 [17]. Unless otherwise stated, the figures refer to the number of women.

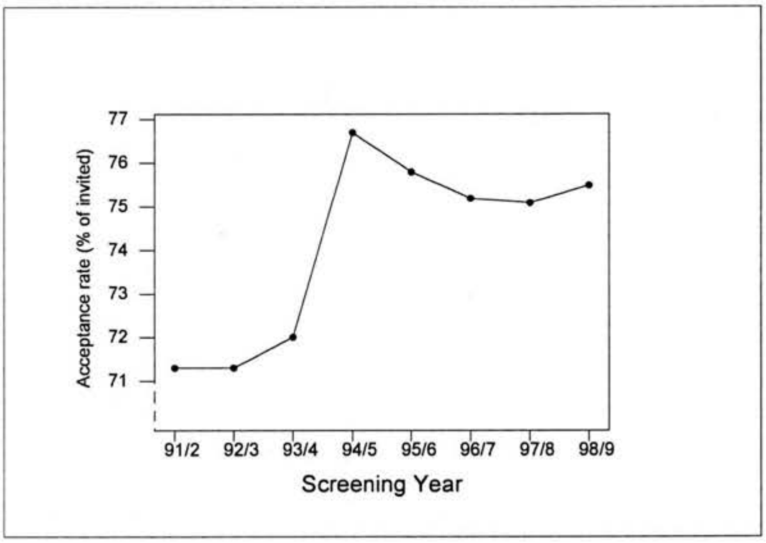


Figure 1.2: Uptake by year [Source: NHS Breast Screening Programme Review 1994 - 2000]

Theoretically, the ratio between first and subsequent screening would be 1:4 (invitation only), if all invitations were accepted. However, not all first time screeners fall into the 'first invite/first screen' category, making them older than the usual 50-52 years of age. The acceptance rate of women who had previously refused an invitation is considerably lower than that of the first screen/first invite group; 22.9% as opposed to 73.8%. Uptake in subsequent screens is encouraging; only 13% found the experience unpleasant enough not to return. Women who were recalled sooner than the usual three years had the highest acceptance rate at 96.4%, reflecting their, quite naturally, higher state of anxiety.

The fastest growing screening group is the over 65 group, who had been previously screened, returning for further screening by self-referral. This number jumped from 66,889 in 1996/7 to 86,214 in 1997/8, although the number of women of this age group attending screening has been increasing steadily since 1992/3 [15]. 90,599 women over the age of 65 (22,185 invited, 68,414 self/GP referral) attended for subsequent screening in the 1998/9 accounting year. This figure is complicated by the introduction of the three pilot sites looking at the extension of the programme, which accounts for some of the invited women. This increase may be due to campaigns by several charities involved in the welfare of the aged, ensuring that those over 65 knew that they had the right to request screening. A recent survey by the charity Age Concern [29] found that many women over the age of 65 believed that they were no longer at risk of developing breast cancer.

1.2.4 Progress through the screening process

As this thesis reaches a more applied stage, it will refer to points within the screening process. Without some idea of the sheer volume of effort that screening the eligible population of women involves, some of the work carried out under the auspices of the PROMAM project may seem somewhat trivial. And so, what follows is a simplified description of how screening is accomplished in practice.

Screening lists are generated by the GPs within each screening region. For instance, if a mobile unit (like a large mobile home, equipped with the photographic equipment for taking mammograms, and staffed by two or more radiographers) is due at a particular region, the GP surgeries that fall into that area send a list of their patients who are between 50 and 52 to the local Screening Centre. From this, and lists compiled by the centre from previous screening rounds, letters are sent out to all eligible women, with an appointment date and time.

These women present themselves at the screening unit (either the centre or a mobile unit), at which point the images of each breast are taken. For women attending for the first time four images are taken; two of each breast. This process takes between five and ten minutes. Mediolateral obliques are taken by compressing the breast diagonally from shoulder to stomach, whereas the the cranio-caudal (CC) view compresses the breast horizontally.

The resulting images are developed at the local Screening Centre, where they are then examined by at least one radiologist. The majority are passed as normal (approximately 95%), with the remaining women being recalled for further examination to a review clinic. In some cases, this may only be for technical reasons, for example if the image was poorly developed or the breast was not compressed enough to make out the mammographic features. For others, the recall is due to a suspicious feature that may be a cancer or may be entirely benign. Further examination may involve a range of techniques - magnification mammograms (focusing on the area of interest only), ultrasound, fine needle aspiration, and so on. Again, most of the recalls are later returned as normal, with only about 10% of suspicious recalls turning out to be genuine cancers.

Surgery usually follows the discovery of a cancer. For some women, pre-operative endocrine therapy (Tamoxifen or some other anti-œstrogen) may be taken to shrink the cancer to a manageable size before surgery, thus improving their chances of breast conservation. In many cases, the cancer is small and may be dealt with by wide local excision (or lumpectomy), the removal of the cancer and a margin of tissue around it only. This method leaves much of the breast untouched, resulting in a more aesthetic outcome. Mastectomy is the more radical surgical method, removing all the breast, usually in cases where the malignancy is large or multifocal. In most instances, surgery is followed by a three to five week course of daily radiotherapy in order to ensure the destruction of all cancer cells, and in some cases, a course of chemotherapy is recommended to improve the chances of the cancer not recurring. Additionally, Tamoxifen or other anti-œstrogen tablets are prescribed for those women who are œstrogen receptor positive, to be taken daily for the following five years. Naturally, women who have had cancer are monitored more often than the population at large.

1.3 Improving cancer detection

The Forrest report [5] recommended that single reading (one radiologist alone views a set of films and makes the decision whether to recall that woman) be adopted as the standard practice within the national service, although this practice tends to vary from clinic to clinic, with clinicians choosing the method they believe to be best, under the restrictions of time, staff and money. A number of ways of improving sensitivity have been implemented [30], several of which are presented below.

At the time that the research for this thesis was conducted, there were a number of methods in use for assessing improvements in sensitivity. Many of these metrics were influenced by the clinical procedures used during the assessment; for example, where the second reader was blinded to the decisions of the first. Others were misapplied, contravening some underlying assumptions. Hence, a paper [31] was published, highlighting these issues, which is also described below.

1.3.1 Double reading

The actual improvement due to double reading (two or more radiologists examining the same films and coming to some sort of consensus decision upon the recall of the woman) quoted in the literature varies dramatically from paper to paper. The quoted improvements in cancer detection reported range from 1.5% [32] to 15% [33], with reported changes in recall rate varying from a decrease of 45% [34] to an increase of 37% [35]. While the lowest figure for cancer detection improvement would suggest that the small improvement in sensitivity due to a second reader is an inefficient use of finite resources, the higher figure would seem to justify double reading.

Simple double reading

Simple double reading is merely two radiologists reading each set of films and noting their decision (usually 'routine recall', 'return for further assessment' or 'technical recall') on a form. However, even this simple configuration can have its variations.

Blinded Blinding radiologists, despite its emotive term, simply means that the second reader is unaware of the decision of the first. Both readers are then completely independent of each other and may return substantially different recalls (that is, the women that they decide require further assessment).

Non-blinded In this instance, the second radiologist is aware of the first reader's decision. When this is the case, the second reader tends to agree with the first radiologist more frequently than had they been blinded - in other words, the second reader is using information from the first reader to inform their recall decision.

One consequence of the absence/presence of blinding means that various strategies become difficult to compare. A commonly used statistic for the calculation of the improvement due to double reading strategies is the Mean Second Screener Contribution (MSSC) [33, 36] . In effect, this is the average number of cancers detected by only one of the radiologists, divided by the average number of cancers found by each radiologist. The number of cancers reported by each radiologist is usually summarised in the following manner, where R_1 refers to the first reader, and R_2 refers to the second reader. 'Cancers detected' is indicated by a '+', 'cancers missed' by a '-'. For example, R_1+ simply means 'number of cancers detected by the first reader'.

	R_2+	R_2-	
R_1+	a	b	a+b
R_1-	c	d	
			n

Table 1.2: The structure used to calculate the MSSC

From this, the mean second screener contribution is defined as:

$$\mathrm{M} = \frac{\frac{\mathrm{b+c}}{2}}{\frac{\mathrm{(a+b)+(a+c)}}{2}}$$

which is generally presented as

$$\mathrm{M} = \frac{\frac{\mathrm{(b+c)}}{2}}{\mathrm{a} + \frac{\mathrm{(b+c)}}{2}}$$

This can be rearranged to give the slightly simpler form of:

$$\mathrm{M} = \frac{\mathrm{b+c}}{\mathrm{2a+b+c}}$$

For example, if we make the assumption that each reader has the same chance of detecting cancers missed by their counterpart when they read blind, then we potentially have the situation illustrated below. In the first case, the second reader is not blinded, and can therefore be influenced by the decisions of the first reader. In the second, the second reader has no information as to the first reader's decision[1].

	R_2+	R_2-	
R_1+	170	2	172
R_1-	19	0	19
	189	2	191

Table 1.3: Not blinded: M = 5.82%

	R_2+	R_2-	
R_1+	153	19	172
R_1-	19	0	19
	172	19	191

Table 1.4: Blinded: M = 11.05%

Despite the same number of cancers being detected overall *(a+b+c)*, the MSSC gives considerably different results, all due to the fact that the second reader is using the information from the first radiologist to increase the probability that they will recall a particular case if it has already been recalled by the first reader. The proof that the MSSC of an unblinded reader will always be less than or equal to that of a blinded reader is given in Appendix B.

[1] Data based on Anderson's paper [37]

Variations on a theme

Many breast screening centres use variations of the two basic methods, blinding and non-blinding, of double reading.

1. Removal of first reader's recalls - this is a variation on non-blinding. The first radiologist removes all the cases that s/he considers suspicious and worthy of recall, and so the second reader knows that the cases that are left are the ones that the first reader considered to be normal. As with simple non-blinding, the fact that the second radiologist knows the decisions of the first reader causes problems when faced with calculating the relative improvement due to the second reader under some of the methods used. An assumption must be made about the agreement between first and second readers, since the option of the second reader disagreeing with the first reader's recall decision is removed along with the films.
2. Consensus on recalls. Either the recalls where first and second reader disagree (when one reader recalls and the other does not) or all the cases recalled by either radiologist are discussed. Usually this just involves the two readers who read the cases, but may also include up to the entire complement of radiologists in a centre [38]. This method is an excellent way of reducing false recalls (false positives), but may lose potential cancers.
3. Ombudsman system. This refers to a third radiologist making a decision on any cases where the two screening radiologists disagree. This third reader is usually a senior radiologist, often the Director. As with consensus, this method reduces false positives, but has the potential to miss cancers.

As mentioned earlier, much of the available literature uses differing methods of calculation for estimating the increase in cancer detection due to the influence of the second reader. Table 1.5 demonstrates the vast procedural differences reported by various studies, most of which were retrospective.

[2] All Scottish breast screening service radiologists

	Sample size	Age range	Radiologists	Cancers	Randomised trial	Recall criteria	Blinded
Denton [32]	62.5% of 36,320	>50	2	225	No	Worst case	Yes
Thurfjell [33][36]	11,343	40-74	2	76	Unknown	Discussion of flagged cases	Yes
Anttinen [34]	15,547	50-59	4	68	No - reader always first or always second	Flagged cases reviewed by both	Yes
Deans [35]	Not given	>50	35 over 4 years[2]	2473	No - varies across clinics	Worst case (except Glasgow, third reader)	No
Warren [39]	33,734	>50	3	269	No - by chance	Consensus or review by senior radiologist	Yes (on initial reading)
Anderson [37]	28,170	>50	3	191	No - first in usually first reader	Worst case	No
Ciatto [40]	18,817	50-70	4	125	Not clear	Worst case	Yes

Table 1.5: Main procedural differences between the studies examined

The following two tables (tables 1.6 and 1.7) illustrate the difficulties involved in attempting to arrive at a cohesive estimate of both increase in detection rate and change in recall rate.

	Method of calculating improvement	Stated improvement (%)	MSSC (%)
Denton [32]	Double reporting - single reporting (R_1 or R_2)	1.5 - 4.2	Not calculable[3]
Thurfjell [33][36]	MSSC	15	15.2
Anttinen [34]	MSSC	8.9	8.8
Deans [35]	c/(a+b+c)	10.5 (12.3)[4]	6.4 (7.6)
Warren [39][5]	c/(a+b)	14	7.1
Anderson [37]	(b+c)/(a+b+c)	10.4	5.8
Ciatto [40]	MSSC	4.6	4.6

Table 1.6: Methods of calculating the improvement in cancer detection due to double reading

	Method of calculating change	Stated change (%)
Denton [32]	Recalls not mentioned	Not given
Thurfjell [33][36]	Recalls not mentioned	Not given
Anttinen [34]	Mean reduction of post discussion cases per reader	-45
Deans [35]	Reader 1 compared with double reading (figures excluding Glasgow)	+37 (4.2 raised to 6.6)
Warren [39]	First reader recall rate compared to post discussion recall rate	-39.1 (6.9 lowered to 4.2)
Anderson [37]	Specificity	-1.8
Ciatto [40]	MSSC	+15

Table 1.7: Methods of calculating the change in recall rate due to double reading (all figures are percentages)

[3] First and second readers not identified

[4] The estimates in parentheses are the figures for Scotland excluding Glasgow

[5] The results quoted in this article are R_1 versus post-discussion and cannot be strictly be compared with the other articles

The wide variety of reported gains makes an accurate evaluation of the possible improvement in detection rate difficult. It may be that some of this variation can be explained by the different methods of calculating the increase that are employed, which vary on almost a study by study basis. The picture is then further complicated by the fact that these evaluations are mostly done in clinical context by looking back over a period of time and analysing the results retrospectively. Hence, there is no experimental procedure followed, to control potential biases and errors.

1.3.2 A standardised measure

An alternative method of calculating the improvement due to the second reader is given in the article by Williams *et al* [31], based upon this work. This also examines the various methods of calculating the increase in cancer detection due to the second reader and proposes an alternative based on the marginal total of $a + b$ (the total number detected by R_1) and c, the actual number detected by R_2 that were missed by R_1.

This measure is the proportional increase in cancer detection rate due to the second reader; $c/(a+b)$. This measure is not influenced by the relative sizes of a and b (R_1+R_2+ and R_1+R_2− respectively), just the number of cancers discovered by R_1 and the additional cancers discovered by R_2. As such, it avoids the problems due to blinding/not blinding posed by the MSSC.

The alternative measure also has the property that it is easier to calculate both the point estimate (the actual increase) and the standard error of the increase. Whereas the MSSC has a complex formula for the standard error [33] the SE of $c/(a+b)$ is given, via a logarithmic transformation, as:

$$SE(log_e \tfrac{c}{a+b}) = \sqrt{\tfrac{1}{a+b} + \tfrac{1}{c}}$$

Example If we look back to tables 1.3 and 1.4 and use the same figures to calculate the proportional increase, then the improvement due to the second reader is 11.05% in both cases (the same as the estimate obtained with the Mean Second Screener Contribution with equal numbers in R_1+R_2- and R_1-R_2+), with a 95% CI of (6.9%, 17.7%). Thus it is irrelevant whether the second reader had access to the decisions of the first.

If we expand our example to cover the articles we have already examined, we get the results as seen in table 1.8.

	Method of calculating improvement	Stated improvement	MSSC	$c/a+b$	95% CI
Denton [32]	Double reporting - single reporting (R_1 or R_2)	1.5 - 4.2	Not calculable	Not calculable	Not calculable
Thurfjell [33][36]	MSSC	15.0	15.2	8.6	3.7, 19.7
Anttinen [34]	MSSC	8.9	8.8	6.25	2.3, 17.2
Deans [35]	c/(a+b+c)	10.5 (12.3)[6]	6.4 (7.6)	11.7 (14.1)	10.3, 13.3 (12.0, 16.6)
Warren [39]	c/(a+b)	14.0	7.1	13.75	9.6, 19.8
Anderson [37]	(b+c)/(a+b+c)	10.4	5.8	Not calculable	Not calculable
Ciatto [40]	MSSC	4.6	4.6	7.7	3.9, 15.3

Table 1.8: Calculating the improvement in double reading by proportional increase (all figures are percentages)

With the stated increases, we have a range of 4.6% to 15% for the estimated effect of double reading (from those with calculable proportional increases). With the proportional increase, the range is much smaller at 6.25% to 11.7% (with the exclusion of the Warren paper [39] from these calculations, as their comparison was between the first radiologist and the post-discussion result, rather than between first reader and second reader). The confidence intervals for these improvements are rather wide, reflecting the relatively small numbers of patients with cancer. A weighted average of the improvements due to the second reader is 11.4% with a 95% confidence interval of (10.0%, 12.9%). This figure is highly influenced by the results of the Deans [35] paper, which is based on 2473 cancers detected - nearly 10 times greater than any other study.

This measure and the MSSC have been presented in terms of determining the increase in cancer detection, but they can also apply in precisely the same way to the determination of increase in recall rate (table 1.9).

[6] The estimates in parentheses are the figures for Scotland excluding Glasgow

	Method of calculating change	Stated change	$c/a+b$	95% CI
Denton [32]	Recalls not mentioned	Not given	Not calculable	Not calculable
Thurfjell [33][36]	Recalls not mentioned	Not given	Not calculable	Not calculable
Anttinen [34]	Mean reduction of post discussion cases per reader	-45	46 (45.3)[7]	39.8, 53.2 (36.5, 56.3)
Deans [35]	Reader 1 compared with double reading (figures excluding Glasgow)	+37 (4.2 raised to 6.6)	37.3	36.1, 38.6
Warren [39]	First reader recall rate compared to post discussion recall rate	-39.1 (6.9 lowered to 4.2)	43.9 (16.4)[8]	40.8, 47.3 (14.7, 18.3)
Anderson [37]	Specificity	-1.8	Not calculable	Not calculable
Ciatto [40]	MSSC	+15	18.7	15.4, 22.8

Table 1.9: Methods of calculating the change in recall rate due to double reading

1.3.3 Two views

In addition to the usual mediolateral obliques, which take an image diagonally across the chest from the shoulder to the stomach, the cranio-caudal (CC) view was introduced for women at their first screen in 1995 following the results of the UKCCCR randomised trial [41]. The CC takes an image through the breast from top to bottom, thus adding another dimension for the radiologists to examine.

At the moment, CCs are only offered to prevalent screenings at the majority of screening centres, substituting for the additional information usually gained by comparing the current mammogram to the previous one. Trial results have been good, with a significant improvement in cancer detection shown, particularly in the detection of small invasive cancers (<15mm). Blanks *et al* [42] reported a 42% improvement in the detection of these cancers, whose early detection can lead to an improved prognosis for the patient. In 1998, Blanks *et al* [43] surveyed the 87 screening programmes in England and Wales on their reading protocols and estimated that double reading with arbitration and two views improved cancer detection by 73% (95% CI 40% to 113%) over single reader single view.

The UKCCCR randomised controlled trial of one and two view mammography [41] established that two views detected 24% more cancers than one view, with a 15% drop in recall rate. Cost analysis showed that, although the second view was more expensive

[7] The figures in parentheses are post-discussion results

[8] The figures in parentheses are the R_1 versus actual decision

per examination (£26.46 compared to £22.00), the cost per cancer detected was similar.

These encouraging results have induced trials of two view mammography at incident screenings. Results from this have also been encouraging, although the improvement has been on a smaller scale, which was not unexpected [44]. Unfortunately, the limiting factor in implementing two views across the board is, as usual, cost. Although it takes very little extra radiologists' time to add CCs to the regular screening information, the cost of the high quality film and processing is prohibitive in all but a few centres [36, 45, 46, 47, 48, 49].

1.3.4 Radiographers reading

Roughly one and a half million women pass through the NHS Breast Screening Programme *every* year, with about 5% being recalled for further assessment. With an average number of films per woman per visit at five (obliques, CCs for the first visit, previous and the occasional woman who requires two films per breast), this is a considerable amount of work for a radiologist, and set to rise with the introduction of CCs for incident round screening. Add to this a shortage of radiologists willing to go into mammography that is increasing every year, and the situation becomes somewhat fraught. And so, some clinics have turned to radiographers to take up the shortfall.

Pauli *et al* [50] reported that an experiment conducted at the Jarvis Breast Screening Diagnostic and Training Centre with radiologist/radiographer pairings improved sensitivity on a similar level to other reported double reading experiments (6.4%), although radiographer specificity (the defining as normal a non-cancer) was significantly lower than that of the radiologists. However, this decreased specificity could be the result of insufficient experience – Warren's paper [39] discusses the improvement over time of radiologists double reading, over a period of nearly 4 years (33,734 women) in one centre. As that improves from 82% sensitivity with a 3.8% recall rate to 97% sensitivity and a 3.0% recall rate, it could be suggested that only insufficient time prevented the radiographers from improving to a similar level of expertise. Since the Jarvis experiment was conducted with 17,202 women over three centres, the level of exposure was not similar to that in the Warren paper.

Currently, Jarvis is one of the few training centres in the UK that regularly train radiographers to read mammograms. Part of this is reluctance on the part of the radiographers; the unwillingness to be responsible for decisions of this magnitude. The decision to pass a woman as clear is a hard one, particularly if there is even the slightest amount of doubt. This is another reason for the poor specificity that radiographers exhibit – they are not sufficiently confident in their own judgement to pass as normal all those that they should.

Bassett *et al* [51] report an experiment involving eight 'radiologic technologists' (radiographers) and seven radiologists, where the 'radiologic technologists' underwent an eight hour training session between two reading tests. It was shown that the sensitivity and specificity improved between the two reading sessions, indicating that formalised training could improve the 'radiologic technologists' ability to interpret mammograms such that they could be used to increase the number of breast cancers detected at screening.

1.3.5 Reducing the interval

Although not something that screening centres can control, one highly regarded method of improving cancer detection is to reduce the screening interval from three years to two. Woodman [23] and his colleagues [52] investigated this question by examining the number of interval cancers that appeared in each of the three years after a screening visit, and found the proportion that manifested in the third year to approach that which would have been expected in the absence of screening. This, he attests, implies that the screening interval is too long. This was also the conclusion of Sylvester *et al* [53].

A recent article by Shapiro *et al* [13] examined the screening guidelines in 22 IBSN

(International Breast Screening Network) and European Network countries, as surveyed in 1995. This showed that the UK, along with Uruguay, had the highest annual breast cancer death rate with 27.7 per 100,000 (age adjusted). However, Uruguay has no organised national screening programme, with less than 25% of the target population covered by the programmes that are in place. What appears to distinguish the UK national programme from the others in the survey is the screening interval. While the majority are inviting their screening population for examination every two years, with some countries screening annually, the UK is unique in screening once every three years.

A computer simulation by Boer *et al* [18] examining the effect of reducing the interval to two years, suggested that a shorter screening interval would result in a reduction in mortality from breast cancer of 15.3%, where the current programme achieves 12.8%, a relative improvement of 20%. Another simulation [54] suggested that a screening interval of two years would reduce the spread of cancer to other parts of the body (distant metastases) by 22%, reducing it to one year would result in a 51% reduction and an interval of six months would give an 80% reduction.

1.3.6 Conclusions

It is quite clear from the evidence presented in this chapter that there is much scope for improvement in the cancer detection rate, without a corresponding increase in the number of false recalls. As we will see in successive chapters, technology may offer one way of improving the number of cancers detected, thus potentially reducing the number of people who die from breast disease every year.

Chapter 2

21^{st} Century mammography

2.1 Aims and Objectives

Technology has made many advances over the last century but, despite this, mammography has changed little since the days of Albert Salomon, the German surgeon who first demonstrated the efficacy of x-rays in highlighting signs of breast cancer in 1913 [55]. Improvements in imaging detail, and in the reduction of exposure to radiation have occurred, and, indeed, made the process safer and less prone to error, but the system of mammographic reading is still, at the beginning of the 21^{st} Century, reliant on human observation and detection. This human interpretation has been shown to be highly variable [56, 57, 58, 59], with Beam *et al* [59] suggesting that certain radiologist pairings can even be detrimental to the true positive/false positive rates.

This chapter will briefly touch upon the methods currently under investigation for technologically improving detection using a variety of computer-aided techniques, before introducing the PROMAM project, its aims and experimental methods. Methods of designing and analysing a trial of this type will be discussed, as will the possible ramifications of introducing a prompting system into clinical context, before a design is proposed and introduced as the basis of future experiments.

2.2 Computer-aided systems

The rapid progress of technology at the end of the 20$^{\text{th}}$ century has caused the expansion of breast cancer detection into areas that rely less on human perception, removing the potential for fallibility that is present whenever humans are involved in a process.

Digital mammography is a major part of this expansion, with many varied algorithms and techniques being applied to the problem; neural networks [60, 61], adaptive algorithms [62], discriminant analysis [63] and so on. Most involve the digitisation of mammographic films, although some are experimenting with direct digitisation [64], and producing markers or prompts to suspicious areas on either a paper or directly onto a video display unit (VDU).

Current algorithms rely on the human reader to keep the specificity to a satisfactory level, as they are not, as yet, discriminating enough to discard suspicious features that may be benign [65]. Sensitivity is good for many of these algorithms, around 90% for Fuji Computed Radiology [66], although the false positive rate is high at 1.35 FP per image for tumours and 0.4 FP per image for microcalcifications. Observer performance studies have shown that computer-aided detection (CAD) systems can improve reader performance significantly [66, 67, 68, 69]. However, no system will be accepted into the mass screening system until it can cope with the daily load of a Breast Unit, up to 100 women per day [15] .

2.3 PROMAM

In 1994, the Royal Observatory, Edinburgh (ROE), in conjunction with the University of Edinburgh (Departments of Computer Science and Public Health Sciences), proposed a research project into the use of the SuperCOSMOS high-resolution scanner in mammography, in line with the then Government's White Paper into the transfer of technology from the academic world into mainstream use.

SuperCOSMOS [70] is an advanced photographic plate digitising machine, created at the ROE to scan and digitise glass photographic plates and films of the night sky. These images are digitised at a resolution of ten microns, *smaller than the human eye can see.* As such, it was seen as an ideal candidate to aid in the mass breast screening programme.

Mammography is a highly demanding task, which can only be conducted by experienced and trained personnel; radiologists and, as has lately been suggested in response to the shortage of radiologists specialising in mammography, radiographers [50]. Breast screening is widely accepted as the most effective method of detecting early signs of breast cancer, but performance could still be improved in the following areas:

- a reduction in the number of false negative interval cancers
- an improvement in the detection of small cancers

Some clinics utilise a second reader to address the problems mentioned above. The improvement in detection rate that this generates varies greatly, but appears to be approximately 11% [31] (see page 16). However, this ties up a second radiologist's time that could be better put to use in the assessment clinics (where recalled women are examined more carefully for signs of cancer).

PROMAM (PROmpting for MAMmography) aims to detect the cancers that are missed by the radiologists, thereby reducing errors. It is hoped that the prompting system could take the place of the first reader, thus potentially reducing the variability detected between pairs of readers while maintaining or exceeding the double reading detection rate.

The PROMAM system consists of a digitising scanner, a DEC Alpha workstation, image-processing algorithms and a paper prompt system.

2.3.1 The scanner

Despite the initial desire to use the SuperCOSMOS technology in the project, it was found that the extremely high resolution was not required, and was, in fact, detrimental to the analysis of the image, as very small features created too much 'noise' for the algorithms. Hence, the project turned to a DBA Imageclear film digitiser, with a 42 micron resolution[1]. This, it was claimed, took 20 seconds to scan a standard 8" × 10" film.

2.3.2 The workstation

The image processing system was a customised DEC Alpha, with extra memory and additional processing power. The aim was to store images from the scanner and run the feature detection algorithms at a rate of one film every two minutes. It was estimated that this level of throughput was required to enable a single system to support most UK breast screening clinic workloads.

2.3.3 The algorithms

Three algorithms were developed for cancer detection – ill-defined lesions, microcalcification clusters and stellate lesions. A pectoral muscle mask was also developed as part of the pre-processing. These algorithms were produced by other members of the PROMAM team, and I shall only give a brief description in order to aid understanding of the experiments that were later conducted.

The pectoral muscle mask The purpose of masking the pectoral muscle (the brighter triangular part at the back of the image on page 190) was to exclude it from the noise estimation stage of the analysis of the image. The scaling code performed more accurately on dense tissue areas when this region was excluded. Most of the work in this area was done by Neville Ramsay and later Ally Hume.

[1] from PROMAM internal documentation, commercial development publicity

The ill-defined lesion algorithm An ill-defined lesion is an area of increased density with a fuzzy edge. The algorithm is designed to prompt for suspicious features between 5 and 34 mm in diameter. The mass (opacity) may constitute the nidus of a stellate feature, be associated with tentacles or spicules, or may have no associated structure. These features form the majority of cancer types.

The image is analysed in three stages:

1. Multi-resolution analysis – The image is separated into a series of sub-images, each encompassing the whole mammogram, but only containing structures that fall into a specified range of sizes, using a maximum entropy technique. This removes the clutter of features not in the scale of interest.
2. Segmentation – The scale images are segmented into candidate regions by growing areas of similar brightness. A number of parameters are calculated at this stage, reducing the huge volume of data. The parameters include area, brightness, shape, contrast, texture values and isolation.
3. Classification – This is a supervised learning process to achieve clustering of the malignant and normal regions of the parameter space.

This algorithm was developed by Lance Miller, Steven Heddle and Ally Hume [71].

The microcalcification algorithm Individual microcalcification flecks are not on their own indicative of cancer. However, several microcalcifications are highly suspicious, especially when in the presence of an ill-defined lesion. A cluster is labelled suspicious by the algorithm if a certain number of calcifications occur within a particular distance from one another. Several permutations of these (and other) parameters have been assessed. Unfortunately, every permutation is subject to errors in detection, either lack of sensitivity or excess of false positives. One of the main causes of false positives is vascular calcification, small calcium deposits along veins in the breast tissue. The algorithm was the work of Ally Hume [72], based on work done by Nico

Karssemeijer [73].

The stellate lesion algorithm This algorithm was developed at Manchester University, but unfortunately never met our criteria for inclusion in any of the system trials. It was designed to examine the digitised image for straight structures that would indicate the radials of the stellate lesion [74]. Stellate lesions occur with and without central nidii.

2.3.4 The paper prompt

Once the algorithms have individually produced their suspicious regions, they are superimposed over a low-resolution image of the film. This is then printed onto a sheet of A4 paper, along with the subject number and the number of suspicious features each algorithm has produced. This *paper prompt* can then be used to locate the congruent region on the film. An example is included in Appendix D.

2.4 PROMAM in the clinic

PROMAM is not intended to replace the skill and experience of a radiologist, merely to enhance. It is a tool, another source of information to be given relevant weight as with other sources of information in making a suspicious/normal classification. It is not possible to design a system that can detect all abnormalities with sufficient sensitivity and specificity to replace a human reader.

2.4.1 Types of cancers

Not all the work undertaken under the auspices of PROMAM was conducted in conjunction with the algorithms. Much work was required to help us set the various parameters that would give the team targets for which to aim; in other words, the 'gold standard' that would indicate a system worthy of a trial in a clinical context. This will be discussed in this chapter, while work that was done with functional, yet unfinished, algorithms will be covered in the next chapter.

The first major piece of work was an investigation into the relative make-up of the types of lesions that form the cancers that are detected in routine screening, in order to be able to predict the overall accuracy of the system, composed of the combined algorithms. It was known, from discussions with the radiologists at Ardmillan House, the South East Scotland Breast Screening Centre (SESBSC), that the various types of cancers are not present in equal amounts in a population of malignancies. However, it was not known what the relative frequencies were, thus necessitating an investigation into this question.

The tables below are from data drawn from Ardmillan House's archives by Pat Dixon, the PROMAM radiographer. The cancer definitions are those used by Ardmillan House and may not correspond exactly with those at another clinic. 'Microcalcification' refers to microcalcification with no other features present.

	1990	1991	1992	1993	1994	1995	Total
Stellate	7	11	10	12	10	5	55
	13%	15%	9%	9%	9%	11%	10%
Spiculated	7	4	24	22	24	9	90
	13%	5%	21%	17%	22%	19%	17%
Tentacled	5	8	15	25	18	9	80
	9%	11%	13%	20%	16%	19%	15%
Irregular	10	17	22	29	12	4	94
	18%	23%	19%	23%	11%	9%	18%
Smooth/	4	6	5	5	8	5	33
Lobular	7%	8%	4%	4%	7%	11%	6%
Vague/	5	9	18	12	8	4	56
Fuzzy	9%	12%	16%	9%	7%	9%	11%
Asymmetry	0	0	3	4	4	0	11
	0%	0%	3%	3%	4%	0%	2%
Dist.	1	2	6	6	4	3	22
Archit.	2%	3%	5%	5%	4%	6%	4%
Micro-	17	16	13	13	23	8	90
calcification	30%	22%	11%	10%	21%	17%	17%
Total	56	73	116	128	111	47	531

Table 2.1: Frequencies and percentages of each lesion by year. Percentages are only calculated for proportion of lesions within a year

Stellate, spiculated and tentacled are types of lesion with radial structures emanating from a medial point which may or may not have a central nidus. Irregular, smooth/lobular and vague/fuzzy are also lesions, but without the radial structures. They are all ill-defined lesions, and highly suspicious features when detected within a breast. Asymmetry refers to the asymmetry between the two breast images; should the two mammograms be sufficiently different in appearance, this is also indicative of a potential cancer. Dist. Archit. is the abbreviation for distorted architecture, where structures within the breast tissue appear to be distorted or twisted. The final category, microcalcification, refers to tiny specks of calcium that have formed within the breast tissue. Although not necessarily an indication of cancer in themselves, when seen in clusters, or in conjunction another suspicious feature, they can be an indication of a problem. These features are described in more technical detail in the glossary (Appendix A).

Using the conventional 5% level of significance, the frequencies within table 2.1 are not significantly different across the years. However, the χ^2 value is 54.71, with 40 degrees of freedom giving a p-value of 0.06. Hence there may be some evidence to suggest that lesion and year are not independent. This may be due to changes in emphasis, definition of lesion type, film quality, experience, or many other factors that may affect a screening service. Early detection of certain lesion types would also remove them from later detection and, hence, inclusion in later calculations.

Microcalcification is often associated with other cancer types. The following chart (figure 2.1), shows the relative frequencies of malignancies with and without associated microcalcification, with the following table (table 2.2) giving the relative proportions within lesion type.

Microcalcification with a lesion is often a sign that what may appear benign is actually cancerous or pre-cancerous. Thus if algorithms are able to detect the microcalcification clusters this is an additional cue to the cancer, especially if the mass algorithms fail to

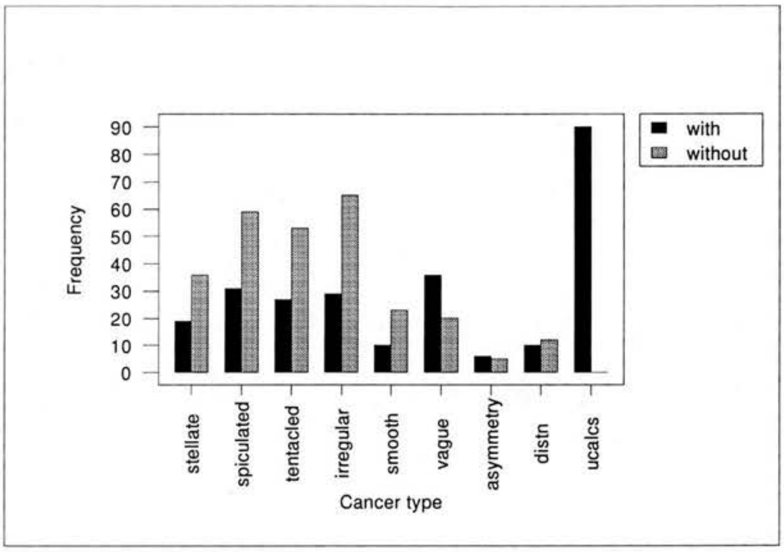


Figure 2.1: Frequency of cancer types with and without microcalcification

	With microcalcification	Total number of lesions
Stellate	35%	55
Spiculated	35%	90
Tentacled	34%	80
Irregular	31%	94
Smooth/Lobular	30%	33
Vague/Fuzzy	64%	56
Asymmetry	55%	11
Dist. Archit.	45%	22
Microcalcification	100%	90
Total	49%	531

Table 2.2: Proportions of lesions with associated microcalcification

detect the lesion.

2.4.2 Interval cancers

Interval cancers are cancers which appear during the interval between screening rounds, typically three years in the UK. However, some of these cancers may not be *true* interval cancers (cancers that have entered the pre-clinical detectable phase between screening rounds with no evidence visible on the previous mammograms). A significant proportion of cancers detected outwith screening are false negatives (FN); that is, there is a visible

	Sample size	True interval rate	False negative rate	Occult	Other
Heddle [75]	91	65%[2]	35%	0%	0%
Duncan[3][76]	50	46%	34%	20%	0%
Simpson [77]	167	46%	26%	11%	16%
Asbury [52]	130	66%	31%	3%	0%
Burrell [78]	90	57%	22%	8%	13%
Jones[4][79]	133	77%	23%	0%	0%
Sylvester [80]	134	50%	16%	9%	25%

Table 2.3: Studies of interval cancers

sign of cancer on the previous mammogram. This failure to detect the cancer may be due to numerous factors; fatigue, having a locational 'blind-spot', distraction, amongst others. If a system could be devised to detect these FN cancers, it would have the potential to substantially increase the sensitivity of the screening programme.

Work has been done on analysing these cancers by other members of the PROMAM team in preparation for testing the algorithms [75]. Their definition of 'interval cancer' included *delayed diagnosis* (the woman was seen at an assessment clinic, the decision was made to recall a year later, whereupon she was diagnosed), *missed diagnosis* (the woman was seen at an assessment clinic but was discharged without further action), *occult* (the cancer was not visible on the mammogram) and *true interval.* As it would not have been possible to affect the outcome of these cancers with a prompting system - since the suspicious feature was either noticed and subsequently returned to the population untreated, or not visible at all - they were combined into one category, which composed 65% of the 259 'interval cancers' examined. The remaining 35% were divided into two types of false negative; FN(1) and FN(2). FN(1) was defined as a cancer that was visible on the mammogram in retrospect only (6.5%), and FN(2) was defined as a cancer that was readily visible on the mammogram (29%).

Similar studies in both the UK and elsewhere have yielded figures much like those noted by the PROMAM team (see table 2.3). Despite the variety in actual FN rate, all studies agree that the number of cancers missed is too high.

[2] Includes delayed diagnosis, missed diagnosis, occult and true interval

[3] By group consensus

[4] From cancers detected at the first incident screen (second screen)

Very little can be done about the true interval cancers, other than reduce the screening interval (see page 19), and occluded cancers can be dealt with by adding a second view (CC). However, there is often no reason why FN cancers should not be detected.

2.4.3 CAD and FN interval cancers

Great effort is currently being put into creating algorithms that can detect early signs of cancer in digitised mammograms. Entire conferences are dedicated to the various aspects of this aim; for example, the International Workshops on Digital Mammography that are held every two years.

One of a computer-aided detection system's main strengths are that it never suffers from fatigue, can never be distracted, and has no spatial 'blind-spots'. Using the interval cancers mentioned above [75], a fully working version of the algorithm set detected 44% of the cancers defined as false negatives (FN(1) and FN(2)). As the system is only designed to be an attention cue, it is possible that the radiologist could still classify the cancer as normal or benign, but it would have at least been seen. PROMAM was not designed to overrule the radiologist [81], and the final decision to recall/not recall will always remain with the reader [82].

2.5 Designing a trial for the system

The UK Breast Screening Programme has a cancer detection rate of approximately six per 1000 women, and a recall rate of around 5.4% (54 per 1000 women) [17]. Evidence from interval cancer studies suggest that the present system may fail to detect around 25% of those cancers which could, in principle, be detected on a mammogram (see table 2.3). As described earlier, PROMAM intends to target these cancers especially.

Once the algorithms were performing to acceptable levels, it was decided that we would need to test the system under screening conditions before PROMAM could be presented as a worthwhile system. Given our prior investigation into double reading improvement rates (see section 1.3), we selected a relative improvement of 6% as the minimum improvement we would wish to detect in order to justify PROMAM as a potential replacement for the second reader.

2.5.1 Normal screening metrics and their disadvantages

When introducing any new technique, it is necessary to compare it to the method currently in practice, in order to ensure that the new method is actually an improvement or, at the very least, is not worse than current practice. Unfortunately, the usual metrics for these comparisons do not easily lend themselves to breast cancer screening.

The three most common metrics are *sensitivity*, *specificity* and *the kappa statistic* [83].

- Sensitivity – a measure of how good a method of detection is at detecting the condition.
- Specificity – a measure of how good the method is at excluding those without the condition.
- The kappa statistic – this measures the amount of agreement between the new and existing methods over that which would be expected by chance, possibly with a weighting function, when the data are paired categorical ordered responses.

2.5.2 Sensitivity

Sensitivity is an absolute measure of a system's ability to detect a 'positive', be it a cancer or any other outcome defined as a 'hit'. In screening, it is impossible to know whether all the cancers have been detected without rigorous follow up of all women screened. Many centres now have interval cancer sessions (re-examining interval cancers for signs that might have been visible on the mammogram) as part of their quality assessment. However, it is still possible for cancers to be missed by both readers, and hence sensitivity is often measured against the combined total of cancers detected when that radiologist was one of the pair reading. In other words, the sensitivity of a radiologist is given as the number detected by him/her, expressed as a percentage of the maximum detected to which they were exposed - co-positivity.

2.5.3 Specificity

Similarly, specificity is also difficult to quantify exactly, for much the same reason. If a cancer is missed and the woman is passed as normal, then the calculation of specificity would include that case as a normal passed as normal. The true specificity would be difficult to establish, although the relative specificity will be close to the true specificity as the data are dominated by normals. Again, specificity is usually calculated as a percentage of the 'normal' cases to which the radiologist was exposed - co-negativity.

2.5.4 Kappa

Kappa (κ) measures the agreement between observers (in this case, our two readers) of subjects (the mammograms) on a categorical scale (recall, technical recall or not recall). Thus, for a randomly selected mammogram, π_{ij} is the probability that the first observer will place the mammogram in category i and that the second will place the mammogram in category j. Then $\Pi_o = \Sigma\pi_{ii}$ is the probability that the observers agree on the category, and $\Pi_e = \Sigma\pi_{i+}\pi_{+i}$ is the probability of agreement if the observers' ratings are independent.

So, if Π_o is the observed probability that observer 1 and observer 2 agree (i.e. $= \Sigma n_{ii}/n$, where n_{ii} is the number that both observers placed in category i and n is the total number of observations), and Π_e is the expected probability, based on the marginal totals (i.e. $= (\Sigma n_{i+}n_{+i})/n^2$, where n_{i+} is the number that observer 1 placed into category i and n_{+i} is the number that observer 2 placed into category i), then kappa is given by:

$$\kappa = \frac{\Pi_o - \Pi_e}{1 - \Pi_e}$$

and

$$\kappa = \frac{\Sigma\pi_{ii} - \Sigma\pi_{i+}\pi_{+i}}{1 - \Sigma\pi_{i+}\pi_{+i}}$$

where $\pi_{ii} = n_{ii}/\mathrm{n}$ etc.

$\kappa = 0$ if the agreement equals that of chance, and $\kappa = 1$ if there is perfect agreement. For the values in between, Landis and Koch [83] suggest the following:

Value of κ	Strength of agreement
< 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

The *weighted kappa* value (κ_w) uses weights to describe the closeness of agreement with $0 \leq w_{ij} \leq 1$ with all $w_{ij} = w_{ji}$ and $w_{ii} = 1$.

Hence,

$$\kappa_w = \frac{\Sigma\Sigma w_{ij}\pi_{ij} - \Sigma\Sigma w_{ij}\pi_{i+}\pi_{+j}}{1 - \Sigma\Sigma w_{ij}\pi_{i+}\pi_{+j}}$$

A common choice of weighting is, $w_{ij} = 1 - \frac{|i-j|}{g-1}$, where g is the number of categories. The divisions above are often applied to κ_w, although they are not strictly intended for the weighted kappa.

Unfortunately, kappa is influenced by the prevalence of the disease, approaching zero as the prevalence approaches zero or one. As the appearance of breast cancer in a screening population is approximately 0.6%, any kappa values derived would be subject to some suspicion.

2.5.5 Designs for assessing the efficacy of PROMAM

In order to accurately measure PROMAM's efficacy, it is necessary to directly compare the system (radiologist with prompts) with a method currently in practice. Thus, our design of choice was a randomised controlled trial, with each woman's mammograms being seen by either the system in clinic or by one radiologist supported by the prompting system. Ideally, the system in the clinic would be double reading, as practised as standard in Scotland, and voluntarily in some centres in England and Wales. However, this design was calculated to require approximately 750,000 women in each group (using an underlying cancer detection rate of 0.6% and a 6% relative detectable improvement), necessitating a total throughput of 1,500,000 women. Given that this is more than are screened by the entire NHSBSP in one year [17], this design was abandoned as unfeasible.

Next, we examined the possibility of matching the data by having each mammogram read by both systems; the dual radiologists (the standard to which the PROMAM system aspired) and the PROMAM assisted single reader. In this case, we were attempting to show that the two methods were clinically equivalent. Unfortunately, clinical problems limited this option in practice. In many cases, the 'bottleneck' in the screening programme is the radiologist performing the second read. Thus most of the clinics approached were reluctant to add a third reader to the sequence (two blinded double readers plus one reader with the prompts produced by the PROMAM system). And so, this design also had to be abandoned.

With our options limited by sample size and clinical procedures, the following design formed the basis of all further experiments with the system in a clinical context:

Each woman would have her mammograms read by both a single unprompted reader and a single prompted reader. The second reader would be blinded to the decision of the first, and readers would be randomly assigned to be prompted. As far as possible, radiologists would also be randomised to be first or second reader. Cases would be assigned to radiologists in blocks, as is currently the practice. Recalls would be made on the 'worst case' basis; i.e. if either radiologist deemed the mammogram suspicious and asked for recall, then that woman would be recalled.

2.5.6 Variables of importance

Obviously, the most important variables in any trial of this type are the relative cancer detection rates and the recall rates. However, there are many other variables that may be of interest, including subject information (such as age, whether this is a prevalent or incident screening, previous history of cancer etc.), information from the various algorithms (false prompt rate, true prompt rate), economic data, and subjective data from the radiologists, such as whether they believe the system to be useful in a screening clinic.

2.5.7 Evaluation of the results

As mentioned above, the principal endpoints of the study are the detection of cancers and the unnecessary recall of women. Both of these variables are binary, and since each woman's films are seen by both a prompted and unprompted reader, we have within-woman comparisons. Hence, the data can be summarised in the form of the table below.

		Clinic system	
		Positive	Negative
PROMAM system	Positive	a	b
	Negative	c	d

The a and d entries denote the $(a + d)$ cases where both methods produce the same outcome. As such, they contribute no useful information on the differences between the methods. This information comes from the b and c cases, where the two methods produce different actions. If the recall/cancer rates are the same under both methods, then b and c should differ from each other only by chance.

The appropriate statistic in such situations is McNemar's Test, a paired-binary test [84].

The Null Hypothesis for such a test is that there is no difference between the two factors under consideration (i.e. prompted and unprompted) with respect to cancer detection and recall rates.

McNemar's test compares the b and c elements of the table, by making the assumption that if the NH is correct, then b (or c) is a binomial variable where $b \sim B(b+c, 0.5)$. Where $b+c$ is small, and if $b < c$, say, an exact two-sided test gives the p-value as $[P(X \leq b) + P(X \geq c)]$, where $X \sim B(b+c, 0.5)$. Where $b+c$ is not small, we may use an approximate test, under the Central Limit Theorem. Hence, $b \sim N\left(\frac{b+c}{2}, \frac{b+c}{4}\right)$, under NH. With continuity correction, this gives the test:

$$z = \frac{|b-c|-1}{\sqrt{b+c}}$$

The point estimate for the difference between the two recall/cancer rates is given by $\frac{b-c}{n}$, with the estimated standard error given as $\sqrt{\frac{1}{n}\left\{\frac{b+c}{n} - \left(\frac{b-c}{n}\right)^2\right\}}$.

2.5.8 Calculation of sample size

As the design chosen required that we analyse the data with McNemar's Test for paired binary data, certain approximations had to be made concerning the relative accuracies of the two detection methods. Additionally, the agreement between the two methods is specified as part of the calculations. Since the usual practice is a reading without prompts, this was defined as the standard, and the agreement was calculated accordingly (number of cancers discovered by both methods divided by the number of cancers discovered by the non-prompted method).

		Unprompted		
		Correct	Incorrect	
Prompted	Correct	θ_{11}	θ_{10}	θ_2
	Incorrect	θ_{01}	θ_{00}	$1-\theta_2$
		θ_1	$1-\theta_1$	

Table 2.4: Calculating sample size

Table 2.4, shows the proportions in each cell; e.g. θ_1 is the proportion of correct responses given by the unprompted reader. Hence, θ_{11} is the proportion of the cancers that the unprompted reader found that the prompted reader also found., i.e. $\theta_{11} = \text{agreement} \times \theta_1$ [84].

The sample size required for an experiment of this type can be calculated by:

$$n = \frac{(\theta_{01} + \theta_{10}) - (\theta_{01} - \theta_{10})^2}{(\theta_{01} + \theta_{10})^2} \times f(\alpha, \beta)$$

where $f(\alpha, \beta) = \{\Phi^{-1}(\frac{\alpha}{2}) + \Phi^{-1}(\beta)\}^2$ the magnitude of which depends on the size of the power $(1 - \beta)$ and the significance level α .

Three tables in the appendix (Appendix C) show the sample size requirements for three different cancer detection rates (0.5%, 0.6% and 0.7%), with 80% power to detect statistically significant differences at the 5% level. The estimates of PROMAM's improvement range from -5% (5% worse than a single reader) to +10% (10% better than a single reader). Although the national average is approximately 6 cancers detected per 1000 women, for prevalent screens the cancer detection rate is 7.8 with the incident screening rate at 3.7 per 1000 [15]. For example, using the estimates mentioned earlier (0.6% underlying rate, 6% estimated improvement), with an agreement of 90%, we would need to recruit 95,085 women into the trial. However, as the agreement, estimated improvement and underlying cancer detection rate increase, the sample size decreases. Therefore, should any of these values be underestimates, the required sample size will decrease.

2.5.9 The expected impact of PROMAM

The potential impact of PROMAM on the functioning of a screening service was expected to come from three sources:

- blind double reading
- use of prompting information

- disruption of normal reading practices (from regimented reading, lack of feedback, physical presence of scanner and associated equipment)

Moving to blind double reading

Although most of the centres approached to participate in these trials currently run non-blinded double reading (where the second reader has access to the first reader's decision), it appears that the second reader, in most cases, makes their decision without reference to this information. So, in practice, this is a non-rigorous form of blinding, resulting in informal, independent decision making. In order to conduct these experiments in a statistically valid fashion, we merely wish to make the process more stringent by denying the second reader the opportunity (whether they would have used it or not) of examining the first reader's decision. Since we are only formalising an existing practice (albeit a loose one), we expect blind double reading to have little impact on the recall rates.

Use of prompting information

The prompting system was not designed to replace the radiologist, merely to enhance the performance. The radiologist will have the final say in any decision concerning recall, whether a suspicious region is prompted or not.

Disruption of reading practices

Radiologists would be randomly assigned to first/second reader and to prompted/not prompted. There is no provision in the design to allow a radiologist to be consistently first, or to always be unprompted. Any indication that this is happening would result in a request that the numbers be re-balanced.

Recalls are made on a 'worst opinion' scenario. In other words, if at least one radiologist decides for recall, then that case is recalled.

We anticipate that there will be a learning curve as radiologists settle into the new system, and this will be taken into account when compiling data for comparison.

2.6 Conclusions

Direct digitisation would appear to be the aim for the future. PROMAM, with its database set up and ability to store and magnify digitised images, would have been ideally suited to take advantage of this technological leap forward.

With the introduction of an extended period of invitation to breast screening, the number of women likely to present themselves for screening is set to increase. In addition, multiple campaigns by age-related charities have raised the profile of self-referral for the over 65s. Hence, the load on radiologists is also going to increase. Such a system as PROMAM, would lower the load of the radiologists by taking the place of the first reader.

Chapter 3

Subjective reaction to prompting experiment

3.1 Aims and Objectives

In this chapter, we shall look at an experiment that was conducted with the cooperation of the radiologists of the Glasgow Screening Units. This experiment involved the first instance of the algorithms being used in a clinical setting, analysing four days' worth of output from the Edinburgh Screening Centre. The readers were asked to complete a series of questions after each session, to gauge their opinions on the three accuracy settings of the algorithms; low, medium and high.

The aim of this experiment was, initially, to investigate the radiologists' views on the three levels of prompting, in order to set the algorithms for the next big test of the system, the pre-clinical experiment (see next chapter). However, fears had been raised by the Directors of Screening Units who had been approached to participate in the multi-centre trial, about whether the recall rate would rise when the prompting system was in use. Hence, information on recall rate, time taken to complete a session and behaviour with the prompts were all to be analysed, with a view to reassuring participants that prompting did not significantly alter behaviour.

3.2 Introduction and background

Before PROMAM was to be considered ready for testing in a clinical setting (i.e. with actual throughput cases under normal screening conditions), the algorithms had to pass a series of target performance criteria. In the main, this involved maximising the True Positive rate (TP rate) while keeping the False Positive rate (FP rate) as low as possible. This is usually done by plotting various operating points on an ROC (receiver operating characteristics) curve [85] [86]. Normally, the point where the TP rate is maximised for a minimised FP rate is the point on the curve closest to the top left-hand corner (figure 3.1). However, should this lie beyond the upper limit of the acceptable FP rate, or if the TP rate was lower than the minimum acceptable TP rate, then an alternative point along the line could be chosen.

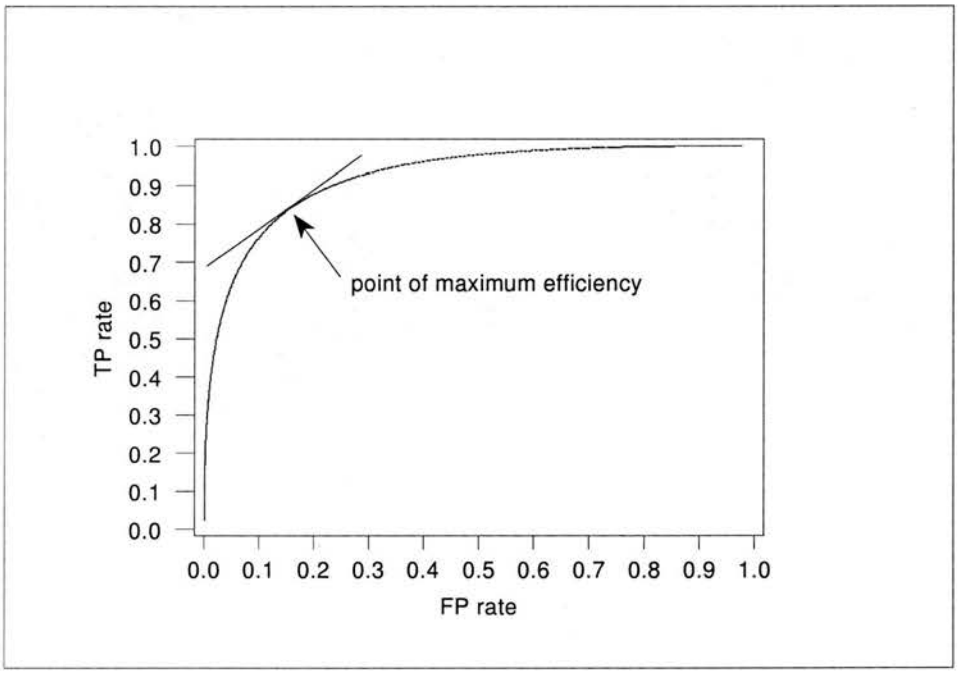


Figure 3.1: Theoretical ROC curve

Although the algorithm developers knew what their algorithms were capable of, there was no clear idea about the levels of true and false prompts that a radiologist would tolerate when using the system for mass screening. Dr Ian Hutt, of Manchester University

[87] had done some work along these lines, with an experiment involving 30 experienced radiologists in 11 screening centres viewing films on equipment that would normally be used for mammograms. Prompts were simulated, since no algorithms currently existed to detect all the types of malignancies in the experimental set, and also to rigidly control the TP (which was fixed at 90% over all conditions) and FP rates. The experimental mammogram set consisted of 100 pairs of mammograms (one per breast from each woman) taken during routine screening. 20 of these film pairs contained a malignancy on one breast film only. Each radiologist was shown the set of 100 film pairs, 50 of which were 'prompted' and 50 'unprompted', with the 20 malignancies distributed equally between the prompted and unprompted sets. The prompted sets had FP rates determined by the ratio of the FP rate to the TP rate. Half of the radiologists in each condition (i.e. 1:1 FP: TP rate, etc., see table 3.1) saw the prompted films first, the other half saw the unprompted films first. Radiologists recorded their responses to the films on a six-point scale (0=Normal to 5=Malignant), enabling the detection performance to be compared between prompted and unprompted conditions, by means of ROC analysis. The conclusion reached by the authors of this experiment was that radiologists would accept a maximum ratio of 3 false prompts to 2 true prompts (table 3.1).

FP:TP	Number of True Prompts	Number of False Prompts	Overall prompt rate	Improvement in detection
1:1	18	18	36%	Yes
3:2	18	27	45%	Yes
2:1	18	36	54%	No

Table 3.1: Improvement in cancer detection by relative prompt rates (experiment conducted by Dr Ian Hutt) [87]

These results, however, caused some concern to the PROMAM team. Since the national malignancy rate at mammographic screening is approximately 6 in 1000, with the recall rate nine times higher (54 in 1000), this would correspond to upper bounds of 0.9% (if only cancers count as true positives) and 8.1% (if recalled cases count as true positives) FP rates respectively. If the latter was not acceptable, it would imply the need to build algorithms capable of performing substantially better than any radiologist.

Ian Hutt's experiment, however, was heavily biased towards the TPs, with 20 out of the 100 cases being 'true', and a sensitivity of 90%. Given that the overall prompt rate of his lowest condition (36%) is higher than an average radiologist's recall rate (usually between 5 – 10%), it is not unreasonable to suggest that it is not only the FP:TP ratio that is causing this effect. It is feasible that it is either the proportion of false prompts or the overall prompting rate that is causing this non-improvement in detection ability. If this is the case, then these results suggest that an improvement is seen up to an overall prompting rate of 54%, suggesting that the point at which the prompting rate becomes more of a nuisance than an assistance is somewhere between 45% and 54% prompting rate. This translates to approximately 1 in 2 women prompted. Alternatively, this could be given as between 27% and 36% false prompt rate (approximately 1 in 3 women falsely prompted). Further concerns about the validity of the false prompts were also broached; worries that the prompts were simulated and were not representative of the type that would be created by a computer-driven prompting system.

Due to this, which the team felt was a potential bias, plus the fact that his experiment had not been completed under screening conditions, the following experiment was suggested. Since it was not known which point on the ROC curve the radiologists would find the most satisfactory, it was proposed by the team that a group of radiologists be shown a selection of prompting levels, including a 'control' level, where no prompts would be given at all.

Although the major aim of this experiment was to canvass opinion on the optimal TP/FP rates to set each algorithm, there was another potential use for the data. Fears had been raised by the directors of the clinics scheduled to be trial centres that the recall rate would rise when using the prompting method. These fears were perfectly reasonable and accepted by the team as valid causes for concern. Hence, the secondary aim was to demonstrate the extent to which the normal workings of a clinic would be greatly disrupted by the inclusion of a prompting system. This work was discussed at the Medical Image Understanding and Analysis Conference in Oxford, in 1997 [88].

3.3 Experimental Design

After discussion with the algorithm developers, it was decided to provide the radiologists with a choice of three prompting levels, similar to Ian Hutt's experiment, thus giving a total of four conditions (not prompted plus the 3 prompted conditions). These were referred to as *null* (not prompted), *low*, *medium* and *high* prompting conditions. Radiologists' time was also at a premium, and so the fewer radiologists involved in the experiment, the better (although not from an experimental analysis point of view). It had been discovered, in conversation with radiologists prior to the experiment, that their ability to remember cases, and especially cancers, was particularly high. Hence, a different set of cases had to been shown to each radiologist for each prompting level. And so, the design had to allow for each reader to see each set of films only once and each condition only once, over four sessions. Thus, given that four radiologists were available, and with four settings for the prompting level (no prompting and three levels of prompts), this led to a Graeco-Latin square experimental design.

3.3.1 The Graeco-Latin Square

A Graeco-Latin square design is a comparative design with one treatment factor (prompting levels) and three 'nuisance factors' (radiologist, set, session). These nuisance factors are used as blocking variables. The size of the n by n grid which comprises the design is dependent on the number of levels in the treatment factor; in this case, four. The design is then restricted such that each blocking factor level must appear only once in each row and column, and each factor-factor combination must appear only once.

	Session			
Radiologist	I	II	III	IV
$\mathrm{R_A}$	A1	B2	C3	D4
$\mathrm{R_B}$	B4	A3	D2	C1
$\mathrm{R_C}$	C2	D1	A4	B3
$\mathrm{R_D}$	D3	C4	B1	A2

Table 3.2: A 4 by 4 Graeco-Latin square

The table above (table 3.2) is an example of a four by four Graeco-Latin square with one treatment factor (prompting level - 1, 2, 3, 4) and three nuisance factors (set of cases - A, B, C, D; radiologist - R_A, R_B, R_C, R_D; session - I, II, III, IV).

Latin squares and, by extension, Graeco-Latin squares have the advantage of being able to handle multiple nuisance factors and be completed with relatively small samples. However, they are more complicated to randomise, and assume that there are no interactions between the nuisance factors and the treatment factor. Greater access to radiologists may have allowed the model to examine whether or not the assumption of no interaction was valid, but, pragmatically, this design was accepted as the most efficient, despite its unverifiable assumptions.

3.3.2 Final Design

The eventual design was altered randomly from that in table 3.2, to prevent R_A from reading four sets that increased in prompt rate over the four sessions, and R_B from reading four sets that decreased over sessions, to that in table 3.3.

	Session			
Radiologist	I	II	III	IV
R_A	B0	A1	C3	D2
R_B	A3	B2	D0	C1
R_C	D1	C0	A2	B3
R_D	C2	D3	B1	A0

Table 3.3: The final design for the experiment

In this table, 0 indicates the *null* prompt rate, with 1 being the *low* prompt rate and so on. The cases were gleaned from four days' worth of typical throughput from Ardmillan House, the South East Scotland Breast Screening Centre, and included two pathology proven cancers. These were divided into four sets, balancing the sets for numbers recalled or not recalled, and the site where mammogram was taken (static, i.e. Ardmillan House, or mobile unit). Each set was composed of 111 test cases and five 'warm-up' cases, with each set having six cases that were deemed worthy of recall by the Edinburgh radiologists during the original screening. The two cancers were categorised as 'recalled' for the purpose of this experiment; one appeared in set A, the other in set B.

The differing prompt rates for the three prompted conditions were chosen by the algorithm designers to represent relative prompt rates of 1 in 6 (*low*), 1 in 3 (*medium*) and 2 in 3 (*high*) films prompted. After some discussion by the team, it was decided to vary both algorithms, rather than fix one and allow the other to vary. Table 3.4 shows the actual prompt rates that were used in the experiment, giving overall rates of 19% (*low*), 36% (*medium*) and 64% (*high*). Prompt rate is defined as the number of prompts generated over the four sets divided by 464 (the total number of cases used in the experiment). The figures are not strictly additive, since some cases had prompts for both microcalcification and masses.

	Low			Medium			High		
	Calc	Mass	Total	Calc	Mass	Total	Calc	Mass	Total
Set A	9	15	23	19	29	44	37	55	76
Set B	12	11	22	23	22	40	40	51	70
Set C	7	17	23	15	31	41	30	61	76
Set D	9	13	21	18	31	44	35	63	74

Table 3.4: The number of women prompted by each algorithm at each condition level

The TP rates at the *low*, *medium* and *high* prompt rates were 22%, 37% and 62% for the mass algorithm and 76%, 86% and 94% for the micro-calcification algorithm, based on a test set of pathology proven cancers. The algorithms were distinguished on the prompt sheet by the shape of the prompts; an ellipse denoted a suspicious feature detected by the mass algorithm, and a polygon delineated the edge of a suspected micro-calcification cluster. An example of a prompt sheet is given in Appendix D.

3.3.3 Experimental Protocols

Radiologists

The four radiologists were recruited from the Aberdeen and Glasgow Woodside screening centres, two from each, in an attempt to avoid utilising staff who might participate in the full trial. Written instructions were supplied prior to the start of the experiment (see Appendix E), detailing what we expected of them. The cases were given in three stages; the test reading set (five cases), group one (56 cases), group two (55 cases). There was a 15 minute break between groups one and two. At the start of each session, the radiologists were informed of the expected prompt rate and sensitivity (*low*, *medium*, *high*).

Before beginning their first session, the radiologists were asked to complete a questionnaire about their attitudes to prompting (see Appendix G.1). This was repeated at the end of their final session (Appendix G.2). They were also asked to complete a more specific questionnaire at the end of each prompted session, asking about the prompting level of the set they had just completed (Appendix G.3).

Prompt sheet

Since the experiment was to take place under screening conditions, the prompt sheet was included with the usual film bag, attached to the reporting form in such a way that the reporting form had to be lifted to examine the prompt sheet. It was suggested that films should be examined in the usual way before checking the prompt sheet. A prompt sheet was produced for each case, regardless of whether the algorithms had found any suspicious features or not. This was thought to be the safest option, as the absence of a prompt sheet could be due to the absence of a suspicious feature, lack of paper in the printer, or the sheet getting lost in the process.

Reporting

Reporting was done in the usual clinic way of entering the results on a standard SBSP reporting form, recording tech recall, recall or normal.

Films

Copy films were used in each case, since the films in question would be away from Ardmillan House for substantial lengths of time. It was not possible to supply previous and

CC films for the experiment, as each case consisted of two films, making a total of 928 films that had to be transported to the clinics by Pat Dixon, the project radiographer. There were also cost considerations to take into account, as even including only the CC views would double the cost of copying the films.

3.4 Results - Recalls

The following results sections contain output from various SAS procedures, where variable names have been shortened due to the nature of the software. Variable names are consistent and reasonably intuitive.

- COND denotes the prompting conditions, a four level factor (*null, low, medium, high*)
- RAD denotes the radiologists ($\mathrm{R_A}$, $\mathrm{R_B}$, $\mathrm{R_C}$, $\mathrm{R_D}$)
- SET denotes the four sets of films used in the experiment (A, B, C, D)
- SESSION denotes the session number of a particular reading session (1, 2, 3, 4)

3.4.1 Exploratory analysis

As with most analyses, we will begin with some simple exploratory analyses of the factor of greatest interest; the prompting condition. Figure 3.2 illustrates the four factors which were employed in the graeco-latin square design, in relation to the number of recalls made.

Prompting Condition

This is the factor in which we are most interested; whether the addition of prompts increases the recall rate, and if so, whether the increase in the number of prompts has a corresponding increase in the number of recalls. The mean and standard errors for each level are included in the table below (table 3.5).

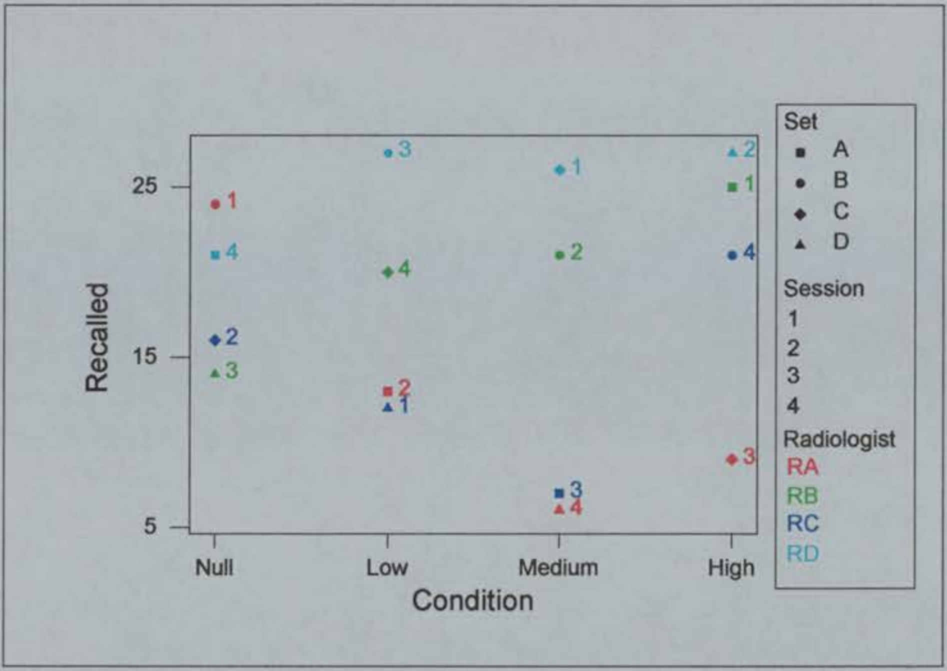


Figure 3.2: Number of recalls by condition, with radiologist, set and session identifers

Condition	Mean recalls	SE Mean
Null	18.75	2.29
Low	18.00	3.49
Medium	15.00	5.02
High	20.50	4.03

Table 3.5: Mean recalls per condition

Simple analysis of these results using ANOVA gives a non-significant model F-value and non-significant between condition differences.

```
                        Sum of
Source          DF     Squares    Mean Square   F Value   Pr > F
Model            3  63.1875000     21.0625000      0.36   0.7843
Error           12 705.7500000     58.8125000
Corrected Total 15 768.9375000
```

This would imply that there is little influence on the recall rate from the prompting levels.

Given that the expected model proposes that there would be an increase in the recall rate as the prompt rate increases, the drop in mean recalls at the *medium* condition is, at first glance, unexpected. However, the variation in recall rate over the four radiologists for the condition is large (6, 7, 21 and 26 recalls, see also figure 3.2), giving the largest standard error at 5.02 and a 95% confidence interval of 6.6 to 23.4. We also encounter the problems of multiple testing, which would suggest that this is a spurious result. A χ^2 test for trend against the condition is non-significant. That is, there is no increase/decrease from the *null* condition to the *high* condition.

Another alternative is to recode the condition into prompted and not prompted (12 prompted and four not prompted). This gives a mean over the prompted conditions of 17.83, and is also not significantly different from the *null* recall rate (p=0.83).

These exploratory methods are all pointing towards the same conclusion; that the prompting condition had little influence on the recall rate.

Radiologist

With the conclusion that prompting condition is not responsible for the great variability between the number of recalls, we must examine other potential sources. From the examination of figure 3.2, it can be seen that radiologist differences appear to be supplying most of the variability (table 3.6). Not only are there great differences between the readers, but their recall behaviour, in terms of whether recalls increases with numbers of prompts, is also different between readers (see table 3.7).

Radiologist	Mean recalls	SE Mean
R_A	13.00	3.94
R_B	20.00	2.27
R_C	14.00	2.97
R_D	25.25	1.44

Table 3.6: Mean recalls per radiologist

Analysis of variance in this case is significant (p=0.03), with significant differences between R_A and R_D (95% CI 3.6,20.9) and between R_C and R_D (95% CI 2.6, 19.9).

Examination of the two factors together produces table 3.7. Since the sets are the same size (i.e. 111 cases), only the actual number of recalled cases will be noted, rather than

	Condition			
Radiologist	Null	Low	Medium	High
R_A	24	13	6	9
R_B	14	20	21	25
R_C	16	12	7	21
R_D	21	27	26	27

Table 3.7: Number of recalls

the recall rate.

Analysis of variance of the influence of both factors yielded the following:

```
                                Sum of
Source               DF        Squares     Mean Square    F Value    Pr > F
Model                 6    453.3750000      75.5625000       2.16    0.1446
Error                 9    315.5625000      35.0625000
Corrected Total      15    768.9375000


          R-Square     Coeff Var      Root MSE    recall Mean
          0.589612      32.78261      5.921360       18.06250


Source               DF       Anova SS     Mean Square    F Value    Pr > F
COND                  3     63.1875000      21.0625000       0.60    0.6306
RAD                   3    390.1875000     130.0625000       3.71    0.0550
```

With an F-value of 3.71, the radiologist effect is still influential, once the prompting condition has been accounted for.

However, radiologist and condition are not the only factors that may have some influence on the recall rate. Once the set and session factors have been included in the model, radiologist differences attain significance at the 5% level (p=0.03).

```
                                Sum of
Source               DF        Squares     Mean Square    F Value    Pr > F
Model                12    737.7500000      61.4791667       5.91    0.0848
Error                 3     31.1875000      10.3958333
Corrected Total      15    768.9375000

          R-Square     Coeff Var      Root MSE    recall Mean
          0.959441      17.85056      3.224257       18.06250
```

```
Source          DF        Anova SS     Mean Square    F Value    Pr > F
COND             3      63.1875000      21.0625000       2.03    0.2884
RAD              3     390.1875000     130.0625000      12.51    0.0334
SET              3     161.6875000      53.8958333       5.18    0.1049
SESSION          3     122.6875000      40.8958333       3.93    0.1451
```

3.4.2 The theoretical model

Given the nature of the data (proportions, with a number of positive 'hits', r, from a potential maximum number of objects, n), analysis of variance was not an appropriate method of analysis, as it assumes that the dependent variable is Normally distributed. Hence, a generalised linear model was fitted, using the logit link function [89]. Traditional linear models rely on assumptions that may not hold true for certain types of data; in this case, the data are restricted to a range of values, [0, 1], whereas the linear predictor can take any value.

A traditional linear model is of the form

$$y_i = x_i'\beta + \varepsilon_i$$

where y_i is the response variable for the i^{th} observation, x_i is a column vector of explanatory variables for observation i, β is the vector of unknown covariates, estimated by a least squares fit to the data, and ε_i are the errors and are assumed to be independent, normal random variables with a mean of zero and a constant variance.

The expected value of y_i is denoted by μ_i, and

$$\mu_i = x_i'\beta$$

A generalised linear model, however, consists of the following components.

- The linear component: $\eta_i = x_i'\beta$
- A monotonic link function (in this case, the logit link): $\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$

- The set of response variables y_i are independent and have a probability distribution from an exponential family.

The SAS procedure GENMOD (see Appendix F) fits a generalised linear model to the data by maximum likelihood estimation of the parameter vector β, using an iterative fitting process.

Results

Although only the level of prompting rate as a determinant of recall rate is of interest, the other factors must be included in the analysis in order to discount for any effect these may have on the recall rate. In the following section, the recalls (as a binomial variable) will be analysed using the GENMOD procedure in SAS version 8.2. The following excerpts are from the analysis of the recalls, as generated by the program in Appendix F.1.1.1.

As before, RAD=radiologist, COND=prompting condition, SET=which particular set of cases was being read and SESSION=the order in which the sets were read.

```
Wald Statistics For Type 3 Analysis

Source        DF    ChiSquare  Pr>Chi

RAD            3      29.2525  0.0001
COND           3       7.3865  0.0605
SET            3      14.7659  0.0020
SESSION        3      10.8314  0.0127
```

Type 3 analysis considers each of the factors in turn as the last factor after all the others have been fitted. From these results, we can see that radiologist differences account for much of the variability, followed by set and session differences. Condition (i.e. prompting rate) is the least contributory with a χ^2 value of 7.39, which is non-significant at the 5% level. In other words, of all the factors which contribute to the variability of the results, condition contributes the least.
The following is a more detailed breakdown of the results.

```
Parameter           DF    Estimate     Std Err    ChiSquare  Pr>Chi

INTERCEPT            1    -1.4697      0.2454      35.8737   0.0001
RAD        A         1    -0.9022      0.1947      21.4779   0.0001
RAD        B         1    -0.3199      0.1697       3.5536   0.0594
RAD        C         1    -0.7956      0.1888      17.7514   0.0001
RAD        D         0     0.0000      0.0000        .        .

COND       high      1     0.0812      0.1810       0.2012   0.6537
COND       low       1    -0.0719      0.1862       0.1493   0.6992
COND       medium    1    -0.4417      0.2040       4.6890   0.0304
COND       null      0     0.0000      0.0000        .        .
SET        A         1     0.1110      0.2056       0.2916   0.5892
SET        B         1     0.6615      0.1910      11.9886   0.0005
SET        C         1     0.2933      0.2021       2.1054   0.1468
SET        D         0     0.0000      0.0000        .        .
SESSION    1         1     0.3764      0.1860       4.0965   0.0430
SESSION    2         1     0.2111      0.1924       1.2048   0.2724
SESSION    3         1    -0.2411      0.2057       1.3739   0.2411
SESSION    4         0     0.0000      0.0000        .        .
SCALE                0     1.0000      0.0000        .        .
```

In the PROC GENMOD output, each level within a factor (e.g. each radiologist within the factor RAD) is compared to the 'last' level (either alphabetically or numerically). For example, conditions *low* and *high* are not significantly different from the *null* condition, but the condition *medium* is.

The above analysis does not take into account the fact that the films in the sets can be and, indeed, should be treated as repeated measures, with each case being read four times (see Appendix F.1.1.2). In the standard model illustrated above, all error terms are assumed to be independent. Since it is unlikely that the response to a particular film is independent between radiologists, we must account for this correlation in our model by relaxing the assumption of independence. A compound symmetry covariance pattern has been fitted to the model, as this tends to be reliable in small data sets [90].

The results from including these features in the model are similar in effect to the simpler analysis, although differ in degree.

```
                Wald Statistics For Type 3 GEE Analysis

                                        Chi-
              Source          DF      Square    Pr > ChiSq

              RAD              3       39.43        <.0001
              COND             3        9.23        0.0264
              SET              3        6.80        0.0787
              SESSION          3       16.71        0.0008
```

As differences between the sets are now accounted for, in some part, by the repeated measures on the film identifier (CHI), set has become less important to the model, while the other three factors have become more so, with the prompting condition now attaining significance at 5%.

```
                             Empirical  95% Confidence Limits
Parameter         Estimate    Std Err      Lower       Upper         Z  Pr>|Z|

INTERCEPT          -1.4550     0.2484    -1.9417     -0.9682    -5.858  0.0000
RAD       A        -0.9020     0.1581    -1.2119     -0.5922    -5.706  0.0000
RAD       B        -0.3181     0.1373    -0.5871     -0.0491    -2.317  0.0205
RAD       C        -0.7985     0.1572    -1.1067     -0.4904    -5.079  0.0000
RAD       D         0.0000     0.0000     0.0000      0.0000    0.0000  0.0000
COND      high      0.0814     0.1336    -0.1805      0.3432    0.6091  0.5425
COND      low      -0.0745     0.1593    -0.3867      0.2377    -.4678  0.6400
COND      medium   -0.4465     0.1790    -0.7973     -0.0956    -2.494  0.0126
COND      null      0.0000     0.0000     0.0000      0.0000    0.0000  0.0000
SET       A         0.0641     0.2923    -0.5089      0.6371    0.2192  0.8265
SET       B         0.6486     0.2682     0.1230      1.1741    2.4186  0.0156
SET       C         0.3031     0.2534    -0.1936      0.7997    1.1961  0.2317
SET       D         0.0000     0.0000     0.0000      0.0000    0.0000  0.0000
SESSION 1           0.3757     0.1378     0.1057      0.6457    2.7271  0.0064
SESSION 2           0.2082     0.1701    -0.1251      0.5415    1.2244  0.2208
SESSION 3          -0.2441     0.1628    -0.5632      0.0751    -1.499  0.1339
SESSION 4           0.0000     0.0000     0.0000      0.0000    0.0000  0.0000
Scale               1.0016          .          .           .         .       .
```

Although it is preferable to have a particular level within a factor be the 'standard' for comparison (e.g. a placebo versus one or more treatments), it is not always possible to define such a standard. In this experiment, only condition (COND) and session (SESSION) have easily defendable 'standards'. E.g. the first three sessions are compared to the last session - does the recall rate change over time? Or, are *high*, *medium* and *low* significantly different from *null*? Unfortunately, this does not hold for radiologists and sets, and so one is arbitrarily selected to become the standard. It is not of particular interest whether Radiologist A differs from Radiologist C; it only matters that there is some difference between one or more of the radiologists, and so this is not a problem. However, should it later be desired, one may either re-order the levels within the factor, or calculate the covariance matrix and determine the answer from the estimates and standard errors given.

By recoding the levels of condition into 0, 1, 2, and 3 for *null*, *low*, *medium* and *high* (see Appendix F.1.1.3), then allowing these values to be considered ordinal rather than categorical, the following result from the logistic regression are obtained, where the renamed COND is now PROMPT:

```
Wald Statistics For Type 3 Analysis
Source         DF   ChiSquare  Pr>Chi

PROMPT          1      0.0042  0.9483
RAD             3     26.6281  0.0001
SET             3     12.2560  0.0066
SESSION         3      9.4001  0.0244
```

From this we can see that there is again no trend (in either direction) once the other factors have been accounted for, as suggested by the χ^2 test for trend discussed earlier. The significant difference between the *null* and *medium* condition is likely to be a spurious result.

The above analyses all underpin the conclusion that a variety of factors influence a radiologist's recall rate, and that the number of prompts generated is likely to play only a small part in that variation.

3.5 Results - Time taken to complete the experiment

Table 3.8 contains the sum of the times for the two halves of each reading sessions, given in seconds.

	Condition			
Radiologist	Null	Low	Medium	High
R_A	1926	1995	1797	2048
R_B	1535	1789	2219	2509
R_C	1487	1909	1586	1845
R_D	1729	1858	2503	2043
Mean	1669.25	1887.75	2026.25	2111.25

Table 3.8: Time taken to complete a set of cases in seconds

Figure 3.3 shows the time taken to complete a set of cases within each condition. As can be seen, only radiologist B shows any indication of a consistent increasing trend towards the higher prompt rates, despite there being an overall mean increase with increasing prompt rate.

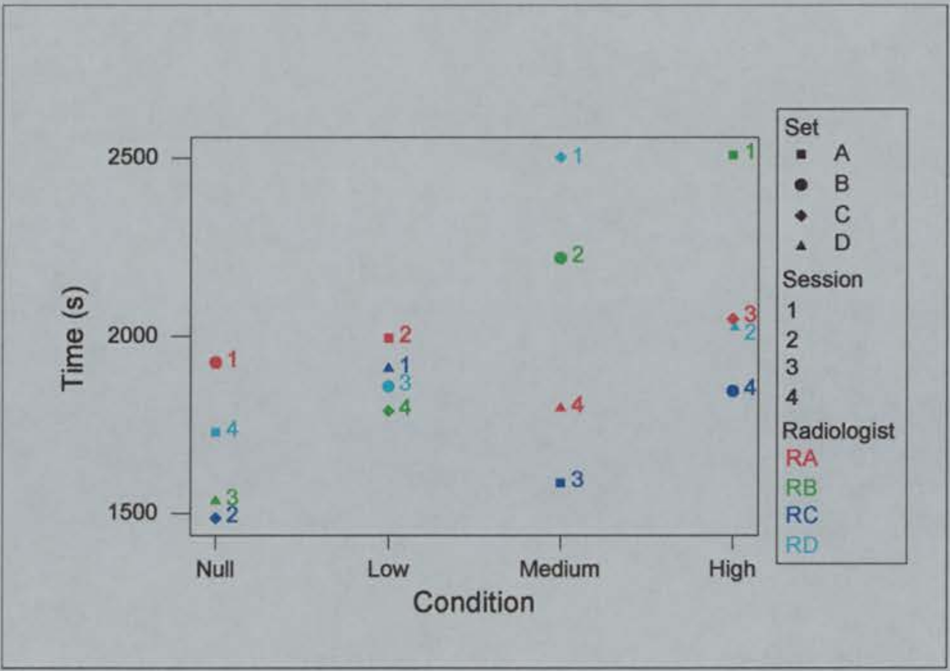


Figure 3.3: Time (seconds) to complete each set by condition

Initial examination of the data revealed the distribution of the times to complete a set to be sufficiently symmetrical to allow the assumptions of Normality to hold without resorting to transformations. As with the recalls, a generalised linear model is used to model the data, although using PROC GLM rather than PROC GENMOD.

Below is the SAS output generated by the generalised linear model (see Appendix F.1.2.1), where TIMES is the continuous variable time in seconds.

```
Dependent Variable: TIMES
                                 Sum of               Mean
Source              DF          Squares             Square   F Value      Pr >F
Model               12      1288124.500         107343.708      6.35     0.0771
Error                3        50681.250          16893.750
Corrected Total 15          1338805.750

        R-Square              C.V.          Root MSE              TIMES Mean
        0.962144          6.756824          129.9760                1923.625


Source              DF      Type III SS        Mean Square   F Value      Pr >F

COND                 3      446914.7500        148971.5833      8.82     0.0535
RAD                  3      269439.2500         89813.0833      5.32     0.1017
SET                  3       56282.2500         18760.7500      1.11     0.4667
SESSION              3      515488.2500        171829.4167     10.17     0.0442
```

In this analysis, only session appeared to make any significant contribution to the time taken to complete a set, although condition is only just non-significant at p=0.054. As the 5% significance level is only a widely held convention and not actually a mathematical rule, there is evidence to suggest that COND is also influencing the time taken to complete a set.

Pairwise comparisons of the levels of prompting illustrate where the significant differences occur:

```
                     General Linear Models Procedure
                           Least Squares Means
      COND            TIMES   Pr > |T| HO: LSMEAN(i)=LSMEAN(j)
                     LSMEAN   i/j      1        2        3        4

      NULL       1669.25000   1    .        0.0978   0.0302   0.0171
      LOW        1887.75000   2   0.0978    .        0.2289   0.0932
      MEDIUM     2026.25000   3   0.0302   0.2289    .        0.4233
      HIGH       2111.25000   4   0.0171   0.0932   0.4233    .
```

The decreasing p-value in the comparisons of *null* with each of the prompted sessions, along with the increasing mean times (table 3.8), suggests that there is an underlying trend. Again setting *null*, *low*, *medium* and *high* to be 0, 1, 2, and 3 and ordinal (see Appendix F.1.2.2), the following is obtained:

```
                  General Linear Models Procedure

Dependent Variable:  TIMES
                              Sum of          Mean
Source             DF        Squares        Square   F Value    Pr > F

Model              10    1270161.800    127016.180      9.25    0.0120
Error               5      68643.950     13728.790
Corrected Total    15    1338805.750

          R-Square           C.V.     Root MSE          TIMES Mean
          0.948727       6.091100     117.1699            1923.625


Source          DF      Type III SS    Mean Square   F Value     Pr > F

PROMPT           1      428952.0500    428952.0500     31.24     0.0025
RAD              3      269439.2500     89813.0833      6.54     0.0350
SET              3       56282.2500     18760.7500      1.37     0.3538
SESSION          3      515488.2500    171829.4167     12.52     0.0092
```

Here, it is clear that PROMPT (the 0, 1, 2, 3 values of COND) has a significant effect on the model, as is suggested by figure 3.3. Thus the length of time required to complete a session increases with the number of prompts generated. However, as this does not coincide with a corresponding increase in the recall rate with respect to the number of prompts, it is of limited concern. Screening time is not a major part of a radiologist's workload [5].

3.6 Results - Observational Data

During the course of the experiment, the radiologists were observed on their usage of the system, in particular, their adherence to the protocol. Radiologists were asked to examine the film, examine the prompt sheet, then record their decision. In the cases where this failed, radiologists either failed to examine the prompt sheet (*nlp* = not looked at prompt) or had marked their response before examining the prompt sheet (*mf* = marked first).

Tables 3.9 and 3.10 clearly illustrate the contention that a *low* prompting rate is as-

Radiologist	R_A		R_B		R_C		R_D	
Condition	nlp	mf	nlp	mf	nlp	mf	nlp	mf
Low	1	20	6	2	3	1	0	20
Medium	0	1	0	10	1	0	0	1
High	0	0	0	5	0	0	1	0

Table 3.9: Failure to adhere to protocol by failure type

Condition	R_A	R_B	R_C	R_D
Low	21	8	4	20
Medium	1	10	1	1
High	0	5	0	1

Table 3.10: Failure to adhere to protocol - summary of table 3.9 by radiologist

sociated with protocol errors, indicating that a low prompting rate is insufficient to maintain the interest of the radiologists. The results from a Fisher's exact test on *nlp* and *mf* are both significant (p=0.03 and <0.001, respectively), which would agree with this conclusion. Work done by Mark Hartswood [91] on the free form questions in Appendix G also agrees, with one of the radiologists feeling that the *low* condition was of such little aid that s/he was tempted to not bother looking at the prompts.

Similarly, if not looking at the prompt or marking the response before examining the prompt sheet are considered as failures to follow the protocol, then *failure=nlp+mf* (table 3.10). Fisher's exact test of the variable *failure* against condition also produces a significant result ($p < 0.0001$).

3.7 Results - Questionnaires

3.7.1 Pre- and post-experiment questionnaires

This section deals with the radiologists' responses to the questionnaires put to them before and after the experiment and, additionally, after a prompted reading session (see Appendix G). The questions were composed by Dr Mark Hartswood.

Q1. Would you prefer a system which has a high sensitivity with a high FP rate or a system with a lower sensitivity and a low FP rate?

As can be seen from table 3.11, radiologist B changed his/her opinion after the experiment. This question was asked in order to see whether the high number of false prompts was compensated for by the higher sensitivity. All four readers agreed that the higher TP/FP rate was preferable to a low FP/TP rate by the end of the experiment.

Radiologist	Before	After
$\mathrm{R_A}$	High	High
$\mathrm{R_B}$	Low	High
$\mathrm{R_C}$	High	High
$\mathrm{R_D}$	High	High

Table 3.11: *Q1. Would you prefer a system which has a high sensitivity with a high FP rate or a system with a lower sensitivity and a low FP rate?* Before and After

Q2. Rate the distracting effect of algorithm output (1 = Useful, 5 = Distracting)

Before	Rating					Average	
Feature	1	2	3	4	5	Rating	Rank
Vascular calcification					4	5	9
Benign clusters			2		2	4	5.5
"Popcorn" calcification			1		3	4.5	7.5
Film artefacts			1		3	4.5	7.5
Lymph nodes		1		2	1	3.75	3.5
Well defined masses	1	1		2		2.75	1.5
Composite shadows	1	1		2		2.75	1.5
Nodular glandular structure			1	3		3.75	3.5
Cysts				4		4	5.5

Table 3.12: *Q2. Rate the distracting effect of algorithm output (1 = Useful, 5 = Distracting)*: Before

As expected, vascular calcifications top the 'must remove' list in table 3.12, where the highest rank indicated the prompting feature the radiologists would most like to see removed. The only other feature that had such agreement were cysts, the other features were more divided.

After the experiment, vascular calcifications were again considered the worst, this time sharing the 'top spot' with film artefacts, such as scratches on the film (table 3.13).

After	Rating					Average	
Feature	1	2	3	4	5	Rating	Rank
Vascular calcification				1	3	4.75	8.5
Benign clusters		2		1	1	3.25	5
"Popcorn" calcification			2		2	4	6
Film artefacts				1	3	4.75	8.5
Lymph nodes			1	1	2	4.25	7
Well defined masses	2		1	1		2.25	2
Composite shadows	1		3			2.5	4
Nodular glandular structure	1	1	2			2.25	2
Cysts	1	1	2			2.25	2

Table 3.13: *Q2. Rate the distracting effect of algorithm output (1 = Useful, 5 = Distracting)*: After

A couple of other features were mentioned by two of the radiologists; asymmetry and parenchymal distortion/spiculated masses. Since the algorithms were not designed to detect the presence of spiculated/tentacled/stellate masses without a nidus, they were not features that caused much concern to the algorithm developers. The information in table 3.13 was subsequently used by the algorithm developers to determine the areas which needed development, and which could be given lower priority.

Q3. Rank the false positives to be removed first

Before	Rank								
Feature	1	2	3	4	5	6	7	8	9
Vascular calcification	3	1							
Benign clusters		2	1	1					
"Popcorn" calcification	1	2	1						
Film artefacts	3	1							
Lymph nodes	1	1	1		1				
Well defined masses			3						1
Composite shadows		1	1		1		1		
Nodular glandular structure	1	1			1			1	
Cysts			2	1		1			

Table 3.14: *Q3. Rank the false positives to be removed first*: Before

Unfortunately, this question was not clearly defined, with some radiologists ranking the features in order, and others allowing multiple features the same ranking. Hence the preponderance of low ranks (first to be removed). It is still fairly clear, however (see

After	Rank								
Feature	1	2	3	4	5	6	7	8	9
Vascular calcification	3	1							
Benign clusters				2	2				
"Popcorn" calcification			3	1					
Film artefacts	1	3							
Lymph nodes		1	1	1	1				
Well defined masses			1			1			2
Composite shadows			1			1	2		
Nodular glandular structure				1				2	1
Cysts				1		1	1	1	

Table 3.15: *Q3. Rank the false positives to be removed first*: After

tables 3.14 and 3.15), that vascular calcifications are the most annoying feature to the radiologists.

Q4. In cases where you are unsure, would the presence of a prompt make you more inclined to recommend recall? (Strongly agree to Strongly Disagree)

Although only one radiologist changed their opinion, it was a positive change.

Radiologist	Before	After
$\mathrm{R_A}$	Agree	Strongly agree
$\mathrm{R_B}$	Agree	Agree
$\mathrm{R_C}$	Disagree	Disagree
$\mathrm{R_D}$	Agree	Agree

Table 3.16: *Q4. Would the presence of a prompt make you more inclined to recommend recall? (Strongly agree to Strongly Disagree)*: Before and After

Q5. In cases where you are unsure, would the absence of a prompt make you less likely to recommend recall? (Strongly agree to Strongly Disagree)

Radiologist	Before	After
$\mathrm{R_A}$	Agree	Uncertain
$\mathrm{R_B}$	Uncertain	Agree
$\mathrm{R_C}$	Agree	Disagree
$\mathrm{R_D}$	Uncertain	Uncertain

Table 3.17: *Q5. In cases where you are unsure, would the absence of a prompt make you less likely to recommend recall? (Strongly agree to Strongly Disagree)*: Before and After

This result is a little more difficult to interpret, although the uncertainty is likely to be due to the fact that the masses algorithm was only 62% sensitive at its highest setting.

Q6. Rate the possible configurations (1=Most useful to 5=Least useful)

Before	Rating					Average
Configuration	1	2	3	4	5	Rating
High prompt rate, high sensitivity	2	1		1		2
Low prompt rate, low sensitivity	1			1	2	3.75
Micro-calcification clusters, but no other types of calc	3	1				1.25
All types of calcification			1	3		3.75
Opacities usually dismissed with previous or multiple films	2	2				1.5

Table 3.18: *Q6. Rate the possible configurations (1=Most useful to 5=Least useful)*: Before

After	Rating					Average
Configuration	1	2	3	4	5	Rating
High prompt rate, high sensitivity	3	1				1.25
Low prompt rate, low sensitivity					4	5
Micro-calcification clusters, but no other types of calc	3	1				1.25
All types of calcification		1		2	1	3.75
Opacities usually dismissed with previous or multiple films	1	2	1			2

Table 3.19: *Q6. Rate the possible configurations (1=Most useful to 5=Least useful)*: After

The high sensitivity/high prompt rate has become more popular after the experiment, and the low sensitivity/low prompt rate has become less so. The opinions on micro-calcification clusters remain the same, but the spread of opinion on all types of calcification has widened. This, however, reveals nothing about the individual changes. Figure 3.4 shows the changes made by each radiologist (not identified) in each of the five categories.

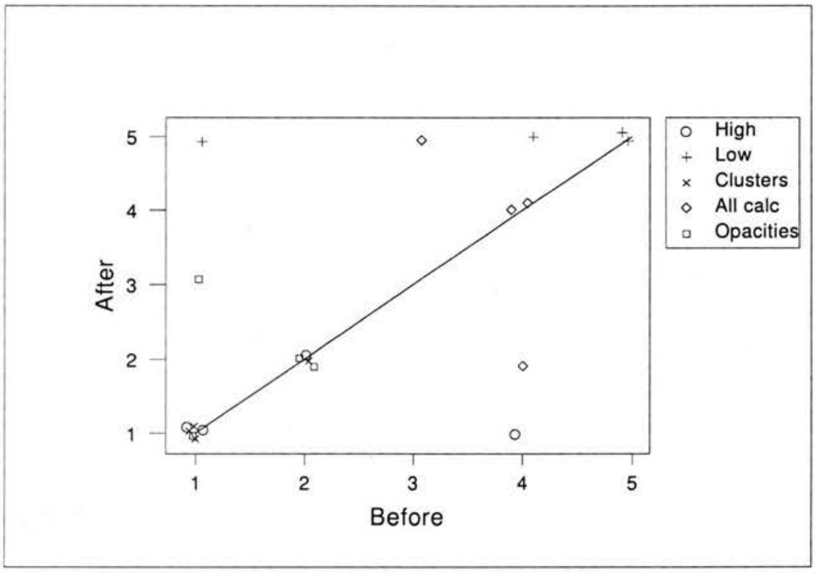


Figure 3.4: Plot of *Q6 Rate the possible configurations*: Before and after

3.7.2 Post-experiment only questions

Q7. Which would be the most useful in a screening context?

	High	Medium	Low	No prompts
Mass prompt rate	3	1		
Calcification prompt rate	4			
Sensitivity	3	1		

Table 3.20: *Q7. Which would be the most useful in a screening context?*

Q8. What is the highest FP rate you would be willing to accept?

	High	Medium	Low	No prompts
Mass prompt rate	3	1		
Calcification prompt rate	4			
Sensitivity	3	1		

Table 3.21: *Q8. What is the highest FP rate you would be willing to accept?*

Q9. What is the lowest sensitivity you would find useful in a screening context?

From the above three questions (tables 3.20 - 3.22), it is fairly obvious that the high sensitivity/prompt rate is the preferred option in most cases, although half of the four

	High	Medium	Low	No prompts
Mass prompt rate	2	2		
Calcification prompt rate	2	2		
Sensitivity	2	2		

Table 3.22: *Q9. What is the lowest sensitivity you would find useful in a screening context?*

radiologists would be willing to accept the medium rate as the lowest useful rate. The most interesting result, however, is the discrepancy between the preference for the masses and micro-calcification prompts; although one radiologist found the medium mass prompts more useful, all four preferred the high micro-calcification prompts.

3.7.3 Post-session questionnaires

The Likert Scores and Radiologists' Rating (Q10 and Q11)

A series of questions was put to the radiologist after each prompted session (not for the unprompted session), with the response rated on a five point scale of strongly agree to strongly disagree [92]. Since each question was biased towards a positive or negative attitude, each question was scored accordingly (1 to 5 for a negative question, 5 to 1 for a positive question). These were then tallied to give an overall Likert score, a value between 20 and 100 indicating the radiologists' opinions of the system. An entirely subjective value was also asked of them; to rate the system on a scale of 0 to 100 (see Appendix G.3). These scores are recorded in table 3.23.

	Prompting rate					
	Low		Medium		High	
Radiologist	Likert	Opinion	Likert	Opinion	Likert	Opinion
$\mathrm{R_A}$	76	20	64	20	63	40
$\mathrm{R_B}$	56	20	59	50	63	40
$\mathrm{R_C}$	72	60	77	80	86	90
$\mathrm{R_D}$	54	20	71	15	81	60

Table 3.23: Table of Likert Score and radiologists' opinions

It would appear that the Likert score and the radiologists' opinions are only roughly related (see figure 3.5), with a correlation coefficient of 0.65 ($p = 0.02$), with a Spearman's

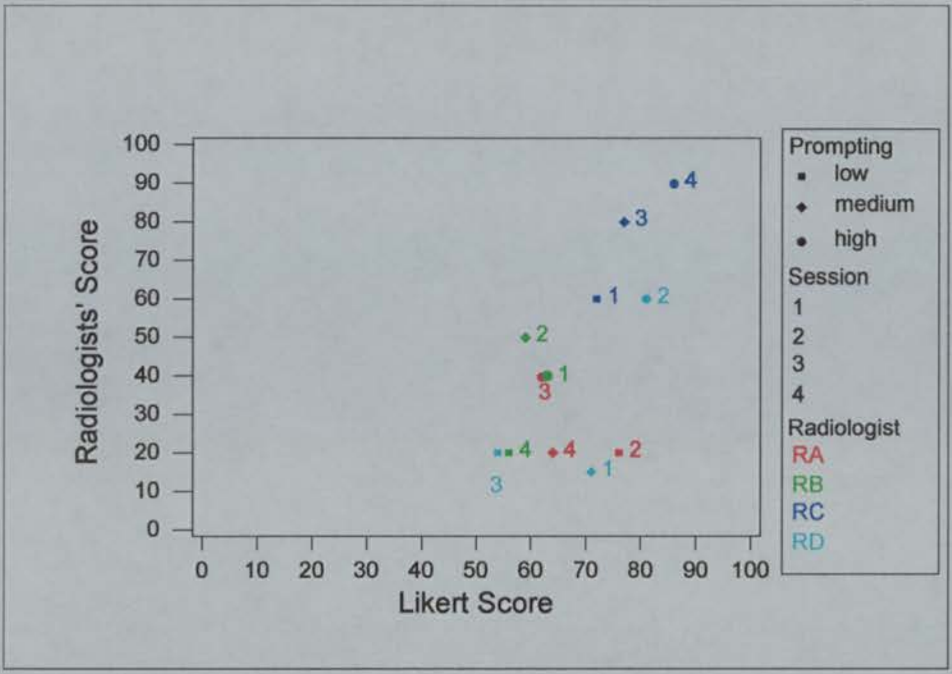


Figure 3.5: Radiologists' score by Likert score

rank correlation coefficient of 0.57 (p = 0.052)

General Linear Models

Since no Likert scores were recorded for the *null* condition, there were insufficient degrees of freedom to fit all four factors, which would lead to SAS fitting a saturated model. Since each level of SET has approximately the same number of prompts, there should not be any difference between sets and, in order to generate a residual term, it was omitted from the model.

The PROC GLM procedure produced the following analysis for the Likert score - see Appendix F.1.3.1:

```
Dependent Variable: LIKERT
                              Sum of           Mean
Source            DF         Squares         Square   F Value      Pr > F

Model              8    1016.2500000    127.0312500      2.91      0.2048
Error              3     130.7500000     43.5833333
Corrected Total   11    1147.0000000
```

```
          R-Square     Coeff Var      Root MSE    likert Mean
          0.886007      9.637617      6.601767       68.50000

Source          DF     Type III SS     Mean Square  F Value    Pr > F

COND             2    156.50000000     78.25000000     1.80    0.3071
RAD              3    778.75000000    259.58333333     5.96    0.0884
SESSION          3    315.41666667    105.13888889     2.41    0.2442
```

None of the factors of interest have a significant influence on the Likert score. However, the power of the experiment is not likely to be great, and the F-value of 5.96 for the radiologist differences may be suggestive of an underlying effect.

In a similar analysis, the radiologists' scores were also analysed with PROC GLM, and produced the following:

```
Dependent Variable: score

                                          Sum of
Source                DF         Squares     Mean Square    F Value  Pr > F

Model                  8     7079.166667      884.895833      60.68  0.0031
Error                  3       43.750000       14.583333
Corrected Total       11     7122.916667

           R-Square      Coeff Var      Root MSE     score Mean

           0.993858       8.898205      3.818813       42.91667

Source     DF    Type III SS    Mean Square    F Value    Pr > F


COND        2    1529.166667     764.583333      52.43    0.0046
RAD         3    5443.750000    1814.583333     124.43    0.0012
SESSION     3     843.750000     281.250000      19.29    0.0183
```

Here, the three factors under examination are all significant (Type III error) at the 5% level, and the three levels of prompting rate are all significantly different.

```
           Least Squares Means for Effect COND
          t for H0: LSMean(i)=LSMean(j) / Pr > |t|

                  Dependent Variable: score

     i/j          high         medium          low
    high                       0.0092       0.0020
  medium        0.0092                      0.0252
     low        0.0020         0.0252
```

However, since there are multiple comparisons, care should be taken in the acceptance of significance. There are many suggested ways of dealing with repeated significance tests; for example, the Bonferroni adjustment (number of tests × $p-$value), although this can be extremely conservative. This does, however, have the advantage that any significant results may be regarded with a fair degree of confidence.

```
       Adjustment for Multiple Comparisons: Bonferroni
            Least Squares Means for Effect cond
          t for H0: LSMean(i)=LSMean(j) / Pr > |t|

                  Dependent Variable: score

     i/j          high         medium          low
    high                       0.0276       0.0061
  medium        0.0276                      0.0755
     low        0.0061         0.0755
```

Using the Bonferroni correction for multiple comparisons, *high* has a significantly higher score than *medium* and *low*, although *medium* and *low* are not significantly different. This is illustrated in figure 3.6.

From the results above, there is a decreasing p-value as the prompting level differences increases, suggesting that there may be a trend for a higher score to be related to a higher prompting level. This may be investigated by recoding the prompting into an ordinal variable (PROMPT) as before.

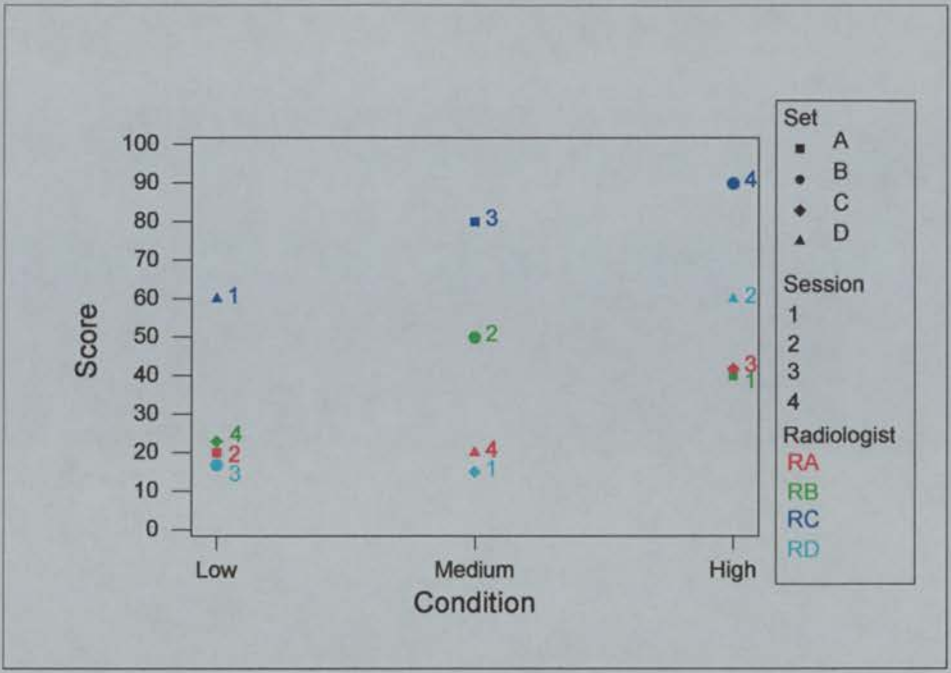


Figure 3.6: Likert score by condition

```
                                  Sum of
Source                 DF        Squares     Mean Square    F Value   Pr > F
Model                   7    7062.500000     1008.928571      66.80   0.0006
Error                   4      60.416667       15.104167
Corrected Total        11    7122.916667

          R-Square      Coeff Var      Root MSE     score Mean
          0.991518       9.055708      3.886408       42.91667

Source               DF      Type III SS     Mean Square    F Value   Pr > F

PROMPT                1      1512.500000     1512.500000     100.14   0.0006
RAD                   3      5443.750000     1814.583333     120.14   0.0002
SESSION               3       843.750000      281.250000      18.62   0.0082
```

This (and figure 3.6) indicates quite clearly that the radiologists' score increases as the prompting level increases. As this corresponds with the conclusion that the *low* prompting conditions are insufficient to sustain a radiologist's interest, it is clear that radiologists are able to tolerate – and actually prefer – a high prompting rate.

Rating the system

- **Q12 Do you believe that the system and its component parts would be useful in a screening context in their present state (yes, no) (figure 3.7)?**
- **Q13 Rate the system and its component parts based on its sensitivity (too low, OK, too high) (figure 3.8).**

What is immediately obvious from these two sets of graphs is that the outcome of the 'system' is identical to the 'masses' outcome. It would appear that the radiologists either feel that the masses part is the most influential or that it is the worst performing component that decides the overall feel to the system. The microcalcification results are encouraging, with three of the four radiologists believing that this component is suitable for screening usage at its *medium* setting. The change in opinions under the *high* condition is probably due to the increase in vascular calcification. Should it prove possible to remove the vascular calcification prompts before being seen by a radiologist, it is likely that the *high* condition would be the most satisfactory. Similarly, in figure 3.8, two radiologists believe that the sensitivity of the microcalcification component is 'OK' at the *medium* level, whereas one of these becomes 'too high' at the *high* condition. Indeed, one opinion of the mass algorithm also became too high at the *high* prompting condition. It was felt by the team, in discussion with the radiologists, that was due to the difficulty in distinguishing between a high sensitivity (lots of correctly identified features) and low specificity (lots of untargeted prompts), since at this stage radiologists are unable to distinguish between cancers and suspicious features.

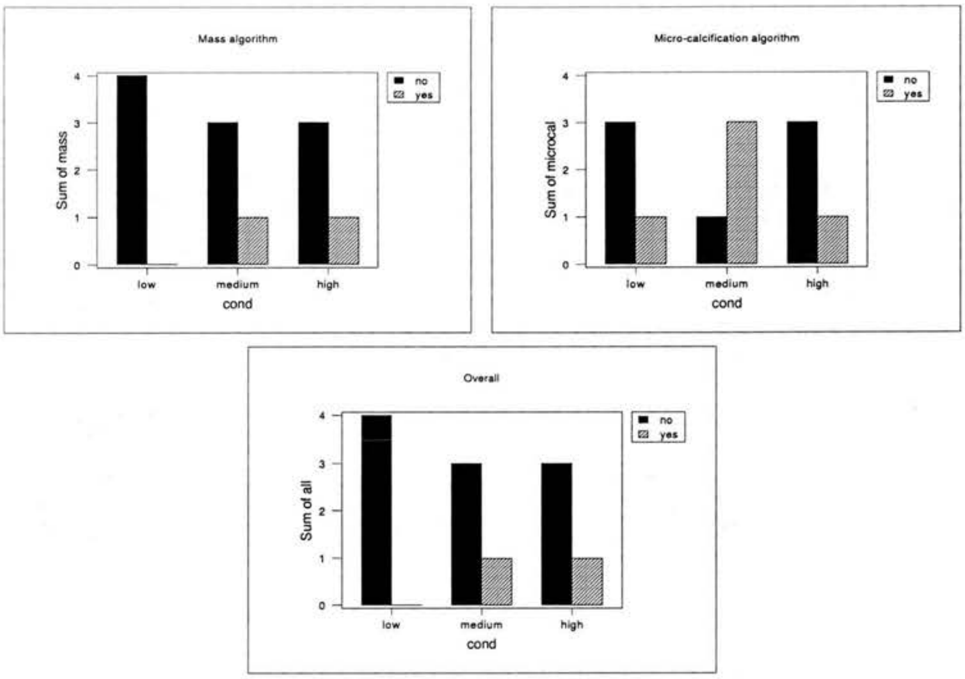


Figure 3.7: The effectiveness of the system at each prompting rate

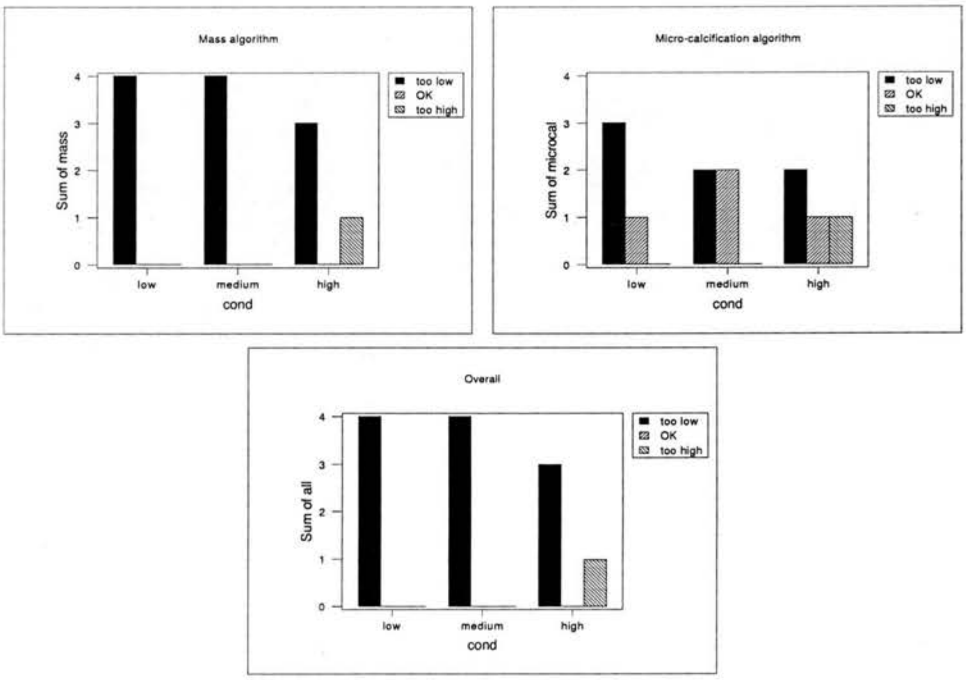


Figure 3.8: The sensitivity of the system components

- **Q14 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 microcalcification clusters being detected). Please tick one box per sensitivity setting. (figure 3.9)**

This question was subject to a great deal of misinterpretation. All four participating radiologists believed that 95% sensitivity was very useful, thereafter opinion was divided. One radiologist believed that the *low* condition was very useful at all levels, despite believing that the *high* condition would be only useful at 85% and 80%.

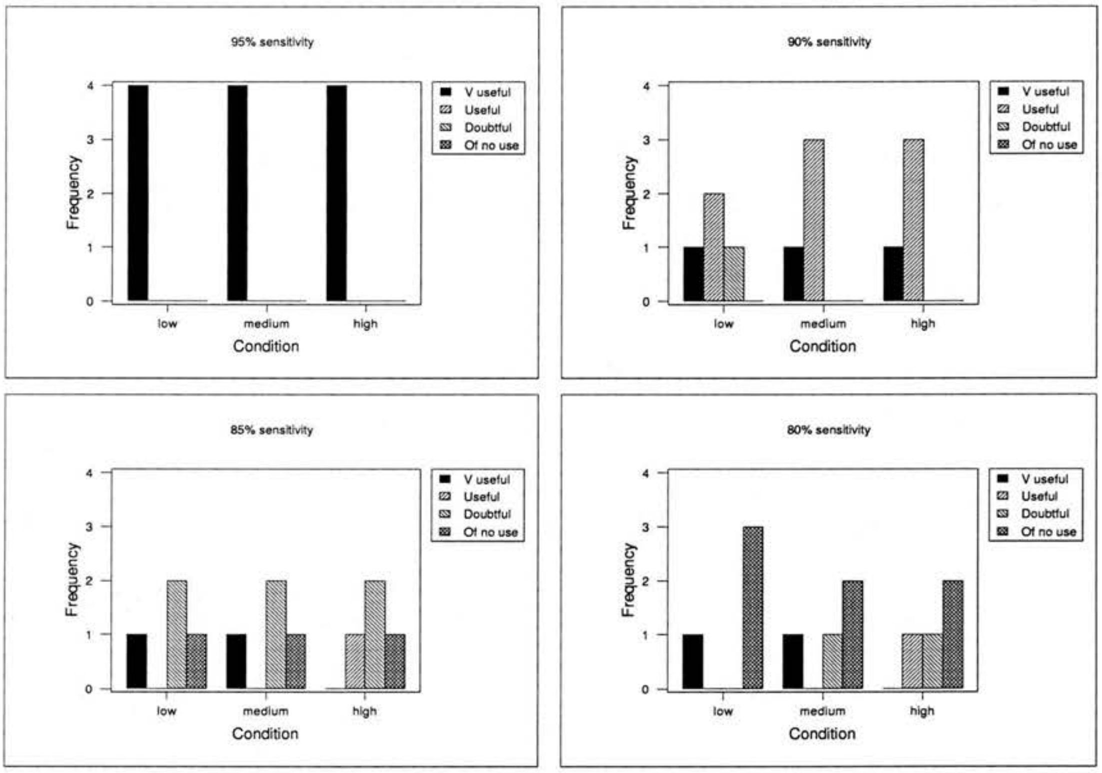


Figure 3.9: Rating the system sensitivities

3.8 Further work

As an extension to this experiment, the prompts from two of the sets were shown to an additional three radiologists, who rated these by whether they were a true prompt, whether they would want to see this feature prompted and how much of an annoyance the prompt was. This work is reported in depth in Mark Hartswood's thesis [91] and at the 1998 Medical Image Understanding and Analysis Conference in Leeds [93]. It will not be reported here.

3.9 Conclusion

3.9.1 Recall Rates (section 3.4)

There is no significant increase in the recall rates as the level of prompting increases. The difference between the *medium* and *null* conditions is only significant when the hazards of multiple testing are ignored.

3.9.2 Time taken to complete the experiment (section 3.5)

As the prompting level increases, so does the average time taken to complete a session. This is fairly intuitive, as it will take longer to investigate a larger number of prompts than a smaller number. However, the increase is only of the order of 26% (calculated from the means at *null* and *high*).

3.9.3 Observational data (section 3.6)

Contrary to previous expectations, radiologists lose interest in the prompts when only a few are generated. A reasonably high prompt rate will sustain their interest, provided that the prompts are of sufficiently high quality.

3.9.4 Questionnaires (section 3.7)

The most obvious conclusion from this section is that the radiologists feel that the ill-defined lesion algorithm in its current form is unsatisfactory, and that the microcalcification algorithm would be better served by the removal of vascular calcification prompts. The radiologists involved were agreed that the higher sensitivity and prompting rate would be the most useful, with the lower sensitivity/prompting rate being considered virtually useless.

3.9.5 Summary

Before beginning this experiment, the accepted wisdom was that only the low prompt rate would be acceptable to screening radiologists. However, this experiment has demonstrated that false prompts are well tolerated if they are 'sensible'; in other words, that the radiologist is able to rationalise the prompt. This will allow the algorithm developers to risk excessive prompts, if this raises the sensitivity of the system.

Chapter 4

The Pre-Clinical Trial

4.1 Aims and Objectives

This chapter describes the 'pilot study', exposing the radiologists at the Edinburgh Screening Centre to the improved algorithms in a more typical setting than in the previous chapter. Twenty batches of films from 100 women were read by pairs of radiologists, who were randomly assigned to be prompted or unprompted. The batches were biased, with 102 cancers in a total of 2002 sets of films, rather than the 12 that would have been expected with a 0.6% underlying cancer rate.

This experiment was to be the first true test of the system in a clinical setting. Although the processing of the films had to be performed off-site for technical reasons, all other aspects of the protocol were kept as close as possible to the human-factors design proposed by Dr Mark Hartswood and Dr Rob Proctor. Thus, the aim was to see how the system and radiologists interacted in as normal a situation as possible.

Initial examination of the results indicated that the recall rate declined with the introduction of the prompting system. As this was contrary to expectations, it was, at first, a surprise and rather worrying. Later, more detailed, models accounted for much of the difference as a radiologist effect. Cancer detection was slightly lower when prompted, although analysis showed that of the eleven cancers missed by the prompted radiologists, six had been correctly prompted by the algorithms. The results of this experiment were reported at the 4th International Workshop on Digital Mammography [94][95].

4.2 Introduction and Background

Prior to the start of a full-scale clinical trial, there were some concerns that the introduction of PROMAM into clinical usage as part of controlled trials might adversely affect recall rates in participating centres. The Directors of the clinics who had agreed to participate were anxious that the recall rate would not exceed a level where assessment clinics were unable to cope, thus causing a 'knock-on' effect to the rest of the service.

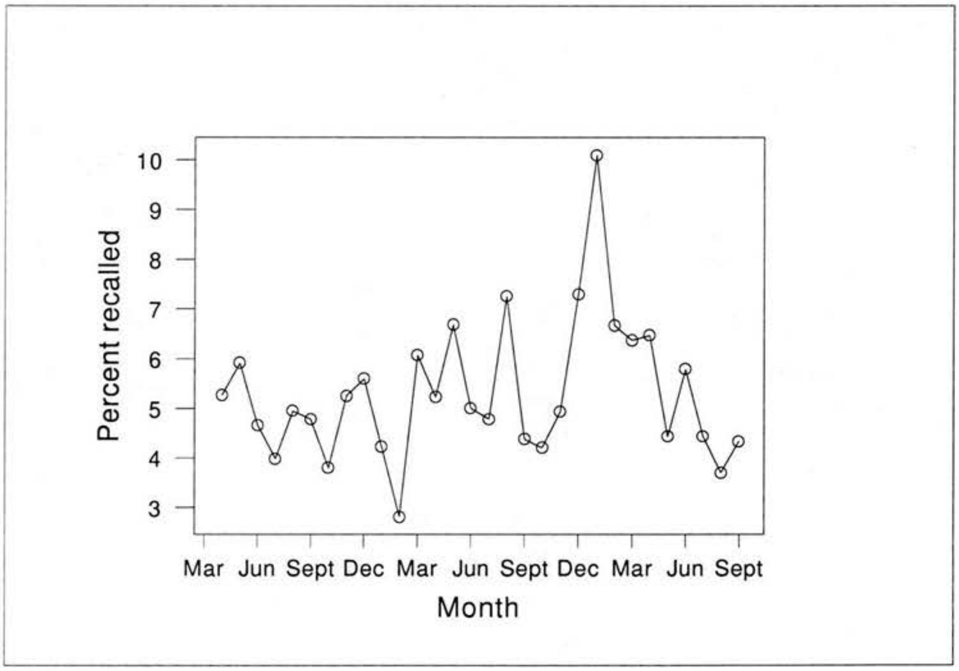


Figure 4.1: Recall rate over April 1993 - Sept 1995 at Ardmillan House, Edinburgh

Consensus by the trial clinic directors appeared to indicate that the recall rates should not increase by more than 10% of current recall rates. Unfortunately, this is somewhat difficult to monitor adherence to in the short term, due to the naturally oscillating recall rate (figure 4.1)

Possible causes of this natural variation include the proportion of first time attendees in the screened population and the area being targeted in a particular month. However, there were further worries that the conversion to blind double readings and the addition of the prompting system would push the recall rate even higher.

Because of these concerns, it was decided that we would conduct a pre-clinical trial as a means of reassuring the Directors, as well as providing vital information on the performance of the system under screening conditions.

4.3 The Pre-clinical Design

This experiment was designed to examine the TP rate under near-normal conditions (where radiologists are unaware whether a film is a malignancy or a non-malignancy), and also to examine the change in recall rate that might be expected during the full trial phase. Due to time and space pressures in Ardmillan House, it was, unfortunately, not possible to conduct this trial on throughput as initially hoped, but instead we asked the radiologists at Ardmillan to blind double read a set of 2002 cases. This set was biased so as to include a measurably large set of malignancies in order for the number of cancers to be sufficient for a reasonable investigation, while still simulating a typical screening session. As was anticipated to be the practice for the full trial, cases were assigned to radiologists on a block minimised method, to ensure that radiologists read equal numbers of prompted/unprompted and first/second sets. This took place outwith normal working practice to limit the disruption to clinic staff as much as possible. From the previous experiment[1], we anticipated that radiologists would lower their threshold during the experiment (as opposed to during screening) and 'recall' more than they would normally. Thus, we decided that the only comparison of relevance was the comparison between the prompted and unprompted radiologists.

[1] Chapter 3

4.4 Experimental Procedures

Due to technical problems with the scanner (described in section 4.5), the duration of the trial was reduced from eight weeks to five weeks, although the same number of cases were processed as had been planned for the longer run. Because of this, the trial protocol was altered very slightly to drop the requirement that feedback be given to radiologists during the experiment. This was to have taken the form of recall rates to date, and mock review clinics. In the latter case, it would have taken the place of a reading session, and there were none to spare.

With the limited time allowed for the experiment, and the requirement that "we take what we could", in terms of radiologist availability, it was not possible to get a complete set of radiologist pairings in either condition. As the experiment was conducted in July and August 1997, many of the radiologists were on holiday during at least part of the experiment. Had more time been available, it would have been possible to select the pairings in such a way that each radiologist read with each of the others. However, this was unachievable in the time available, and so whichever radiologists were available at the time were utilised.

Based on analysis performed by the algorithm developers, the microcalcification algorithm had a sensitivity of 90% with a prompt rate of 1 in 4 women prompted, and the ill-defined lesions algorithm had a sensitivity of 80% with a prompt rate of 1 in 2 women prompted. Both algorithms were measured on test sets prior to the experiment.

4.5 Compilation of the experimental set

With 2002 sets of mammograms, over 5000 films in all, to be digitised by the scanner, it was not considered feasible to leave it in its then current location in Ardmillan House, as the technical design of the scanner required that films be fed into it manually. It had been hoped that a newer model with a hopper would be made available to the team in time for the experiment, but this was not to be. The scanner was taken to the Royal Observatory, Edinburgh (ROE), where the algorithm developers would be close at hand in the event of any problems. Films were retrieved from the archive at Ardmillan House by Pat Dixon, the project radiographer, scanned, and replaced.

Difficulties arose when, after the move to ROE, there was a brief period when the scanner would not calibrate. Fortunately, this was resolved in short order, and scanning could begin. Scanning took place between 15 May 1997 and 11 June 1997, with the algorithms applied as each digitised image appeared in the database. A problem with the software meant that the first set of films produced unacceptable prompts, and the trial was delayed by three weeks while the algorithm developers tracked down the problem.

4.5.1 Compilation of malignancies

Malignancies were selected from the archives at Ardmillan House between the periods of November 1995 and April 1997. Given the low frequency of cancers in the population, this was the maximum possible number available, since prior to the earlier date, the quality of the films was not high enough for the algorithms to read successfully. Cases where the cancer occupied a large fraction of the breast were excluded, since the ill-defined lesion algorithm was unable to target these. Other exclusions included those lesions classified as 'probably benign' at assessment for whom the pathology information was not included at time of selection. 105 cancers were selected in this manner, although three of these were subsequently removed from the test set, two due to problems with the scanner, and another through not being located in the archive. It was believed that this was due to the CHI number being mis-entered into the database.

microcalcification only	33	32%
ill-defined lesion only	47	46%
both	22	22%

Table 4.1: Cancers in the experiment

Table 4.1 shows the distribution of the cancer types used in the experiment. Ill-defined lesion covers all opacities – in other words, all malignancies that were not microcalcification alone. 'Both' refers to the presence of both microcalcification and an ill-defined lesion on the malignant site.

4.5.2 Compilation of non-malignancies

With each cancer selected, 19 non-malignancies were also randomly drawn from the same date. This gave us 102 sets of 20, a total of 2040 sets of films in total, of which 1938 were non-malignancies. Each group of 20 was divided into two sets; ten cases containing the cancer, another ten without a cancer. These were then referred to as the positive set and the negative set. Hence, a batch of 100 was created by assigning appropriate numbers of positive and negative sets. For example, if a batch was called upon to contain four cancers, it was constructed from four positive sets and six negatives sets. The only stipulation required was that none of the sets must come from the same day as another.

4.6 Randomisation of cancers to radiologists

Cases were assigned to batches of approximately 100, which included a randomly decided number of cancers (based on a mean of 5). Radiologists were assigned to batches using a minimisation method that attempted to ensure that each radiologist saw approximately equal numbers of cancers in prompted and unprompted sessions.

Some of the malignancies available for this experiment had already been annotated by the radiologists, for the purpose of testing and training the algorithms. Given the modest length of time since collection began and a radiologist's memory for cancers, it was considered a potential source of bias let a radiologist who annotated a film be one of the readers for that case in the experiment. Since the team possessed the information regarding the identity of the radiologists who annotated, assessed and screened each film, a simple formula was created to minimise prior exposure to such cases.

Each type of exposure to a case was given a score; screening = 2, assessment = 3 and annotation = 4, based on the idea that with each increasing score, the case is one among an increasingly smaller sample. In the case of screening, it is one of many hundreds that a radiologist sees, in assessment, it is still only one tenth of the cases that he or she might see, but by annotation, they are aware that this is a cancer and may remember the case accordingly. These points were additive. The lowest scoring cancers for the available radiologists were then assigned to that batch. Where possible, cancers with no prior exposure to the radiologists in question were used.

4.7 Briefing Radiologists

The principal communication between the radiologists and the PROMAM team directly involved in this experiment was three-fold; an initial briefing and training session, a pre- and post-experiment questionnaire (see Appendix H.1 and H.2), and a questionnaire that was to be completed subsequent to every prompted reading session (see Appendix H.3). That is, for every batch of cases, only the prompted radiologist was asked to complete the questionnaire. These questionnaires were similar in design and content to those given during the subjective reaction experiment (Chapter 3). As before, the content of the questionnaires was composed by Mark Hartswood.

The training session consisted of a series of examples of the system's behaviour when presented with real cases. Illustrations for each of the three important potential results (True Positive, False Positive, and False Negative - when a feature that should have been prompted wasn't) were shown to the radiologists, along with explanations as to why the system had behaved in such a way. Examining the data from the questionnaires and semi-structured interviews later, it may be that this training exercise was not as broad or complete as we would have hoped. This is further discussed in Mark Hartswood's PhD thesis [91].

4.8 The experimental set of cancers

Although the rationale behind wanting to know the history of each cancer was to ensure that no radiologist encountered a cancer to which they had had prior exposure, much interesting information may be gleaned from the data in their own right. In this section, only the results that were obtained when the cancers initially passed through the screening process will be discussed.

Radiologist	Screened (R_1, R_2)	Assessed	Annotated
R_A	63 (40, 23)	39	9
R_B	52 (27, 25)	39	7
R_C	21 (3, 18)	1	0
R_D	58 (29, 29)	18	24
R_E	7 (2, 5)	4	0
Total	201 (101, 100)	101	40

Table 4.2: Exposure at screening, assessment and annotation for each radiologist in the experiment

It may be noticed that, in table 4.2, although it has been stated that there were 102 cancers in the experimental set, only 101 are seen to have been read first (R_1), and only 100 have been read second (R_2). This is due to there being no historical data available for one case (hence only 101 being 'seen'), and one other case was only single read, hence the discrepancy between first and second readers. Radiologists C and E had only recently joined the Edinburgh Breast Screening Centre when this experiment took place, hence the much reduced numbers of screening and assessment cases. All annotations took place prior to these radiologists joining the clinic.

		Second reader		
		Recalled	Not recalled	Total
First reader	Recalled	82	4	86
	Not recalled	14	0	14
	Total	96	4	100

Table 4.3: Numbers of cancers recalled by first and second readers when initially read in normal clinic practice

From table 4.3, we may get a first estimate of the agreement between radiologists when reading under screening conditions. The article by Williams *et al* [57] suggests several methods of calculating inter-observer agreement, recommending that kappa, positive agreement and negative agreement be reported during studies into the agreement between radiologists. However, kappa is not a particularly useful measure in screening mammography, as it is influenced by the prevalence of the disease, approaching zero as the prevalence approaches zero or one. As the appearance of breast cancer in a screening population is approximately 0.6%, any kappa values derived would be subject to some suspicion. Positive agreement,[2] however, is 0.9. Negative agreement[3] is meaningless in this case, as this table deals only with known cancers. Any cancers missed by both radiologists are not included.

As with many Breast Screening Centres, the radiologists at Ardmillan House are not actively blinded to the decision of the first reader. From table 4.3, it can be inferred that this is influencing the recalls made by the second reader, as there are significantly fewer recalls made by the first reader that are not also made by the second reader than the converse $(P(X \leq 4|X \sim B(18, 0.5)) = 0.03)$ 2-sided).

In Williams, Hartswood and Prescott [31], an alternative measure to the mean second screener contribution (see page 11) was proposed to examine the increase attributable to the second reader. Using this, we see that the increase due to the second reader is 16.3%, and the standard error of the $\log_e$ of the increase is 0.29. Back transformed, this gives a 95% CI of (9.3%, 28.6%).

Sensitivity between radiologists varies greatly within this test set of cancers, and also within radiologist, depending on whether they are reading first or second (table 4.4). These figures cannot be taken as completely accurate, as they do not take into account the cancers that were missed by both readers (i.e. false negative interval cancers).

[2] number described as abnormal by both/mean number described as abnormal

[3] number described as normal by both/mean number described as normal

Radiologist	Reading	Recalled	Not recalled	Sensitivity
R_A	First	37	3	92.5%
	Second	23	0	100.0%
	All	60	3	95.2%
R_B	First	25	2	92.3%
	Second	23	2	92.0%
	All	48	4	92.3%
R_C	First	3	0	100.0%
	Second	18	0	100.0%
	All	21	0	100.0%
R_D	First	20	9	69.0%
	Second	27	2	93.1%
	All	47	11	81.0%
R_E	First	2	0	100.0%
	Second	5	0	100.0%
	All	7	0	100.0%

Table 4.4: Recalls per radiologist made on the set of test cancers

4.9 Data collection

The pre-trial experiment was not only an opportunity to observe the algorithm performance in a screening setting, but to also give other aspects of the full trial a dry-run. For example, the team needed to know if the scanner could cope with the volume of through-put that a clinic would be expected to see during a day. Despite the scanner not being totally ideal for this type of continuous work, and the technical trouble we had at the beginning of the experiment (as described earlier), it performed reasonably well, although it needed to be monitored at all times, which is hardly suitable for a busy screening clinic. However, the full trial was anticipated to run using the next generation model, which included a hopper to load many films at once.

Another aspect of the trial that it was possible to test was the database. The database used in this experiment was a simplified version of the one to be used in the full trial, since no information beyond the level of screening was required. Hence, the assessment, cytopathology and pathology forms were rendered redundant, and so were left out of this database. However, this database did hold information that would not be available in the full trial, namely which cases were the cancers, what type of cancers pathology had revealed them to be, and which radiologists had seen them at each stage of the process.

During the experiment, data were entered into only two tables; the batch information and the results tables.

4.9.1 Batch Information

Batch number	**to marry batch information and subject information**
Batch size	the number of cases in each batch
Reader 1	first reader ID
Reader 1 prompted?	whether the first reader had the prompts
Reader 2	second reader ID
Reader 2 prompted?	whether the second reader had the prompts

4.9.2 Results form

ID	**subject ID (case identifier)**
Batch number	**to marry batch information and subject information**
Reader 1 recalled?	the decision made by the first reader regarding this case
Abnormality prompted?	was the recall site correctly prompted?
Reader 2 recalled?	the decision made by the second reader regarding this case
Abnormality prompted?	was the recall site correctly prompted?

The information in bold is the key identifier in each table. This key must be unique, either on its own (as in the batch table) or in combination (the two keys combined in the results table give a single unique identifier).

The trifurcated line from the batch table to the experimental results table (figure 4.2) indicates a 'one-to-many' relationship. That is, for every unique key identifier in the

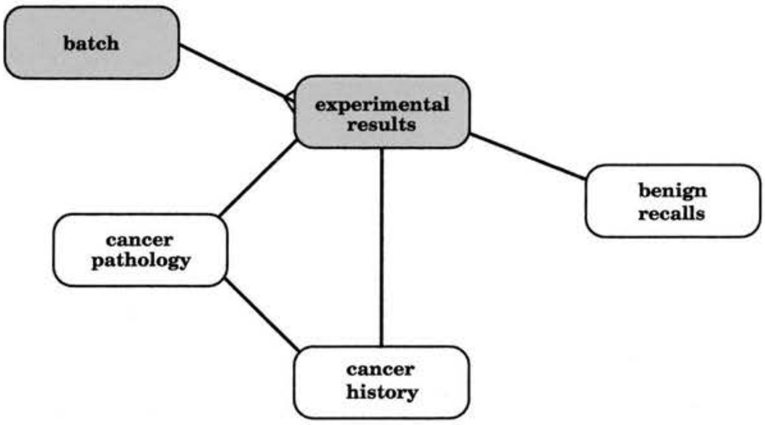


Figure 4.2: The relationship between the tables in the pre-trial experiment database

batch table (batch number), it is matched to an identical identifier that may appear several times in the 'many' table. For example, batch number 3 appears only once in the batch table, but it appears 100 times in the results table. However, each combination of batch number and subject ID is unique.

The shaded ellipses indicate which tables had forms that were used to input the data (see Appendix I). The unshaded ellipses were unchanged by the experiment.

The field 'Abnormality prompted' was where the prompted radiologist was asked to indicate (on any case that they recalled), whether they thought the system had prompted for the feature that had caused the recall decision. From this, it was possible to calculate the perceived accuracy of the algorithms.

4.10 Results

As mentioned earlier, radiologist pairings were limited by the availability of the radiologists. Table 4.5 shows the final tally. Unfortunately, it was not possible to ensure that each available pairing happened at least once, but this should not be a problem in the longer-running multi-centre trial.

	unprompted				
prompted	R_A	R_B	R_C	R_D	R_E
R_A			2		1
R_B			2	1	
R_C		1			3
R_D	2				1
R_E	1	4		2	

Table 4.5: Pairings of radiologists in the pre-clinical experiment

4.10.1 Simple cancer detection and recall rates

Due to the nature of the experiment, it was decided to divide the results into two separate sections; the cancers, and the recalls. The cancer results all derive from the 102 pathology proven cancers, whereas the recall results refer to the 1900 cases that were not pathology proven cancers (normals and benign recalls).

In algorithm terms, a case was deemed to be prompted if there were at least one prompt on any film in the case. The prompted recalled case was deemed to be correct if and only if the suspicious feature (the feature that had inspired the recall) was correctly prompted.

From earlier work, it was believed that the recall rate would increase when radiologists perform under experimental conditions, when there is less pressure to keep the recall rate low, and, as mentioned earlier, the data would be examined in terms of prompted versus not prompted, rather than as absolutes.

4.10.1.1 Overall results

Below are two summary tables, classifying the cancer detection and recall rates by prompting and order.

Prompted	Cancer detection		Recall rate	
Yes	91/102	89.2%	134/1900	7.0%
No	92/102	90.2%	163/1900	8.6%

Table 4.6: Cancer detection and recall rates by prompting

Obviously, there is no significant difference between the prompted and unprompted readers in the case of cancer detection rates (table 4.6). There is, however, a significant difference between the recall rates of the prompted and unprompted readers (see page 95). The fact that, overall, the prompted radiologists had a lower recall rate than the unprompted radiologists was unexpected; the theoretical model of this experiment suggested that the recall rate for the prompted radiologist would rise due to extra features - that the radiologist missed - being brought to their attention. This interesting result will be examined again later in this chapter.

In table 4.7, only the prompted results are examined, to see if there is a significant difference between a prompt being given to the first reader or to the second.

Order	Cancer detection		Recall rate	
First	47/50	94.0%	68/948	7.1%
Second	44/52	84.6%	66/953	6.9%

Table 4.7: Cancer detection and recall rates by order of prompting

Despite appearances, there is no statistically significant difference (by χ^2 test) between either the cancer detection rates or between the recall rates. However, in the case of the cancer detection rates, there may be too few cases to determine if this is a true effect or merely a lack of statistical power.

On an individual radiologist basis, recall rates were not unusually high, with a higher TP rate coupled with a higher recall rate (see table 4.8). This is hardly surprising; the more that are recalled, the more likely it is to discover a cancer amongst them.

Radiologist	Cancer detection		Recall rate	
R_A	27/32	84.4%	34/562	6.0%
R_B	31/36	86.1%	54/770	7.0%
R_C	45/46	97.8%	82/757	10.8%
R_D	25/28	89.3%	49/572	8.6%
R_E	55/62	88.7%	78/1139	6.8%

Table 4.8: Overall cancer detection and recall rates

The figure below (figure 4.3) illustrates the approximately linear relationship between recall rate and cancer detection rate. This simple relationship, however, becomes somewhat more complex when prompting is taken into account (figure 4.4).

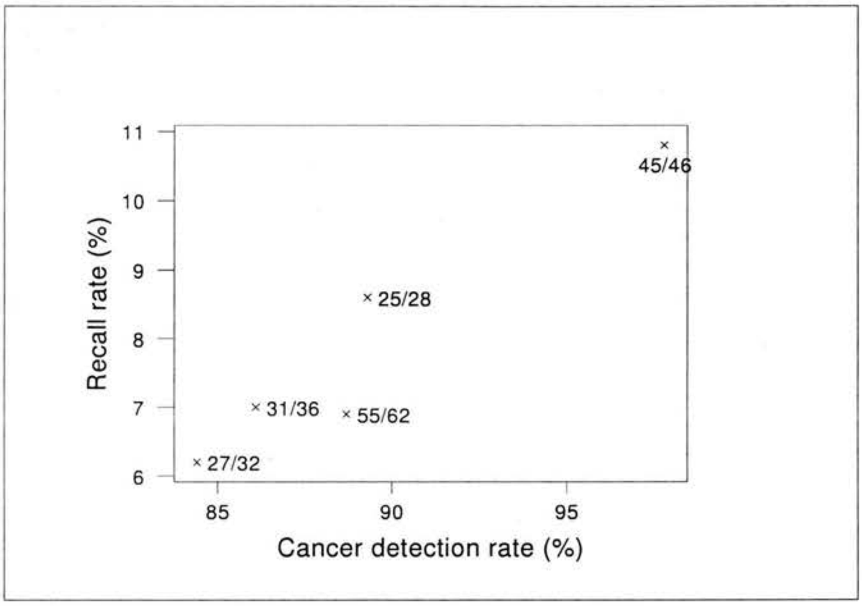


Figure 4.3: Overall recall rate by cancer detection

Radiologist/condition	Cancer detection		Recall rate	
R_A unprompted	10/13	76.9%	16/285	5.6%
R_A prompted	17/19	89.5%	18/277	6.5%
R_B unprompted	20/24	83.3%	36/478	7.5%
R_B prompted	11/12	91.7%	18/292	6.2%
R_C unprompted	23/24	95.8%	40/375	10.6%
R_C prompted	22/22	100.0%	42/382	11.0%
R_D unprompted	13/15	86.7%	32/288	11.1%
R_D prompted	12/13	92.3%	17/284	6.0%
R_E unprompted	26/26	100.0%	39/474	8.2%
R_E prompted	29/36	80.6%	39/665	5.9%

Table 4.9: Cancer detection and recall rates by prompting

Figure 4.4 illustrates the individual recall and cancer detection rates for each radiologist under prompted and not prompted conditions (as seen in table 4.9). In most cases, the results comply with the hypothesis that prompting improves the cancer detection rate, despite a mixed response on the recall rates. The one counter result comes from radiologist E, who has a cancer detection rate of 100% when unprompted which falls to 80.6% when prompted. However, since the set of cases that were prompted is not the same set as those unprompted, there may be other factors at work; for example,

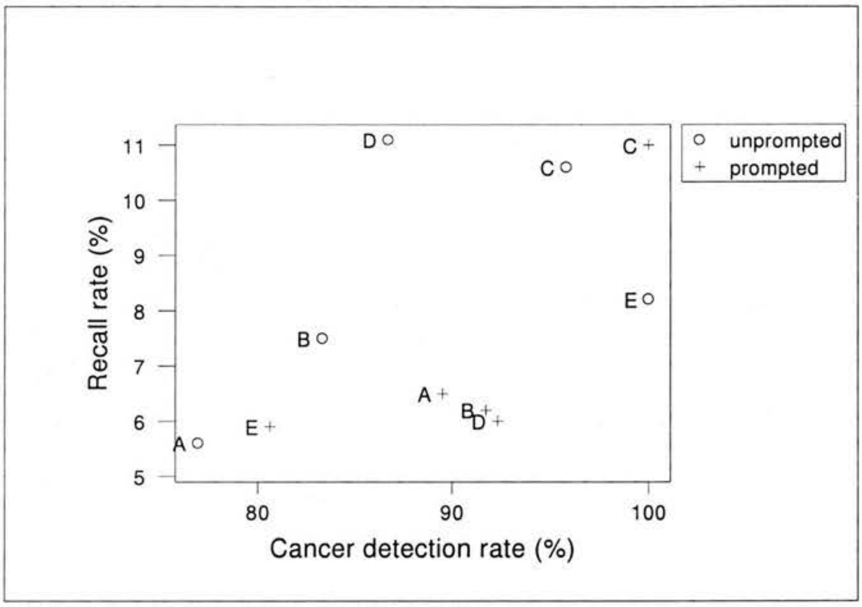


Figure 4.4: Recall rate by cancer detection, by prompting

the cancers in the prompted set may have been more subtle and hence harder to find.

4.10.1.2 Prompting effect

As stated earlier, the main goal of this experiment was to determine whether or not there were any differences between the cancer detection and recall rates when a radiologist was prompted and when they were reading unaided. The most efficient way of examining the data for this type of result is McNemar's Paired Binary Test.

		Unprompted		
		Recalled	Not recalled	Total
Prompted	Recalled	86	5	91
	Not recalled	6	5	11
	Total	92	10	102

Table 4.10: Cancers detected by the prompted and unprompted radiologists

Cancers It is the off-diagonal cells in table 4.10 that are of interest in determining whether there is a statistical difference between the cancer detection rates. The usual method, when sample sizes are small, is to use the Binomial distribution $B(n, p)$ where

n is the sum of the off-diagonal cells and $p = 0.5$. So, in this case, we are testing $P(X \leq 5) + P(X \geq 6)$, where $X \sim B(11, 0.5)$. This gives a p-value of 1; not significant. This disheartening result can be mostly attributed to batch 17, where the unprompted radiologist recalled all six cancers that were present in the set, and the prompted radiologist recalled only three (see Appendix J). It is not known why the prompted radiologist missed these cancers, especially in light of the fact that they were correctly prompted (see Appendix K).

Of special interest, however, is the fact that even under such artifical conditions, nearly 5% (5/102) of the cancers were missed by both readers. Since none of the test cancers had been drawn from the set of interval cancers, it is obvious that factors other than the difficulty of detection were causing cancers to be missed.

From table 4.10, it is possible to calculate the agreement of cancer detection between the prompted and unprompted radiologists as 93% (no. recalled by both/no. recalled by unprompted reader, as in section 2.5.8), which can be used to calculate an improved estimate of the sample size required for a full-scale trial.

Recalls Table 4.11 illustrates the decisions made on the set of 1900 non-malignancies. In normal practice, films are classified into three types; recalls (for immediate recall and further exploration), tech recalls (recalled for a technical reason, e.g. poorly developed film) and normal (considered to be non-malignant and not recalled).

		Unprompted			
		Recalled	Not recalled	Tech recall	Total
Prompted	Recalled	65	69	0	134
	Not recalled	97	1650	10	1757
	Tech recall	1	7	1	9
	Total	163	1726	11	1900

Table 4.11: Recalls made by the prompted and unprompted radiologists

Again, we will consider the data as paired binary, ignoring the technical recalls for the moment. When the sum of the off-diagonal cells ($n = b+c$) becomes large, (69+97=166, in this case), it is easier to calculate an approximation to the Binomial, rather than the exact result, since the Binomial tends to the Normal (by central limit theorem). With the usual continuity correction, $P(R \leq r) = P(X < r + \frac{1}{2})$ where $X \sim N(np, np(1-p))$. From this, we find the p-value for the difference between 134 false recalls and 163 false recalls in table 4.11 to be 0.02; significant at the 5% level. Hence, there is a significantly lower false positive rate in the prompted condition than in the unprompted condition. If we also include the prompted tech recall/unprompted recall with the cell above, then the difference is even more pronounced. However, there are many factors involved in producing the result, and this should not be taken as a straightforward prompted versus not prompted outcome until these other factors have been investigated.

One possible explanation for this lower recall rate in the prompted cases is that the radiologists are using prompt information to classify a suspicious feature; for example, the absence of a prompt may make them less inclined to recall. This will be examined in more detail later, when the results of the questionnaires are discussed.

Agreement on recalls between radiologists for the set of non-malignancies is considerably lower than that on the set of cancers (40%), once again illustrating the widely different criteria that radiologists appear to have when deciding whether a feature is suspicious enough to recall.

4.10.1.3 Order effect

As an aside, we are also interested in whether reporting first or second has any influence on cancer detection and recall rate when the radiologists are blinded to the first reader's decisions.

Cancers In this instance (table 4.12), the difference between first and second reader is non-significant ($p = 0.23$), suggesting that if the second reader is blinded to the first reader's decision, there is no difference between them. However, the point estimate and standard error of the difference (0.049 and 0.0322, respectively) show that the power of the test is too low to detect a difference in means, should one exist.

		Second reader		
		Recalled	Not recalled	Total
First reader	Recalled	86	8	94
	Not recalled	3	5	8
	Total	89	13	102

Table 4.12: Cancers detected by first and second readers

When compared to the historical data (see section 4.8, table 4.3), however, there is a noticeable change, as that showed a significant difference between first and second readers ($p = 0.03$), with the second reader appearing to detect a significantly higher number of cancers. In this blinded experiment, however, there is no significant difference between the reading order, and the first reader detected more cancers than the second. This would suggest that there is a distinct order effect when the second reader is not blinded to the decisions of the first.

Recalls Using the central limit theorem as before, we find the p-value for the difference between the 152 false recalls and 145 false recalls in table 4.13 to be non-significant. Again, this supports the hypothesis that there is no order effect, provided that the second reader has no access to the first reader's decisions.

		Second reader			
		Recalled	Not recalled	Tech recall	Total
First reader	Recalled	65	86	1	152
	Not recalled	80	1650	6	1736
	Tech recall	0	11	1	12
	Total	145	1747	8	1900

Table 4.13: Recalls made by first and second readers

4.10.1.4 Comparison of results with controls

As mentioned before, the results in table 4.9 were initially quite worrying, especially the results from radiologist E, where this radiologist was 100% accurate when not prompted, and only 80.6% accurate when prompted. However, the cases used in these comparisons were not the same, and it is known that batches can vary considerably. Hence, the comparison was made, not between prompted and unprompted within radiologist, but between prompted radiologist and the control readers; the radiologists who saw the same cases, only unprompted.

Radiologist	Prompted		Control (unprompted)	
R_A	17/19	89.5%	19/19	100.0%
R_B	11/12	91.7%	11/12	91.7%
R_C	22/22	100.0%	21/22	95.5%
R_D	12/13	92.3%	11/13	84.6%
R_E	29/36	80.6%	30/36	83.3%

Table 4.14: Cancers by each prompted radiologist

Prompted cancers As we can see from table 4.14, radiologist E's surprisingly low prompted sensitivity is probably more due to the cancers in those sets being difficult to diagnose, rather than any shortcomings on the part of either the prompting system or the radiologist. However, four of the cancers missed by E/prompted were actually highlighted by the prompting system (see Appendix K). Sadly, the numbers involved are really too small to examine on a radiologist by radiologist basis.

Radiologist	Prompted		Control (unprompted)	
R_A	18/278	6.5%	36/278	12.9%
R_B	18/292	6.2%	30/292	10.3%
R_C	42/382	11.0%	27/382	7.1%
R_D	17/284	6.0%	16/284	5.6%
R_E	39/665	5.9%	54/665	8.1%

Table 4.15: Recalls by each prompted radiologist

Prompted recalls As with the cancers, some of the more surprising results in table 4.15 can be explained by the differences in batches; radiologist D's reduction from 11.1% false recalls when unprompted to 6.0% when prompted is less startling when the prompted recalls are compared to the recalls made by the unprompted control readers. There are still differences between prompted and unprompted when viewed this way, implying that other factors beyond batch differences are having an effect. These other factors will be discussed later.

Since the inspection of the prompted results yielded some interesting conclusions, it was decided to also examine the unprompted condition in a similar manner.

Unprompted cancers Again, radiologist E's disparate results appear to be mainly the effect of batch differences, with the 100% accuracy when unprompted matched by the prompted control readers (table 4.16).

Radiologist	Unprompted		Control (prompted)	
R_A	10/13	76.9%	11/13	84.6%
R_B	20/24	83.3%	19/24	79.2%
R_C	23/24	95.8%	21/24	87.5%
R_D	13/15	86.7%	14/15	93.3%
R_E	26/26	100.0%	26/26	100.0%

Table 4.16: Cancers by each unprompted radiologist

Unprompted recalls In this case (table 4.17), it would appear that batch is not the most influential factor when looking at the recall rate for radiologist D, as it is nearly twice the size of the prompted recall rate. However, radiologist D is paired with radiologists B and E, both of whom have lower average recall rates than D (see tables 4.8 and 4.9). Radiologist C has a consistently higher recall rate than his/her fellows (especially A and B, the radiologists reading the prompted cases), so the larger recall rate is not particularly surprising. This larger recall rate is offset by radiologist C possessing the most accurate cancer detection rate (see figure 4.3).

Radiologist	Unprompted		Control (prompted)	
R_A	16/285	5.6%	18/285	6.3%
R_B	36/478	7.5%	29/478	6.1%
R_C	40/376	10.6%	24/376	6.4%
R_D	32/288	11.1%	19/288	6.6%
R_E	39/474	8.2%	44/474	9.3%

Table 4.17: Recalls by each unprompted radiologist

4.10.2 Algorithm results

In addition to the usual recall/not recall to which radiologists were accustomed, they were asked to indicate on the form whether a case that they had chosen to recall had been correctly prompted by the system.

4.10.2.1 Cancers

	correctly prompted	not correctly prompted	total
Recalled by both	72	14	86
Recalled by prompted only	5	0	5
Recalled by unprompted only	4	2	6
Recalled by neither	2	3	5

Table 4.18: Summary of cancers, by algorithm result

As can clearly be seen in table 4.18, six of the eleven cancers that the prompted radiologist passed as normal (recalled by unprompted only and recalled by neither) were actually correctly prompted by the system. The instances where the cancer was recalled by both is not strictly of interest, in terms of the PROMAM system, as these would be the occasions where the single reader would find the cancer with no help from the prompting system. The five cases that were recalled by the prompted radiologist but not by the unprompted radiologist are encouraging, since all five were correctly prompted (see Appendix K). It suggests (but does not prove) that these cases are instances where the prompting system highlighted a feature that the radiologist might have missed.

Overall, the sensitivity of the algorithms is 83 correctly prompted cancers out of a sample of 102. In other words, 81.4%. The perceptible accuracy (that is, the sensitivity as perceived by the radiologist, based on the cancers detected) is slightly higher, at 84.6% (77/91).

4.10.2.2 Potential improvement

If the cancers that had been correctly prompted by the algorithms and yet rejected by the prompted radiologists had actually been accepted, we would have had a cancer detection rate of 95.1% of cancers detected by the prompted radiologist against 90.2% of cancers detected by the unprompted radiologist (see table 4.19). This is still not a statistically significant difference ($p = 0.18$), despite there being a difference of nearly 5% between the conditions. However, it was never expected that significance would be achieved with such a small sample size.

		Unprompted reader		
		Recalled	Not recalled	Total
Prompted	Recalled	90	7	97
	Not recalled	2	3	5
	Total	92	10	102

Table 4.19: Theoretical cancer detection

4.10.2.3 Recalls

The sensitivity of the algorithms for radiologist defined 'suspicious features' is naturally lower (since this definition can vary considerably between radiologists) at 46.3%, with the perceptible accuracy at 56.7% (table 4.20). There is a significant difference between the proportions of correctly prompted cases recalled by the prompted radiologist alone and the proportion of those recalled by the unprompted radiologist alone (χ^2: p=0.015). This would suggest that the prompted radiologist is making use of the prompting information to inform their recall decisions.

	correctly prompted	not correctly prompted	total
Recalled by both	41	24	65
Recalled by prompted only	35	34	69
Recalled by unprompted only	31	66	97

Table 4.20: Summary of recalls, by algorithm result

4.10.2.4 Appearance of the cancers

As information on what type of malignancy each cancer was was available, it seemed wasteful not to use it. And so, a similar analysis was performed on the three basic types; ill-defined lesion (mass), microcalcification (calc) and cases where both were present (both). The results are presented in table 4.21.

The differences between the cancer types is borderline significant at $p = 0.051$, although if the instances where both features are present are removed, the differences

	Correctly prompted	Not correctly prompted	Percent correct
Calc	31	2	93.8%
Mass	34	13	72.9%
Both	18	4	81.8%

Table 4.21: Summary of cancer types, by algorithm results

become more pronounced ($p = 0.015$). The differences are unsurprising and reflect the developers' knowledge in how well the algorithms work for the different types of cancer.

4.10.3 Generalised Linear Models

The previous sections have been concerned only with tabulation and very simple analysis, which, although enlightening in some ways, has raised as many questions as it has solved. For example; why does it appear that prompted radiologists have recalled fewer cases than the unprompted radiologists? Is this an artefact of the prompting system, such that the absence of a prompt leads a reader to not recall a suspicious feature, or is there some other underlying cause? We know that radiologists have varying levels of recalls, and we should also consider the possibility of batch differences.

4.10.3.1 Cancers

As shown before, there is little difference between the prompted and unprompted conditions, and deeper analysis produces inconclusive results. A generalised linear model (PROC GENMOD) was applied as in the previous chapter, with repeated measures on the subject ID (the CHI number). In the case of the cancers, this merely reiterated the fact that there appeared to be no significant difference between prompted and unprompted conditions, and that radiologist C was significantly different to radiologist E.

The following is the results section from the analysis of recall/not recall, conditioned on subject, with the factors radiologist (RAD) and prompting (PROMAM). All cases in this set were cancers.

```
                    Analysis Of GEE Parameter Estimates
                     Empirical Standard Error Estimates

                           Empirical   95% Confidence Limits
Parameter        Estimate    Std Err       Lower       Upper       Z  Pr>|Z|

INTERCEPT          2.0407     0.4465      1.1655      2.9158  4.5702  0.0000
RAD        A      -0.4667     0.4475     -1.3439      0.4104  -1.043  0.2970
RAD        B       0.2065     0.6062     -0.9817      1.3946  0.3406  0.7334
RAD        C       1.4059     0.6166      0.1975      2.6143  2.2803  0.0226
RAD        D      -0.0452     0.6286     -1.2772      1.1867  -.0719  0.9427
RAD        E       0.0000     0.0000      0.0000      0.0000  0.0000  0.0000
PROMAM     no     -0.0026     0.3010     -0.5925      0.5873  -.0087  0.9931
PROMAM     yes     0.0000     0.0000      0.0000      0.0000  0.0000  0.0000
Scale              0.9886          .           .           .       .       .
```

The SAS output with this method compares each factor to the last in the sequence. In other words, PROMAM(no) is compared to PROMAM(yes), and is found to be not significantly different, whereas each of RAD(A), RAD(B), RAD(C) and RAD(D) are compared to RAD(E). In this case, A, B, and D are not significantly different to E, but C is (at the 5% level). However, given the problems of multiple testing, and that the asymptotic 95% confidence interval of the odds ratio is between 1.2184 and 13.6576, this result should perhaps be cautiously interpreted.

4.10.3.2 Recalls

As above, recalls were conditioned on subject, and analysed for differences between radiologist (RAD), prompting (PROMAM) and batch (BATCH).

```
                      Analysis Of GEE Parameter Estimates
                       Empirical Standard Error Estimates

                           Empirical   95% Confidence Limits
Parameter        Estimate    Std Err       Lower       Upper       Z  Pr>|Z|

INTERCEPT         -2.7047     0.3572     -3.4049     -2.0046  -7.572  0.0000
RAD        A      -0.1836     0.2065     -0.5883      0.2211  -.8890  0.3740
RAD        B       0.1356     0.1658     -0.1893      0.4605  0.8180  0.4134
RAD        C       0.5233     0.1647      0.2004      0.8462  3.1765  0.0015
RAD        D       0.2721     0.2039     -0.1274      0.6717  1.3349  0.1819
```

RAD	E	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PROMAM	no	0.1929	0.0967	0.0033	0.3825	1.9940	0.0462
PROMAM	yes	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
BATCH	1	0.7624	0.4470	-0.1138	1.6386	1.7054	0.0881
BATCH	2	0.0738	0.4564	-0.8208	0.9684	0.1618	0.8715
BATCH	3	0.0331	0.4542	-0.8572	0.9233	0.0728	0.9420
BATCH	4	-0.2718	0.5171	-1.2854	0.7417	-.5257	0.5991
BATCH	5	-0.2243	0.4662	-1.1380	0.6895	-.4811	0.6305
BATCH	6	0.6239	0.4378	-0.2342	1.4819	1.4250	0.1541
BATCH	7	0.3489	0.4388	-0.5110	1.2089	0.7953	0.4265
BATCH	8	-0.2358	0.5042	-1.2240	0.7524	-.4677	0.6400
BATCH	9	-0.2943	0.5305	-1.3342	0.7455	-.5548	0.5790
BATCH	10	-0.2498	0.5344	-1.2972	0.7977	-.4674	0.6402
BATCH	11	-0.4503	0.5244	-1.4781	0.5774	-.8588	0.3905
BATCH	12	-0.1553	0.5124	-1.1596	0.8489	-.3032	0.7617
BATCH	13	-0.5111	0.5373	-1.5643	0.5421	-.9512	0.3415
BATCH	14	-0.0088	0.5182	-1.0245	1.0068	-.0170	0.9864
BATCH	15	-0.0399	0.4535	-0.9287	0.8490	-.0879	0.9299
BATCH	16	0.0983	0.4667	-0.8163	1.0130	0.2107	0.8331
BATCH	17	-0.3273	0.5078	-1.3226	0.6681	-.6444	0.5193
BATCH	18	-0.3597	0.4786	-1.2976	0.5783	-.7516	0.4523
BATCH	19	-0.1252	0.4878	-1.0812	0.8309	-.2566	0.7975
BATCH	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Scale		1.0070	.	.	.	.	.

Again, radiologist C is significantly different to radiologist E, whereas prompting is only marginally significant. The non-significant results for the batch differences would indicate that there is not any difference between the number of recalls in each batch, that they are reasonably uniform.

4.10.4 Comparison of results with historical data

All these cases, cancers and normals, were taken from the screening programme, and thus had historical data attached. With this form of hindsight, it is possible to examine how the decisions vary between testing periods.

To simplify table 4.22 further, it is sensible to collapse 'First radiologist' and 'Second radiologist' into a single variable - 'only one radiologist'. In other words, if a case was recalled by either the first or the second radiologist (but not both), then it was considered to be a single radiologist recall. This may be seen in table 4.23.

	Experimental data			
	Both	First	Second	Neither
Historical data	radiologists	radiologist	radiologist	radiologist
Both radiologists	70	7	2	3
First radiologist	5	0	0	0
Second radiologist	10	1	1	2

Table 4.22: Comparison of historical and experimental recalls of cancers

	Experimental data		
Historical data	Both radiologists	One radiologist	Neither radiologist
Both radiologists	70	9	3
One radiologist	15	2	2

Table 4.23: Collapsed version of table 4.22

This is not truly a fair comparison for the experimental radiologists, as no account of the cancers missed by the original screening radiologists has been made. Ideally, a few interval cancers should have been included for comparison, but this proved to be unfeasible.

Individual sensitivities have also varied between screening (when the cancers were first seen at the clinic) and experiment (table 4.24). Indeed, radiologists' view of what is and is not suspicious varies greatly, not only between radiologists, but also within radiologists [56]. Performance can depend on a great many effects; time of day, fatigue, how the day prior to reading had gone.

Other than $\mathrm{R_C}$ (who only missed one cancer in the experiment), every radiologist has a lower sensitivity when reading second than when they were the first reader in this experiment. This, however, was not a significant decrease, probably due to the low numbers of cancers per radiologist. Overall, without taking radiologist into account, it is already known that order is not significant. There is also no significant difference between the sensitivities at screening and those during the experiment.

If we now examine the additional 71 benign recalls (considered suspicious at screening but later found to be non-malignant) that were also in the set of non-malignancies (table 4.25), it can again be seen that the unprompted reader has made the larger

	Screening sensitivity	Experimental sensitivity
R_A first	92.5% (40)	91.7% (12)
R_A second	100% (23)	80.0% (20)
R_B first	92.3% (27)	88.5% (26)
R_B second	92.0% (23)	80.0% (10)
R_C first	100% (3)	95.8% (24)
R_C second	100% (18)	100% (22)
R_D first	69.0% (29)	93.3% (15)
R_D second	93.1% (29)	84.6% (13)
R_E first	100% (2)	92.0% (25)
R_E second	100% (5)	86.5% (37)

Table 4.24: Comparison of sensitivities from screening and experiment. Numbers in parentheses refer to the number of cancers seen

number of recalls from this set of suspicious but benign features, significant with a p-value of 0.02, which tallies with the larger set of all non-cancers.

		Unprompted		
		Recalled	Not recalled	Total
Prompted	Recalled	34	6	40
	Not recalled	17	14	31
	Total	51	20	71

Table 4.25: The number of historically benign recalls recalled during the experiment, by prompting

Similarly, there is no significant difference between the numbers recalled by the first and second readers (table 4.26).

		Second reader		
		Recalled	Not recalled	Total
First reader	Recalled	34	13	47
	Not recalled	10	14	24
	Total	44	27	71

Table 4.26: The number of historically benign recalls recalled during the experiment, by order

4.10.5 Time and observational data

Each reading session, prompted and unprompted, for each radiologists was timed, so that comparisons could be made on the duration of prompted and unprompted sessions (figure 4.5 and table 4.27). In general, radiologists took less time to complete each prompted session as the experiment progressed, but this also held true for some of the radiologists' unprompted sessions.

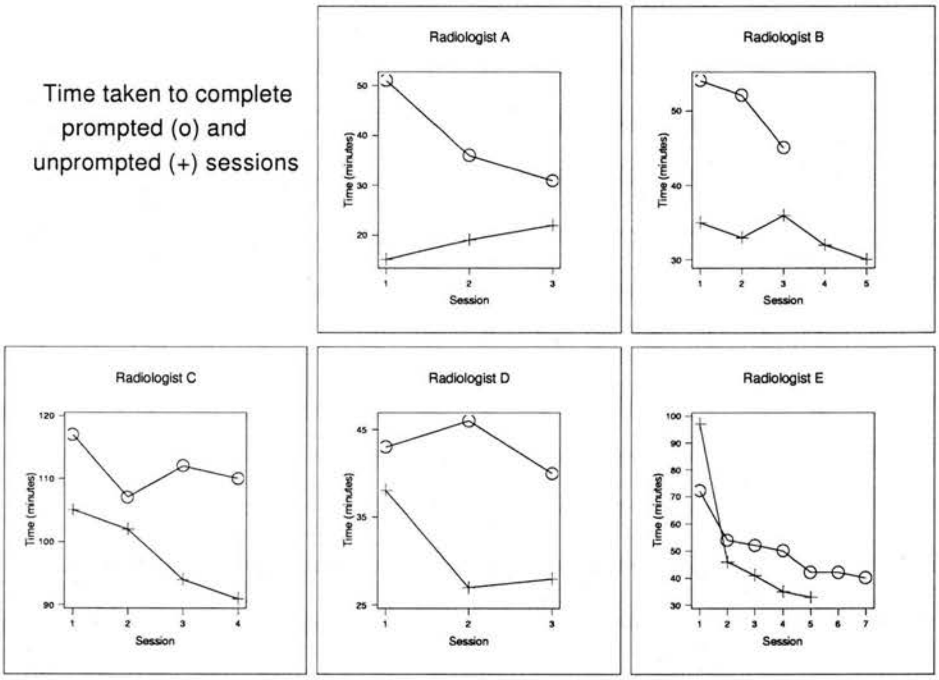


Figure 4.5: Time taken to complete each session

Due to the constraints of the experiment, it was not possible to completely balance the reading sessions in the limited time available.

One of the arguments for recording the time was to assess whether there was any form of learning effect; whether the radiologists would become more comfortable with the system and, thence, spend less time on interpreting and dismissing obvious false prompts. However, given that the time taken during the unprompted session also decreased (in general, not in all cases), this cannot be the case. It may be that the longer duration during the earlier unprompted sessions can be explained by lack of familiarity with the protocol, and the absence of first reader information.

In every session, bar one (R_E session 1 unprompted), the prompted session is always longer than the corresponding unprompted session. Part of this may be due to reading styles; the experimental protocol demanded that the prompted radiologist examine the

		Time in minutes	
Radiologist	Session	Unprompted session	Prompted session
$\mathrm{R_A}$	1	15	51
	2	19	36
	3	22	31
$\mathrm{R_B}$	1	35	54
	2	33	52
	3	36	45
	4	32	*
	5	30	*
$\mathrm{R_C}$	1	105	117
	2	102	107
	3	94	112
	4	91	110
$\mathrm{R_D}$	1	38	43
	2	27	46
	3	28	40
$\mathrm{R_E}$	1	97	72
	2	46	54
	3	41	52
	4	35	50
	5	33	42
	6	*	42
	7	*	40

Table 4.27: Time taken to complete each reading session

film then examine the prompt sheet for that film. Thus, each case was checked and recorded one at a time. For the unprompted sessions, no such guidelines were given, and so radiologists were free to use whichever method of checking and reporting they felt most comfortable with. For two of them, this meant the batch method, where they would either check the films until they found one that they wished to recall, then report all the cases up to that one, or merely check a certain number of films, then report on them. Further discussion of the reading behaviour can be found in Mark Hartswood's thesis [91].

The unexpectedly high value for $\mathrm{R_E}$ session 1 unprompted is most likely due to $\mathrm{R_E}$ starting to read on a case by case basis and then switching to a batch style, and to the fact that this was his/her first session in the experiment.

4.10.6 Questionnaire data

Questionnaires similar to the one given during the experiment covered in Chapter 3 were administered during this experiment. They were not identical to the earlier ones, since the purposes of the experiments were different. Where a question in this section has a corresponding question in the earlier chapter, it will be marked with a (*). The questionnaires can be found in Appendix H.

4.10.6.1 Pre- and post-experiment questionnaires

Before	Rating					Average	
Feature	1	2	3	4	5	Rating	Rank
Vascular calcification				3	2	4.4	8.5
Benign clusters			1	4		3.8	6.0
"Popcorn" calcification			1	1	3	4.4	8.5
Film artefacts		1		1	3	4.2	7.0
Lymph nodes		1	3	1		3.0	4.0
Well defined masses	3	1	1			1.6	1.0
Composite shadows		1	4			2.8	2.5
Nodular glandular structure		3		2		2.8	2.5
Cysts		2	1	1	1	3.2	5.0

Table 4.28: *Q1* Rate the distracting effect of algorithm output (1 = Useful, 5 = Distracting)*: Before

After	Rating					Average	
Feature	1	2	3	4	5	Rating	Rank
Vascular calcification	1			3	1	3.6	5.0
Benign clusters		2	1	2		3.0	3.5
"Popcorn" calcification		1		2	2	4.0	7.0
Film artefacts				2	3	4.6	9.0
Lymph nodes		2	3			2.6	1.5
Well defined masses		3	1	1		2.6	1.5
Composite shadows				4	1	4.2	8.0
Nodular glandular structure			1	3		3.75	6.0
Cysts		1	3	1		3.0	3.5

Table 4.29: *Q1* Rate the distracting effect of algorithm output (1 = Useful, 5 = Distracting)*: After

Q1* Rate the distracting effect of algorithm output (1 = Useful, 5 = Distracting) Separate tables for the responses before and after the experiment (tables 4.28 and 4.29) do not convey the full story. The data are presented graphically in figure 4.6. The feature names have been truncated to fit on the graph, but are reasonably intuitive.

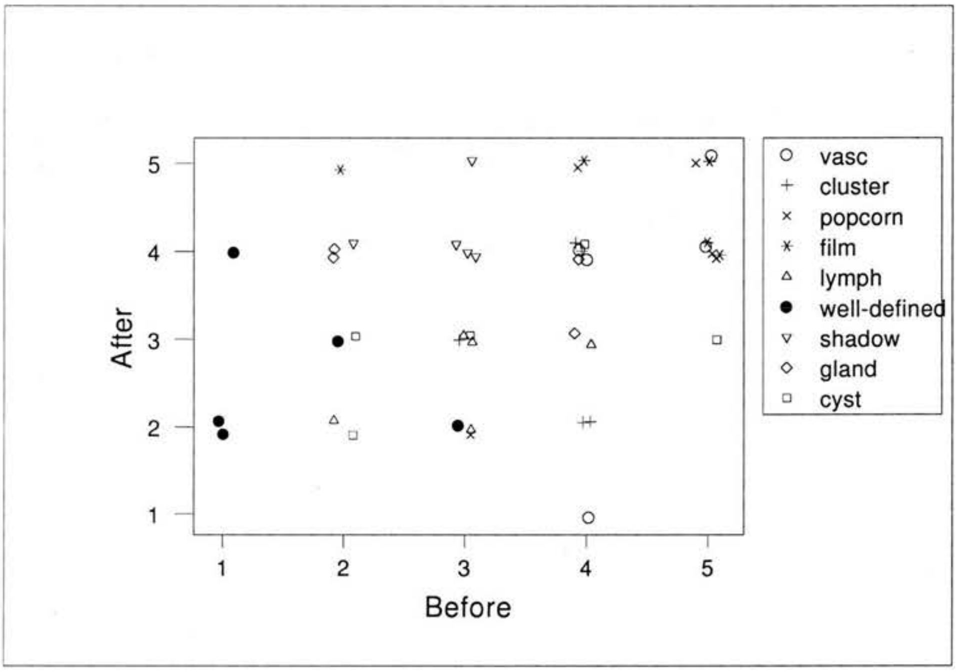


Figure 4.6: Attitude score of false prompt features on a scale of 1 to 5

The graph is a little cluttered, but it is still possible to see that there is very little agreement between the 'before' and 'after' questionnaires. Pearson correlation of the average rating before and after the experiment is 0.52, which is not significant at the 5% level, as would be expected from figure 4.6. Similarly, the Spearman's rank correlation coefficient of the mean scores is only 0.36, also non-significant.

Overall, there appears to be little consistency between the scoring made before and after the experiment, with 14 instances of a positive change (where a high score before the experiment became a lower score after), 15 instances of a negative change, and 15 with no change (one missing value). However, examining these on a feature by feature basis was a little more interesting (table 4.30).

Well defined masses and composite shadows are the features that have had the greatest change in opinion. In the case of composite shadows, all five radiologists have changed

	Positive change	Negative change	No change
Vascular calcification	2	0	3
Benign clusters	2	0	3
"Popcorn" calcification	3	1	1
Film artefacts	2	2	1
Lymph nodes	2	0	3
Well defined masses	1	4	0
Composite shadows	0	5	0
Nodular glandular structure	1	2	1
Cysts	1	1	3

Table 4.30: Change in opinion between the start of the experiment and the end

their opinions from '3' or better to '4' or worse; indicating that prompts for this feature were more annoying than first anticipated. Of the other features, the most interesting are the few that have changed from a negative opinion ('4' or '5') to a positive opinion ('1' or '2') (the lower right quadrant of figure 4.6). This would seem to show that benign calcification prompts are not as redundant as initially expected.

Q2* Please rank the following categories of false positive as to the priority that should be given to their removal (1 = this feature should be removed first, 2 = this feature should be removed second etc.)

Due to some misunderstanding of the question in the subjective reaction experiment (chapter 3), this question was clarified somewhat to ensure that each value was only used once.

Since there are nine potential categories and only five radiologists, it seemed a little redundant to include the tables of responses. The following table (table 4.31) gives the means and ranks before and after the experiment.

As the image below (figure 4.7) shows, opinions have changed between questionnaires, although vascular calcification and film artifacts are consistently considered to have high priority in removal. It is difficult from the individual observations to discern this pattern. Hence, the following graph (figure 4.8) has only the mean priority (the average score over all five radiologists). It is clearer here that vascular calcification and film

	Mean rating		Rank	
	Before	After	Before	After
Vascular calcification	1.6	2.4	1.0	1.0
Benign clusters	6.3	6.4	7.0	9.0
"Popcorn" calcification	3.6	4.8	3.0	3.0
Film artefacts	2.0	3.4	2.0	2.0
Lymph nodes	4.8	5.6	4.0	6.0
Well defined masses	8.3	6.0	9.0	7.5
Composite shadows	6.6	5.2	8.0	4.5
Nodular glandular structure	5.6	5.2	5.0	4.5
Cysts	6.2	6.0	6.0	7.5

Table 4.31: *Q2* Please rank the following categories of false positive as to the priority that should be given to their removal (1=first removed, etc)*

artefacts have the highest priority both before and after the experiment.

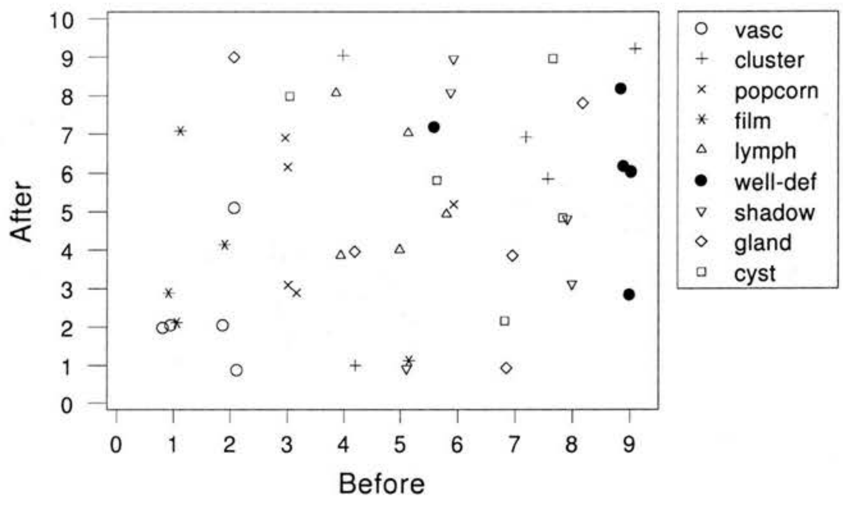


Figure 4.7: *Q2* Please rank the following categories of false positive as to the priority that should be given to their removal (1=first removed, etc)*

Using the mean rating and ranks, we see that the correlation between the responses given before the experiment and again afterwards have higher agreement on the order of removal than they did in question 1 (the amount of distraction). Pearson's correlation has a coefficient of 0.88 (p = 0.002), while the non-parametric Spearman's rank

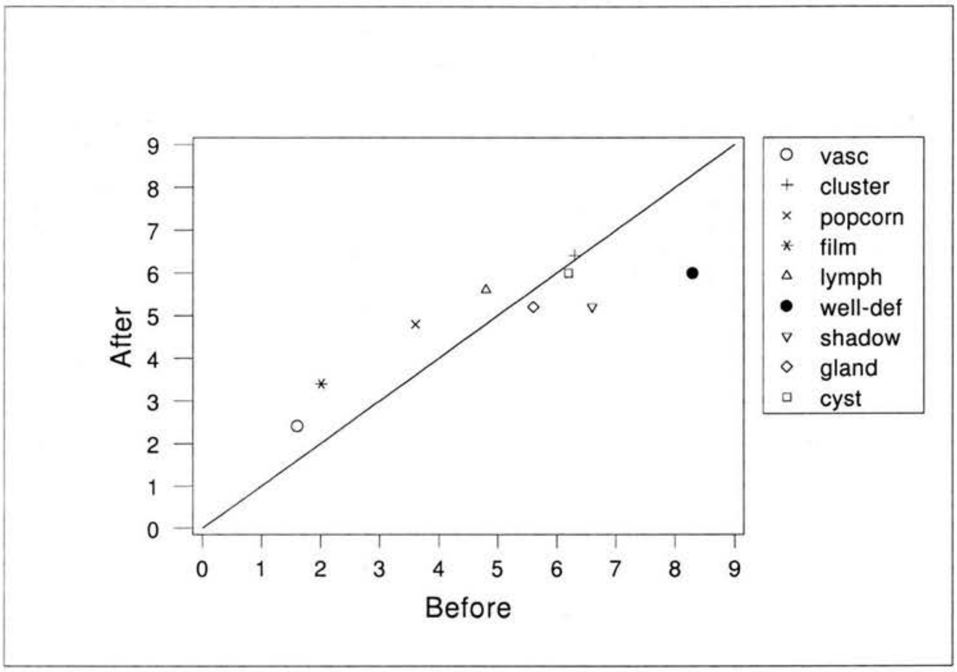


Figure 4.8: Mean ratings for *Q2* Please rank the following categories of false positive as to the priority that should be given to their removal (1=first removed, etc)*

correlation has a coefficient of 0.79 (p = 0.01).

Given the form of these data, it is possible to examine it in many ways. The method used above was merely the simplest and encompassed all the results, albeit in a much reduced way. Since each radiologist is asked to rank a series of features, it is possible to then compare that rank with their opinions after the experiment or with one of their colleagues. Unfortunately, this would give us a large number of comparisons (five before/after and ten radiologist to radiologist) which would be of limited use.

Below are the individual radiologist correlation coefficients for the *before* and *after* responses (table 4.32).

Although the radiologists' opinions have changed over time, they invariably agree that vascular calcification has a high priority for removal. There is one exception to this; R_E changed their ranking of vascular calcification from second to fifth. Radiologist E has, unusually, completely reordered his/her rankings.

Radiologist	Rank correlation
R_A	0.37
R_B	0.64
R_C	0.23
R_D	0.68
R_E	-0.12

Table 4.32: Rank correlations for *Q2* Please rank the following categories of false positive as to the priority that should be given to their removal (1=first removed, etc)*

Q3 Please rate the following tasks according to how difficult or how easy you find them.

This question was, unusually, not directly aimed at the radiologists' response to the prompting system. Rather, its purpose was to ascertain which functions of the prompting system would be most useful as a complementary function.

As figure 4.9 illustrates, the radiologists find the detection of microcalcification relatively easy, although they find the classification of this feature somewhat more difficult. Despite what appears to be perfect agreement in two sub-questions, these actually have poorer agreement between the responses before the start of the experiment and the end than the detection of architectural distortions.

There is very little agreement between a radiologist's opinion at the start of the experiment and the end (table 4.33), with no categories having a greater level of agreement than moderate (see page 34). This suggests that the radiologists' opinions are changing during the period of the experiment. These questions are possibly making them think about what they are doing during reading, rather than proceeding on a more intuitive level.

	Weighted κ
Detection of microcalcification clusters	0.55
Detection of ill-defined lesions	0.38
Detection of architectural distortions	0.58
Detection of asymmetries	0.23
Classification of microcalcifications	0.38
Classification of ill-defined lesions	0.38

Table 4.33: Weighted κ statistic for before and after the experiment for *Q3 Please rate the following tasks according to how difficult or how easy you find them.*

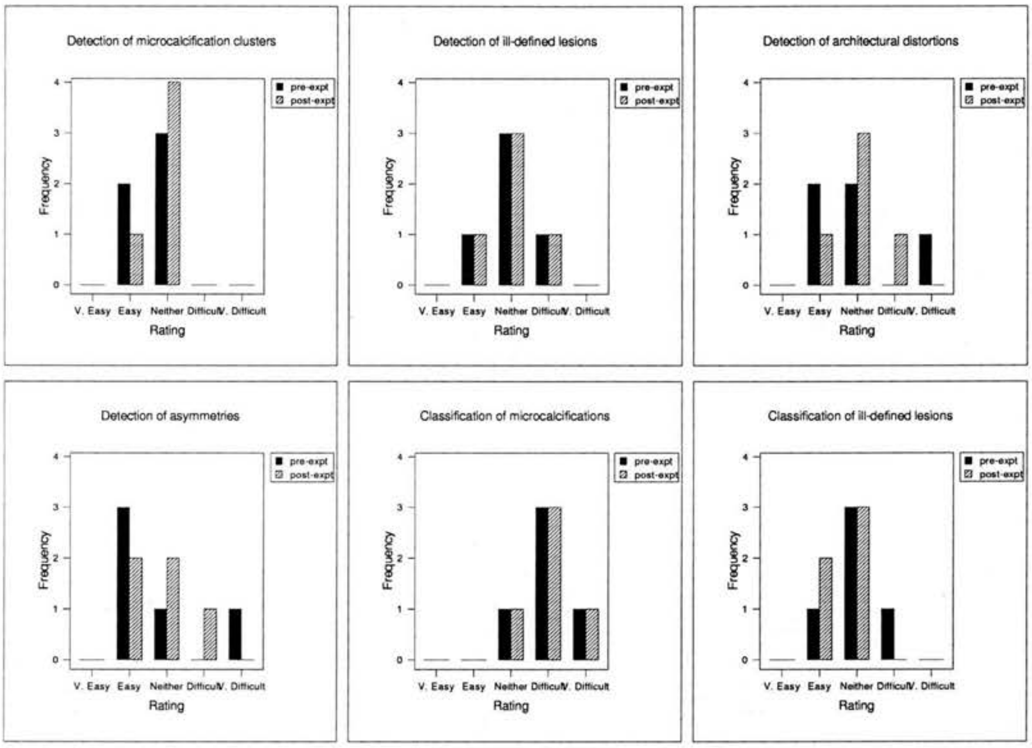


Figure 4.9: *Q3 Please rate the following tasks according to how difficult or how easy you find them.*

Q4 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:

The consensus is that prompting for microcalcification, ill-defined lesions and architectural distortion is essential in any screening prompting system, with asymmetries and the classification of prompts less so (see figure 4.10). It appears that the radiologists would prefer to make their own classification once a suspicious region has been brought to their attention.

Again, there is little agreement between the responses before and after the period of the experiment. The higher agreement is for the microcalcification prompts, as would be expected. Below are the individual κ values for each sub-question (table 4.34).

The negative κ value for the question of prompting for asymmetries indicates that there

	Weighted κ
Prompting for microcalcification clusters	0.55
Prompting for ill-defined lesions	0.38
Prompting for architectural distortions	0.17
Prompting for asymmetries	-0.25
Classification of prompted microcalcification clusters	0.12
Classification of prompted ill-defined lesions	0.17

Table 4.34: Weighted κ statistic for the agreement before and after for *Q4 Rate each of the hypothetical properties of a prompting system in terms of how useful you perceive they might be in a screening practice*

is less agreement than would be expected by random chance. In fact, there is only one instance of a radiologist not changing their opinion in this question (see table 4.35), with three of the rest down-grading their opinion.

Q5 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)

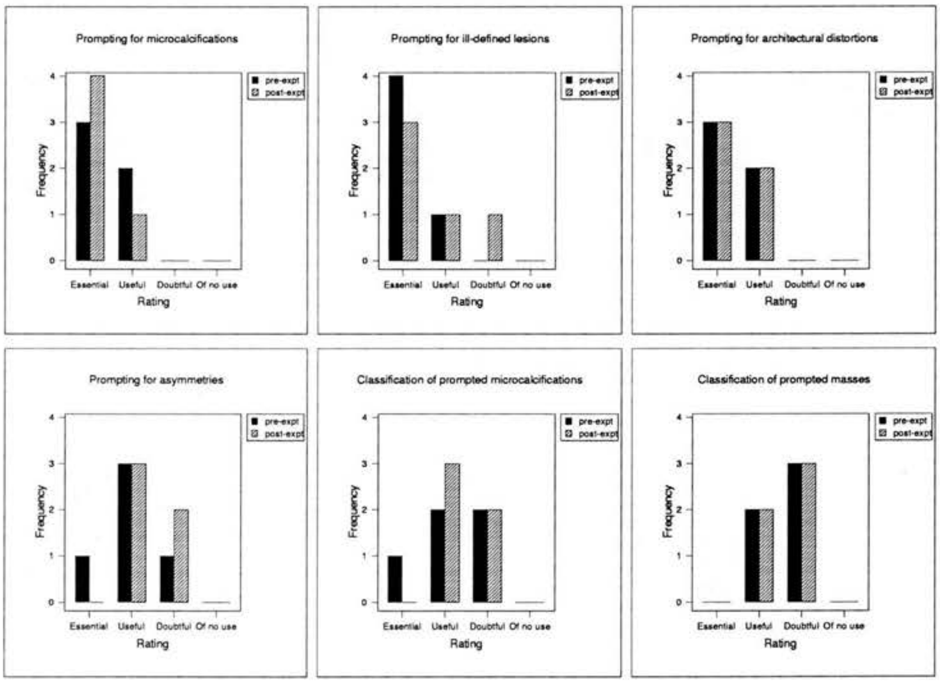


Figure 4.10: *Q4 Rate each of the hypothetical properties of a prompting system in terms of how useful you perceive they might be in a screening practice*

		After			
		Essential	Useful	Doubtful	Of no use
Before	Essential		1		
	Useful		1	2	
	Doubtful		1		
	Of no use				

Table 4.35: Responses to the 'prompting' for the asymmetry sub-question

	Mean rating		Rank	
	Before	After	Before	After
calc	1.8	1.8	1.0	1.0
lesions	2.4	2.4	2.0	3.0
distortion	3.4	2.0	3.0	2.0
asymmetry	4.8	4.6	5.5	4.5
class calc	3.8	4.6	4.0	4.5
class lesion	4.8	5.6	5.5	6.0

Table 4.36: Mean ratings and their ranks for *Q5 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)*

As with the previous question, the classification of features by a prompting system is not particularly welcomed by the radiologists, with prompting for microcalcification clusters, ill-defined lesions and architectural distortion again a high priority. On average, little changes between the start and end of the experiment, despite architectural distortion's rather large jump from fifth to first priority in one instance.

The correlation coefficient between the mean scores is 0.87 (p = 0.023), whereas the Spearman rank correlation coefficient is 0.90 (p = 0.015).

The functions have split into two distinct groups; prompting for microcalcification clusters, ill-defined lesions and distortion; and prompting for asymmetry, classification of calcification and lesions. This does not change from the start of the experiment to the end, although relative position within the groups does (see table 4.36).

Q6 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)

Table 4.37 illustrates the mean ratings of the problems that would be addressed by a prompting system. Here, the correlation coefficient of the mean rating is 0.88 (p = 0.022), with the Spearman's rank correlation coefficient at 0.70 (p = 0.12) This non-significant result may be due to the three equally ranked scores in the 'before' case.

	Mean rating		Rank	
	Before	After	Before	After
Reducing the number of interval cancers (FN)	2.0	2.8	1	3
Improving the detection performance of a single reader (performance)	2.8	1.8	3	1
Improving the consistency of reading (consistency)	2.4	2.4	2	2
Supporting inexperienced radiologists (inexperience)	4.6	4.2	5	4
Addressing resourcing limitations (resources)	4.6	4.6	5	5
Reducing recalls (recalls)	4.6	5.2	5	6

Table 4.37: Mean ratings and their ranks for *Q6 In a screening practice, what problems do you see a prompting system addressing?*

The change in rank between reducing interval cancers and improving the detection performance of a single reader may indicate a slight disappointment in the function of the algorithms by the radiologists. Their belief that the system would be able to detect false negative cancers has been lowered, allowing consistency of reading and improving the performance of a single reader to become more prominent. This may be only a perceptual belief, however. It is possible that the radiologists are not noticing when they are being prompted for something that they would have missed, had they been unprompted. This is supported by the four prompted cancers that were missed by the prompted radiologists, and especially the two cancers that were correctly prompted by the system and were missed by both the prompted and unprompted readers. In a screening situation, those two cancers would have been false negatives, giving the system a 40% success rate in false negative cancers.

Q7 How do you see a prompting system being used in your clinic:

By the end of the experiment four of the five radiologists have formed the opinion that the system would be of more use in enhancing double reading than in having the system 'replace' a reader; in other words, assisted single readers (see table 4.38 and figure 4.11). This implies that the radiologists may have become slightly disillusioned by the system's performance.

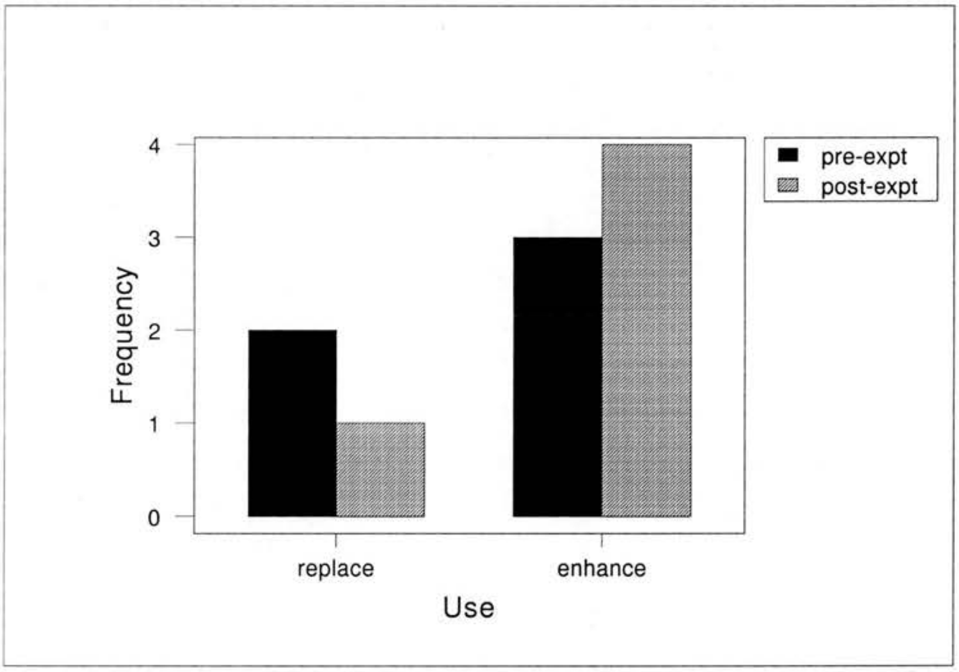


Figure 4.11: *Q7 How do you see a prompting system being used in your clinic*

	After	
Before	Replacing double reading with single	Enhancing double reading
Replacing double reading with single	1	1
Enhancing double reading	0	3

Table 4.38: *Q7 How do you see a prompting system being used in your clinic*

Q8a* In cases where you are unsure, would the presence of a prompt make you more inclined to recommend recall (Strongly Agree to Strongly Disagree)

Radiologist	Before	After
R_A	Agree	Uncertain
R_B	Agree	Agree
R_C	Uncertain	Agree
R_D	Disagree	Uncertain
R_E	Agree	Uncertain

Table 4.39: *Q8a The presence of a prompt will make you more inclined to recommend recall*

Q8b* In cases where you are unsure, would the absence of a prompt make you less likely to recommend recall (Strongly Agree to Strongly Disagree)

Radiologist	Before	After
R_A	Agree	Uncertain
R_B	Strongly Disagree	Uncertain
R_C	Uncertain	Agree
R_D	Disagree	Agree
R_E	Uncertain	Uncertain

Table 4.40: *Q8b The absence of a prompt will make you less likely to recommend recall*

The two questions above (tables 4.39 and 4.40) should not be considered separately. In both questions, there is more uncertainty after the experiment; radiologists are not sure whether the presence/absence of a prompt in borderline cases will influence their decision to recall or not. The results from the recall data, however, suggest that the radiologists are using prompt information to inform a recall decision (see section 4.10.2, page 98). The answers given to part b show that the radiologists are aware of this at some level. The answer to this question should, in theory, have been 'Strongly Disagree'. The absence of a prompt shows nothing other than the algorithms have failed to find anything. This does not mean that there is nothing there to find. Radiologist B's change of response indicates an understanding that the absence of a prompt is falsely reassuring in borderline cases.

Q9* 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)

As with the earlier questionnaire (section 3.7.1), the score for the higher rate of prompting became more popular once radiologists had been exposed to the prompting system (page 65). Similarly, average score for the low prompt rate/low sensitivity has fallen, although only one radiologist changed their opinion on this option. Since the rating on this configuration was poor before the start of the experiment, it was unlikely that it would worsen very much. The score of '2' was given by radiologist B on both occasions (see tables 4.41 and 4.42).

Before	Score					
Configuration	1	2	3	4	5	Ave. Score
High prompt rate, high sensitivity	1		2	2		3.0
Low prompt rate, low sensitivity		2		2	1	3.4
Microcalcification clusters, but no other types of calc	4	1				1.2
All types of calcification		1	3	1		3.0
Opacities usually dismissed with previous or multiple films		4		1		2.4

Table 4.41: *Q9* 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)* Before

After	Score					
Configuration	1	2	3	4	5	Ave. Score
High prompt rate, high sensitivity		4		1		2.4
Low prompt rate, low sensitivity		1		3	1	3.8
Microcalcification clusters, but no other types of calc	3	2				1.4
All types of calcification			2	2	1	3.8
Opacities usually dismissed with previous or multiple films		2	1	2		2.4

Table 4.42: *Q9* 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)* After

Little has changed in the microcalcification clusters only configuration, with only one radiologist changing their opinion. The usefulness of prompting for all types of microcalcification clusters has fallen, as has the configuration that prompts for opacities that can be dismissed with previous and/or multiple films. It appears that radiologists are only interested in features that cannot easily be dismissed as benign.

4.10.6.2 Post-experiment only questions

The following four questions were designed for this experiment only, and did not appear in the previous experiment.

Q10 At the outset of this experiment we gave you an estimate 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?

The mean estimates are quite close to the actual experimental results (table 4.43), indicating that despite the relatively low number of cancers seen by each radiologist (28 to 62), they are able to glean a fairly accurate picture of how sensitive the algorithms are (figure 4.12). This suggests that they are using information other than the cancers.

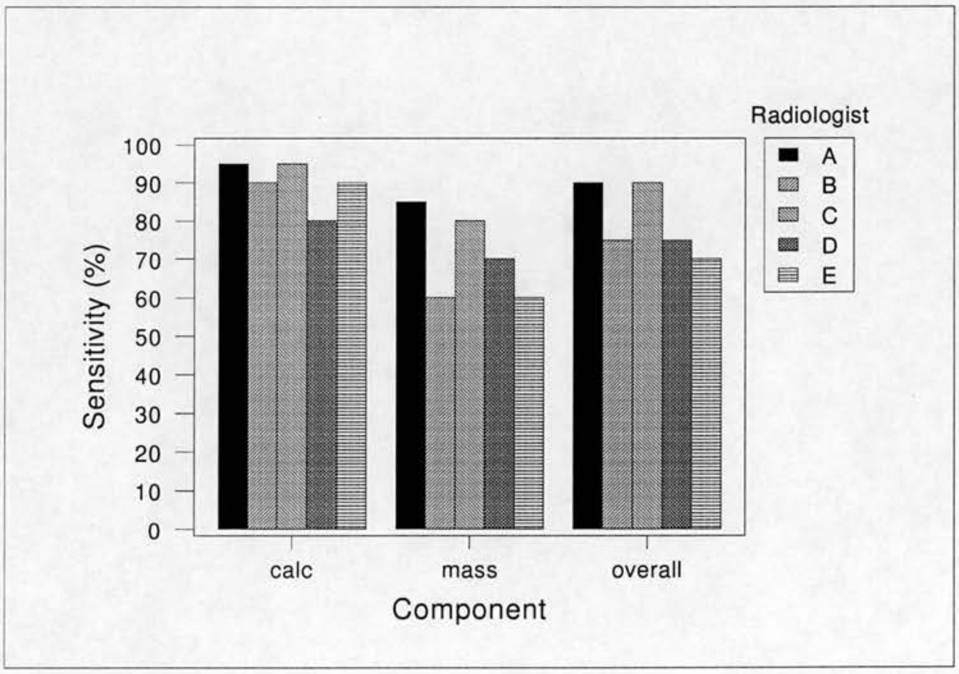


Figure 4.12: *Q10 At the outset of this experiment we gave you an estimate of the sensitivity of the ill-defined lesion and microcalcification algorithms. Based on your experience of using the system, what would be your estimate of the sensitivity of these components?*

			Radiologist estimates		
Algorithm	Sensitivity given at training	Actual sensitivity	Mean	Median	SE Mean
Microcalcification	90%	93.8%	90.0%	90.0%	2.74%
Ill-defined lesion	81%	72.9%	71.0%	70.0%	5.10%
Overall	-	81.4%	80.0%	75.0%	4.18%

Table 4.43: *Q10 At the outset of this experiment we gave you an estimate 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?*

Q11 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).

The radiologists appear mostly confident about their ability to estimate the sensitivity of the algorithms (table 4.44), although R_E admits to some doubt when asked for an estimate of the ill-defined lesions algorithm.

Radiologist	Microcalcification	Ill-defined lesion	Overall
R_A	2	2	2
R_B	3	2	2
R_C	2	3	3
R_D	2	3	3
R_E	1	4	2

Table 4.44: *Q11 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)*

Q12 Do you believe your sensitivity in the prompted session has been better, the same or worse, compared with your sensitivity in the unprompted sessions, for the following types of lesion:

Figure 4.13 shows that the radiologists believed that prompting was either improving or not worsening their sensitivity. In fact, although the sensitivity of the unprompted readers (over all sessions) was slightly higher than that of the prompted readers, for four of the five readers prompting improved their sensitivity. Radiologist E, who's sensitivity declined with prompting, believed that his/her sensitivity with the microcalcification was better with prompting, and was no different for the masses or overall.

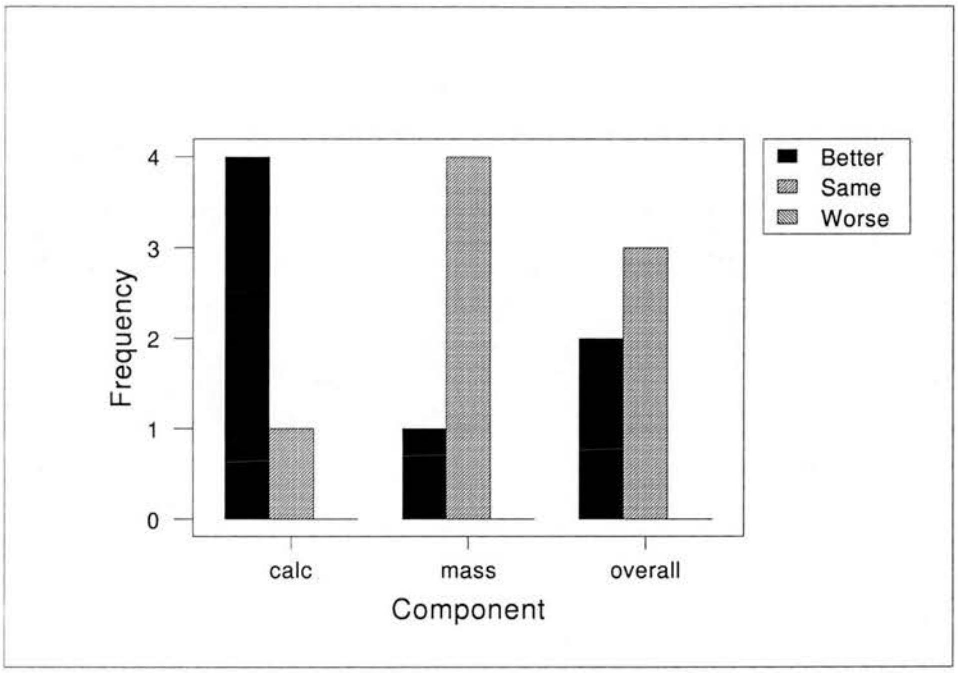


Figure 4.13: *Q12 Do you believe your sensitivity in the prompted session has been better, the same or worse, compared with your sensitivity in the unprompted sessions, for the following types of lesion*

Q13 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:

Opinion here (figure 4.14) appears to be different from the actual results of the experiment. Despite two radiologists having higher false recall rates while prompted (although one was only very slight), and the other three having lower false recall rates (one significantly lower, $\chi^2 = 4.8, p = 0.03$), their view of the system overall was that it was making no difference to their specificity whether they were prompted or not prompted. Radiologist E believed that his/her specificity with respect to microcalcification was better, but that it was worse with the masses (same overall), and yet his/her recall rate was lower with prompting (although not significantly so).

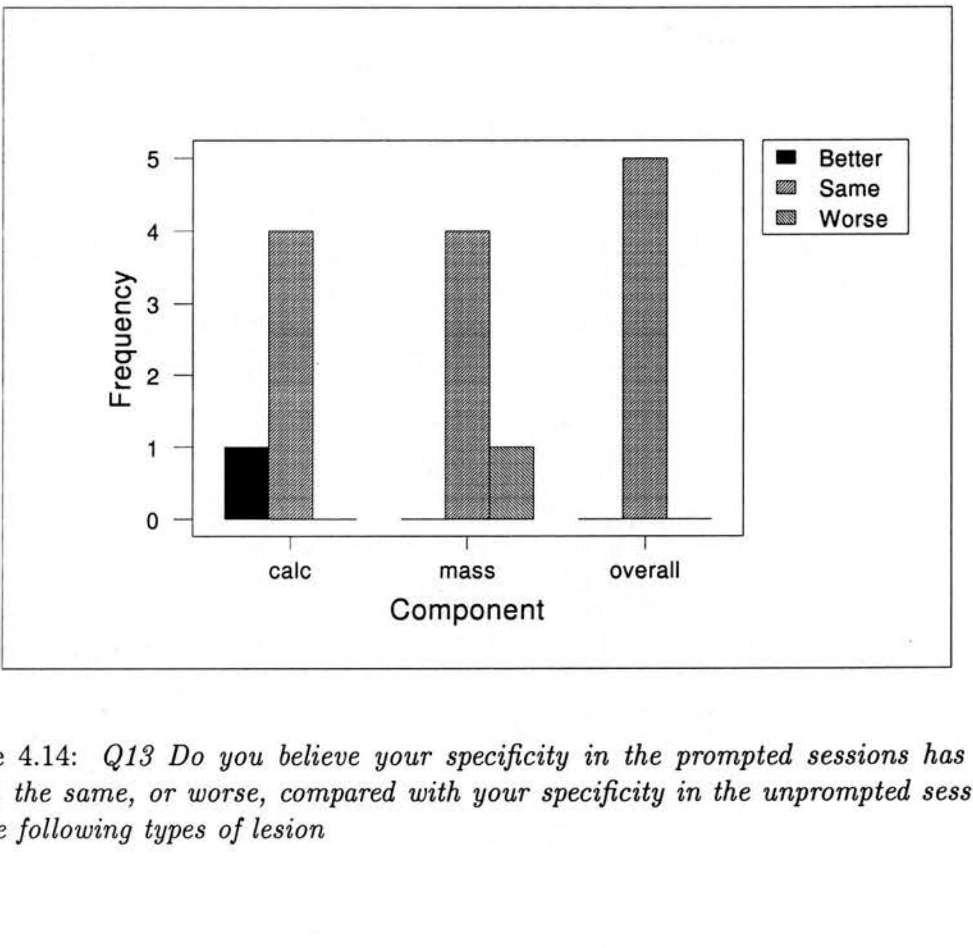


Figure 4.14: *Q13 Do you believe your specificity in the prompted sessions has been better, the same, or worse, compared with your specificity in the unprompted sessions, for the following types of lesion*

4.10.6.3 Post-session questionnaires

Q1* The Likert score and the radiologists' opinions

As before, (see section 3.7.3) a series of questions was put to the radiologist after each prompted session (not for the unprompted session), with the response rated on a five point scale of strongly agree to strongly disagree, and scored according to the negative or positive aspect of the question (see figure 4.15). Similarly, the overall score is the subjective performance rating given by the radiologist (see figure 4.16).

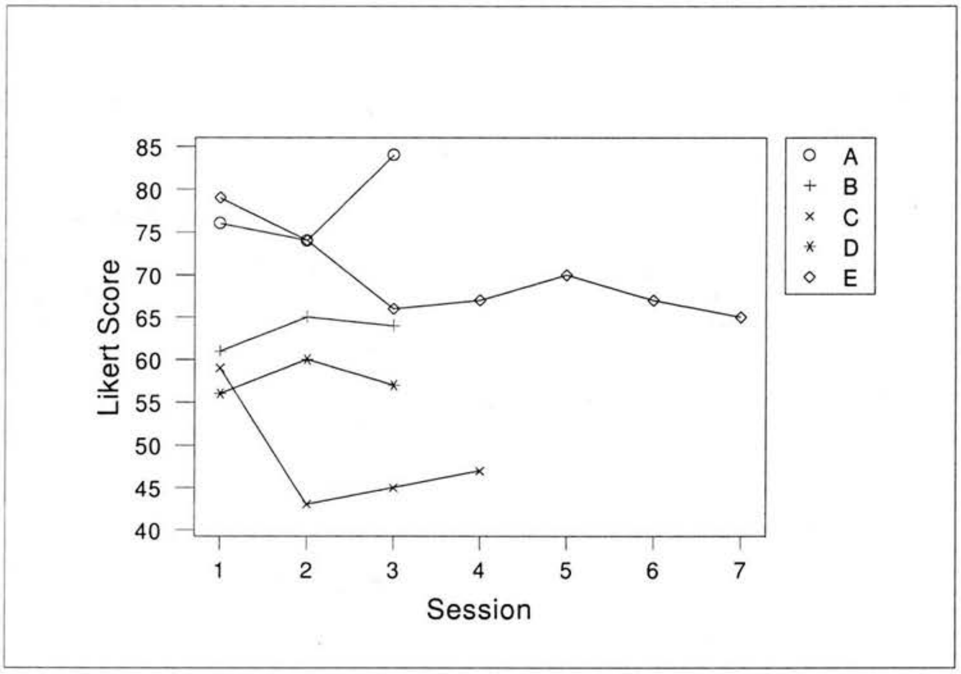


Figure 4.15: Likert score after each prompted session, radiologist markers

Correlation between these two variables is better than that in the subjective experiment (see section 3.7.3), with a correlation coefficient, calculated from all observations (ignoring the lack of independence), of 0.84.

Q2* Do you believe?

This question came in a series of three sub-questions, with the possible responses being only yes or no. A similar question was asked in the previous experiment.

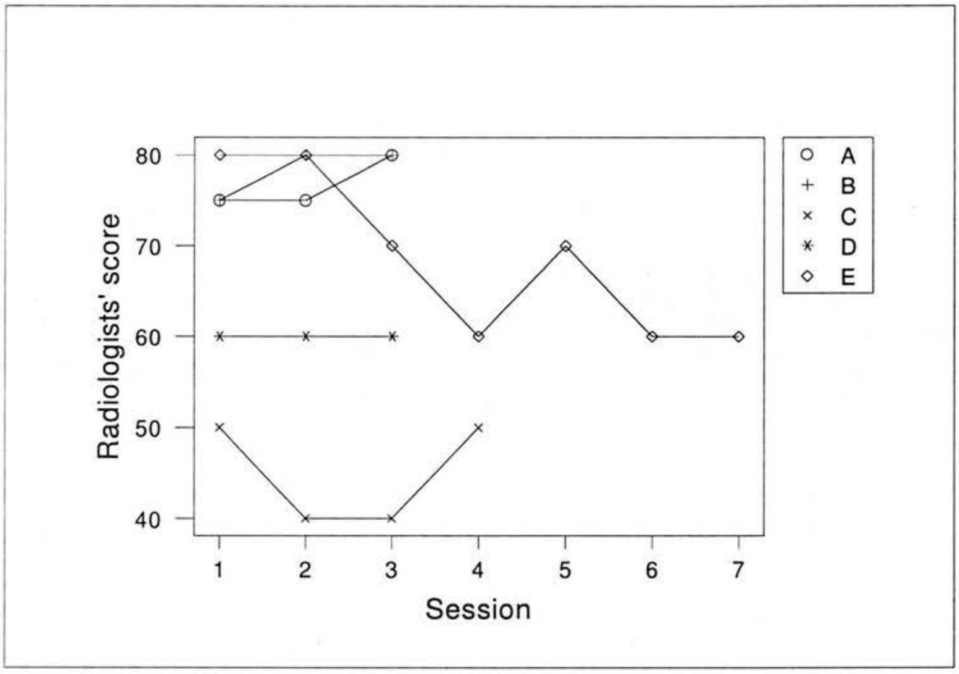


Figure 4.16: Radiologists' score after each prompted session, radiologist markers

Do you believe that:

- Overall, the system would be useful to you in a screening context as it currently stands (total)?
- The mass detection component of the system would be useful to you as it currently stands (mass)?
- The microcalcification detection component of the system would be useful to you as it currently stands (calc)?

Opinion is divided on the usefulness of the system in its current form, although the majority agree that the microcalcification cluster algorithm is useful as it stands (see figure 4.17). Strangely, at the end of the experiment more radiologists believe that the overall system was useful than believed that at the start, despite a decrease in the number of radiologists who believed that the ill-defined lesion (mass) algorithm was of use. Since these changes were made by two different radiologists (R_D and R_E), little

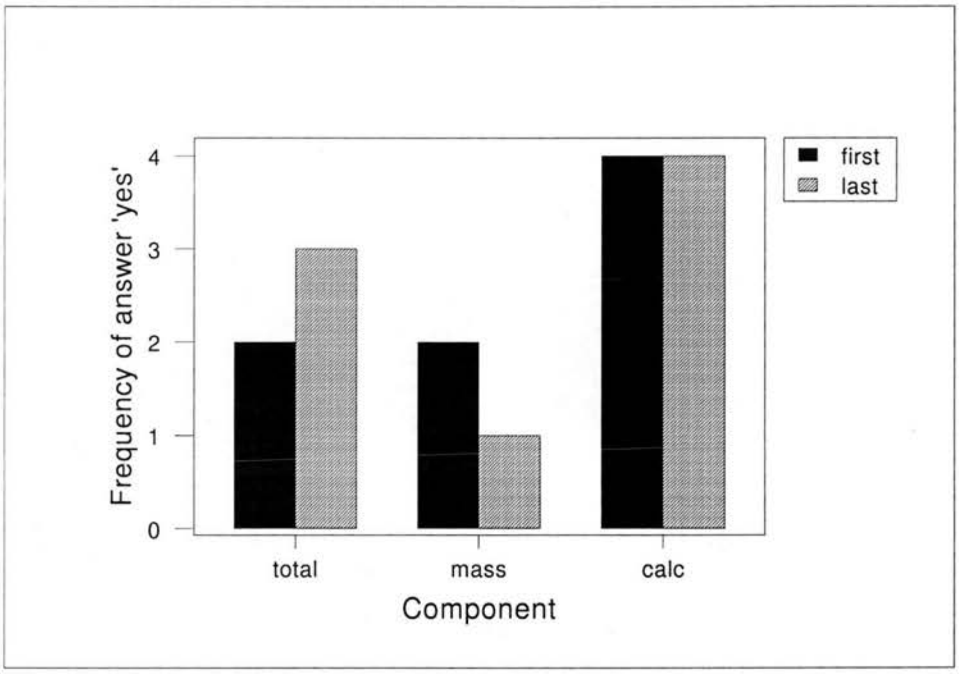


Figure 4.17: *Q2 Do you believe that overall, the system would be useful to you in a screening context as it currently stands?*

can be inferred from this. R_D changed their opinion on the overall usefulness of the system to 'yes', despite believing that the mass part of the system was not useful at all sessions. Looking at R_E's results individually, of the seven times this question was asked of this radiologist, a 'No' response was given for the mass question only twice third and seventh sessions). Hence, this change is not necessarily indicative of a gradual disenchantment with the algorithm; it is more likely that it is due to the performance of the algorithm on the batch of cases read during those sessions.

Q3 Please rate the system's sensitivity

Obviously, the perception that the system is too sensitive once the radiologists have been exposed to more than one session is clear (see figures 4.18 and 4.19). It is likely to be difficult to make an assessment of sensitivity based on one session (approximately 100 cases, between two and six pathology proven cancers), although the post-experiment questionnaires indicated that they were able to form a fairly accurate opinion of the true sensitivity by the end of the experiment. By the last session, the opinions have polarised for the mass algorithm into 'Too sensitive' and 'Not sensitive enough'. This is probably due to differing interpretations of sensitivity. They are possibly attempting to convey the same point; that the algorithm is producing too many false prompts (too sensitive) while not producing enough true prompts (not sensitive enough). This is supported by results from the questionnaire data [91] and the following question. The responses for the microcalcification algorithm have not changed, the 'Too sensitive' responses are due to the prevalence of vascular calcification being highlighted as suspicious clusters.

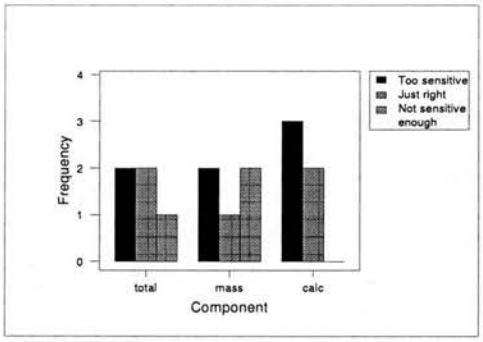


Figure 4.18: *Q3 Perception of sensitivity after first prompted session*

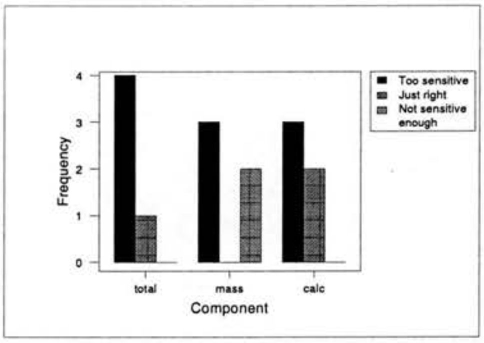


Figure 4.19: *Q3 Perception of sensitivity after last prompted session*

Q4 Please rate the system's specificity

The responses to this question (figures 4.20 and 4.21) indicate that the radiologists believe that the algorithms are prompting far too many false positives, although this may also be subject to misinterpretation of the question. In figure 4.21, there is an incidence of 'Too specific' in the mass algorithm category. This is possibly due to the fact that the algorithm is not detecting enough suspicious features, i.e. it is defining features as normal that are not - hence, too specific, although that should be a function of the previous question.

It is difficult, in computer-aided detection, to completely separate sensitivity and specificity. As mentioned earlier, the usual metrics for measuring performance are only partially valid and are used only rarely.

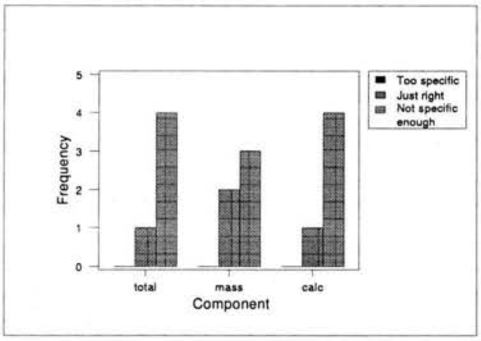


Figure 4.20: *Q4 Perception of specificity after first prompted session*

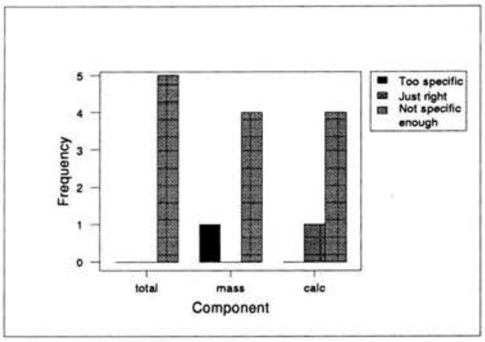


Figure 4.21: *Q4 Perception of specificity after last prompted session*

Q5 Roughly, for what percentage of prompts have you had difficulty in being able to:

Locate the prompted region on the mammogram? (figure 4.22)

Understand why the system has prompted for a particular area? (figure 4.23)

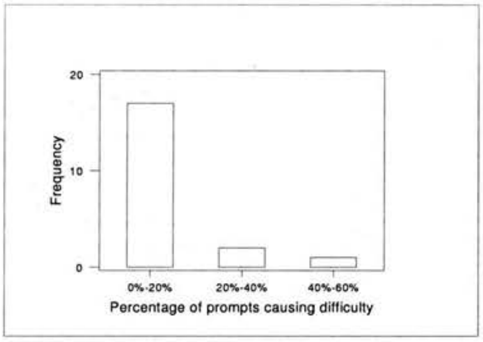


Figure 4.22: *Q5 For what percentage of prompts have you had difficulty in being able to locate the prompted region on the mammogram?*

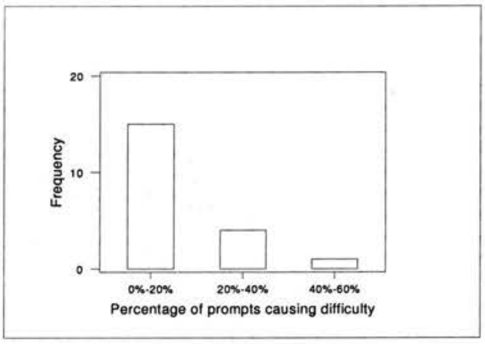


Figure 4.23: *Q5 For what percentage of prompts have you had difficulty in being able to understand why the system has prompted for a particular area?*

In the majority of sessions, the prompts were considered fairly easy to locate and interpret.

4.11 Conclusion

This experiment was the first time a fully working system had been exposed to radiologists and vice versa. Despite the lower than hoped for sensitivity of the mass algorithm, the developers were pleased with the performance of the system on such a large sample of films, as nothing on this scale had been attempted before. Technically, it was a great success.

From the point of view of its purpose, that of improving cancer detection, this too was a success. Of the five radiologists involved in the experiment, four improved their detection rate. Examining the results from the fifth radiologist, it was found that of the seven cancers that s/he had missed while prompted, four had been correctly prompted. Had s/he recalled these cancers, it would have led to an overall (although not significant) improvement in cancer detection when prompting was in use.

Overall recall rates also declined, although the individual results were more varied. Since a major concern of the radiologists had been the potential for the system to increase the number of recalls made, these fears were somewhat allayed by the results. As it is believed that radiologists behave slightly differently in experimental situations than in a clinic, where there is pressure to keep the recall numbers as low as possible, it is feasible that the recall rate would be even lower in a true clinic setting.

Another concern had been the time taken to read a prompted session. Although the differences between the time taken to read a prompted session and the time to read an unprompted session were very different at the start of the experiment, they became closer as the number of sets the radiologist was exposed to increased. By the end of the experiment they were reasonably similar, although the prompting sessions always remained longer. Differences in reading behaviour accounts for much of this, and as radiologists become more familiar with the system, it is possible that more efficient ways of reading with the prompts would be introduced.

This experiment was designed to be the pilot stage before the multi-centre full scale trial of approximately 90,000 women. Unfortunately, this was not to be. The reasons for this, and the steps that were taken to address the problem of having no data from a full-scale trial will be examined in the next chapter.

Chapter 5

Simulation of the PROMAM system

5.1 Aims and Objectives

Following the loss of funding that had been anticipated for the six centre trial of the PROMAM system, this chapter aims to simulate the system in a clinical setting, using a C program to emulate the behaviour of readers when confronted with 'films' that are either normal or contain a cancer, both with and without the prompting system.

In order to describe the program in detail, the component parts will be discussed in some depth. The random number generator, a standard set of C commands, is explained, as is the method of decreasing the processing time when confronted by the trigonometric functions of the Box-Muller transformation. The parameters are also examined, with an explanation for their presence in the model.

Beyond this, the program itself is illustrated in the form of a pseudo-program, where the complexities of the C programming language have been simplified to describe the functions of the simulation.

The results of each simulation are analysed as they would have been in a real clinical trial, with other methods also examined to investigate whether a more complex method of analysis is necessarily an advantage over a more standard approach.

Parameter settings have been varied in order to achieve an ever closer approximation to a 'real' situation, while each stage is examined to determine what effect this increasing complexity has on the results of the analyses. Power curves have been generated for each level of complexity, illuminating the differences in power for each method of analysis

In short, this chapter aims to examine the different outcomes that are achieved by a variety of analysis methods, and the main advantages and disadvantages that are inherent in each method.

5.2 Introduction

Due to circumstances outwith the project's control, funding for the full six centre trial was not forthcoming. Attempts were made to finance the trials through other funding bodies, including forming a company with financial backers from industry. However, the complicated partnership arrangement which, up until that point had supported the project fully, made this extremely difficult to set up in the time allotted and, ultimately, the attempt failed. An attempt to resurrect the project and take the system to trial unfortunately also failed, and PROMAM is once more in limbo.

And hence, this simulation.

5.3 The random number generator

A random number generator lies at the heart of every simulation model that includes any form of unpredictability (such as whether the next film in a simulated series will contain a cancer). True randomness is not possible in computer models, however, and any software that claims to produce random numbers is, in reality, producing 'pseudo-random' numbers. These are not strictly random, since if the seed value and the formula are known, it is possible to calculate the series of numbers that the generator will supply. However, the numbers generated *appear* to be unrelated and random, and would be Uniformly distributed between zero and one.

All pseudo-random number generators are cyclic (i.e. they return to the initial value and repeat) and the number of values generated before the cycle repeats is known as the period. Hence, a good generator will have a large period before it repeats. L'Ecuyer's generator, (from Numerical Recipes in C [96]) with Bays-Durham shuffle (where the cells in an array are filled with random numbers, which are then randomly called and replaced) has a period of more than 2×10^{18}.

5.4 The Normal deviate generator

In the 'real world', many distributions tend to the Normal if given a long enough period. In this study, an assumption is made that each radiologist has a mean cancer detection rate, and that their actual detection rate on any particular batch varies about the mean with a Normal distribution and a given standard deviation. Given this assumption, we need to generate random numbers from a Normal distribution, rather than a Uniform distribution.

The unit Normal deviates, N(0,1), are traditionally produced from pairs of random unit Uniform deviates (x_1, x_2) using the Box-Muller transformation (equations 5.1 and 5.2).

$$y_1 = \sqrt{-2\ln x_1}\cos 2\pi x_2 \tag{5.1}$$

and

$$y_2 = \sqrt{-2\ln x_1}\sin 2\pi x_2 \tag{5.2}$$

However, this method is computationally intensive, requiring multiple calls of trigonometric functions. A common trick when faced with this problem, is to use v_1 and v_2 as the co-ordinates of a random point within a unit circle around the origin. Thus, if we have two successive Uniform deviates (z_1 and z_2) from our random number generator, we can define v_1 and v_2 as

$$v_1 = 2z_1 - 1 \tag{5.3}$$

and

$$v_2 = 2z_2 - 1 \tag{5.4}$$

where v_1 and v_2 are Uniformly distributed between -1 and +1. If we then discard points outside the unit circle, the distance from the origin to the point (v_1, v_2) is R, where $R^2 = v_1^2 + v_2^2$ and is Uniformly distributed between 0 and 1. Additionally, the angle that the point (v_1, v_2) defines with respect to the v_1 axis may be given as θ. As this is simply a Uniformly distributed angle between 0 and 2π, it may be substituted in equations 5.1 and 5.2 for $2\pi x_2$, yielding

$$\begin{aligned} \cos\theta &= \frac{v_1}{R} \\ &= \frac{v_1}{\sqrt{v_1^2 + v_2^2}} \end{aligned} \tag{5.5}$$

Similarly,

$$\sin\theta = \frac{v_2}{\sqrt{v_1^2 + v_2^2}} \tag{5.6}$$

Returning to equations 5.1 and 5.2, we may now generate the Normal (0,1) random deviates by

$$y_1 = \sqrt{-2\ln(v_1^2 + v_2^2)}\frac{v_1}{\sqrt{v_1^2 + v_2^2}} \tag{5.7}$$

and

$$y_2 = \sqrt{-2\ln(v_1^2 + v_2^2)}\frac{v_2}{\sqrt{v_1^2 + v_2^2}} \tag{5.8}$$

thus circumventing the need for trigonometric function calls and vastly increasing the speed of the algorithm.

5.5 Parameters

A number of factors can influence the overall cancer detection rate of a screening programme. Table 5.1 presents the parameters that have been included in this simulation model. The rationale for these decisions follows the table.

Although reading order has not been selected as one of the parameters, and no effect will be imposed, it will be checked for in the analysis of the results.

Parameter	Coding	Distribution
Centre	`centreno`	Multinomial, P(i)=$\frac{1}{6}$, i=[0,5]
Number of readers	`maxrad`	{3,4,5,2,3,3} depending on `centreno`
Batch size	`batchsize`	Square of Normal $\mu = 6.97$, $\sigma^2 = 8.08$
TN(prompted) μ, σ^2	`prob[i][0], prob[i][1]`	Normal, μ_{i0}, σ^2_{i0}
TP(prompted) μ, σ^2	`prob[i][2], prob[i][3]`	Normal, μ_{i1}, σ^2_{i1}
TN(unprompted) μ, σ^2	`prob[i][4], prob[i][5]`	Normal, μ_{i2}, σ^2_{i2}
TP(unprompted) μ, σ^2	`prob[i][6], prob[i][7]`	Normal, μ_{i3}, σ^2_{i3}
Cancer rate	`hit`	Fixed (estimated to be 0.6%)

Table 5.1: Parameters used in the simulation program

5.5.1 Centre

As the original full-scale trial was to have been a multi-centre trial of six centres, this simulation also has six centres, containing differing numbers of readers, as would have been the case during the trial. The number of readers per centre is fixed, as each reader requires specified data on their TN and TP probabilities.

5.5.2 Number of readers

The minimum number of readers in a given run must not be less than two, as the simulation requires that at least a majority of films are double read. The values chosen were arbitrary, although some centres do function with as few as two readers. The value five was chosen as the maximum, corresponding to the number of available readers at Ardmillan House, Edinburgh.

5.5.3 Batch size

The batch size has been based on data gathered over a period of 7.5 months (from 22 January to 1 August 1997) from the Breast Screening Centre in Edinburgh. The data have been collated from 613 batch reading slips to 190 batch reading pairs. This is due to many of the smaller sub-batches being read by the same readers on the same day. Since these smaller sets were read as part of a batch of sets, it is the batch that we are interested in. For example, the raw data from table 5.2 can be reduced to a smaller number of batches, such as in table 5.3.

Screened	Type	Number	Reader 1	Date 1	Reader 2	Date 2
23/01/97	T	14	AEK	31/01/97	MEC	31/01/97
23/01/97	M	12	AEK	31/01/97	JM	31/01/97
22/01/97	M	38	AEK	31/01/97	JM	31/01/97
23/01/97	M	59	MEC	31/01/97	AEK	31/01/97
27/01/97	T	7	BBM	05/02/97	JSW	06/02/97
28/01/97	M	24	JSW	05/02/97	BBM	05/02/97
27/01/97	L	2	JSW	05/02/97	BBM	05/02/97
28/01/97	T	10	BBM	05/02/97	JSW	06/02/97
28/01/97	M	52	JSW	05/02/97	BBM	05/02/97
27/01/97	T	4	BBM	05/02/97	JSW	06/02/97
28/01/97	M	16	BBM	05/02/97	JSW	06/02/97
27/01/97	S	47	BBM	05/02/97	JSW	06/02/97

Table 5.2: Sample of data from Edinburgh Breast Screening Centre (T=technical recall, M=screened at a mobile unit, S=screening at the static unit)

The collation of these 613 batch slips into 190 batch pairs resulted in a heavily skewed distribution (see figure 5.1)

Although neither of the standard transformations (square root and logarithm) gave completely straight lines when drawn on a probability plot, the square root transformation was slightly better than the logarithm (see figure 5.2 for the normal plot of the square root transformed data). There were several large values, with four being greater than 200. Given that the next lowest value was 165, and that it would be very unlikely that any radiologist would read that many sets of films in one sitting; it is possible that these resulted from more than one batch being read by the same pairing on the same day. Unfortunately, due to the nature of the raw data, it is impossible to confirm this. Hence, the mean and standard deviation of the transformed data were calculated with

Number	Reader 1	Date 1	Reader 2	Date 2
14	AEK	31/01/97	MEC	31/01/97
50	AEK	31/01/97	JM	31/01/97
59	MEC	31/01/97	AEK	31/01/97
84	BBM	05/02/97	JSW	06/02/97
78	JSW	05/02/97	BBM	06/02/97

Table 5.3: Collated batch data

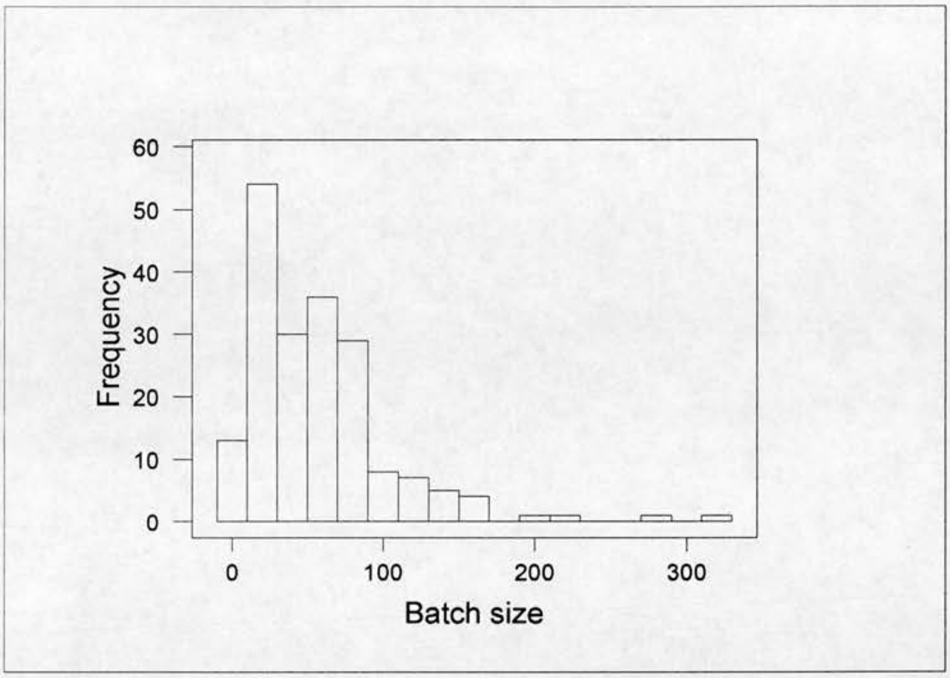


Figure 5.1: Frequency of batch sizes

the extreme values. Removing them had only a small effect on the resulting statistics.

Thus, for simplicity, the square root transformation to the Normal distribution was taken, with a mean of 6.97 and standard deviation of 2.84

5.5.4 True positive and true negative rates

There are four possible outcome probabilities for a reader viewing a subject's films: recording a malignant subject as positive when a cancer is present and the reader is prompted (TP(prompted)), recording a non-malignancy as normal when the reader is prompted (TN(prompted)), recording a malignancy as positive when a cancer is present and the reader is not prompted (TP(unprompted)), and finally, recording a non-malignancy as normal when unprompted (TN(unprompted)). Given that each reader may behave differently each time they read a set of films, these probabilities have been simulated to have truncated Normal distributions. Each reader has a set of eight values that are preset and in a pre-determined order: TN prompted mean (TNp

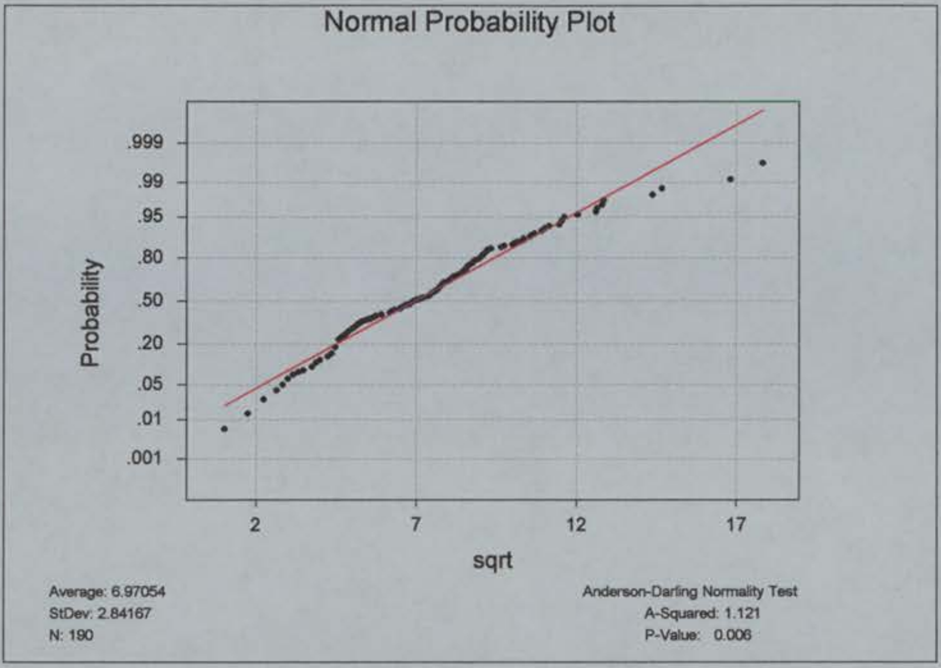


Figure 5.2: Normal probability plot of the square root transformed data

μ), TN prompted standard deviation (TNp σ), TPp μ, TPp σ, TNu μ, TNu σ, TPu μ, TPu σ. To avoid extreme TP and TN rates, the random values will be truncated at three standard deviations; i.e. no value will be lower than $\mu - 3\sigma$ or greater than $\mu + 3\sigma$. Occasionally, the upper limit may exceed 1.0 if the value of μ is large, but this will be uncommon and be treated as if the value were 1.0, giving that reader a 100% chance of detecting cancers in that batch.

It is these values and their relative magnitudes that will be altered during these simulations. Several models will be examined, each becoming successively more complex.

Although all the simulated cancers will be known, the analysis will treat the data as 'real'. In other words, only *detected* cancers will be analysed as cancers, with cancers missed by both readers being labelled as normal. This will give higher estimates for TP and TN rates than is actually true.

5.5.5 The underlying cancer rate

The figures published by the NHS Breast Screening Programme at the time of creating the simulation [15] put the cancer detection rate at 5.9 per 1000 women (0.59%). For simplicity, the figure in the program has been fixed at 0.6%. Thus approximately six 'cancers' should be produced by the program for every 1000 'women' generated. This figure does not take into account the cancers that will inevitably be missed by both readers, so using a cancer rate of 0.6% will produce a detected cancer rate closer to the 0.59% calculated by the NHSBSP.

5.6 The simulation program

The simulation program can be thought of as three nested sections for simplicity.

1. Level 1, the control level, sets all counters to zero, reads in the probability values for the TP and TN rates, and sets the underlying cancer rate. This level is only called once for each run of the program.

2. The second level is where most of the processing occurs. The size of a batch of subjects is generated from the square of N(6.97, 8.08) A centre is randomly selected from the six available, thus dictating which readers are available. The two readers are selected from those available and assigned to read first or second. Their simulated sensitivity and specificity for this batch are also generated from their corresponding mean and standard deviation. And finally, prompting is assigned to first or second reader using the minimisation method, to retain balance between the readers, so that some readers are not being prompted more or less than others. The values for these parameters changes for each batch, although they remain constant within a batch.

3. Finally, the core routine. This is where the individual subjects are created and where the accuracy of the readers is tested. The result for each subject is recorded,

along with the identification of the readers, who was prompted and to which centre the readers belonged.

The following is a pseudo-program that defines the program in language that is easier to follow than the programming language itself (C), where *rad* refers to the unprompted radiologist, and *PROMAM* refers to the radiologist prompted by the PROMAM system. The 3 levels of the program are separated by spaces. The program itself is in Appendix L.2.

Section 1: Sets the parameters (hyperparameters for the TP and TN rates were preset for the individual readers)

```
define all counters==0
read prob[i][j] for all i=0,19, j=0,7 from data set of probabilities
let hit=0.006
let centres={3,4,5,2,3,3}
let runsize=100000
```

Section 2:

(1) Generates a Normal unit deviate to create the size of the batch.

(2) Samples a random centre

(3) Randomly selects first reader from the selected centre

(4) Presets counter for number of times reader1 is first prompted and unprompted. This is later used in the minimisation routine

(5) Randomly selects second reader from the same centre as the first, checking that first and second readers are different

(6) Compares the number of times prompting has been given to the first reader with the number of times it has been given to the second reader - the minimisation routine (see appendix for details) - and assigns prompting to the lowest total. If totals are equal, then prompting is randomly assigned

(7) The TP and TN rates are generated for the readers of this batch (and only this

batch)

```
generate random number=random (1)
generate Normal unit deviate from random=gasdev
let batchsize=(gasdev*2.843+6.97)^2
generate random number=random (2)
let centreno=int(random*6)
let maxrad=centres[centreno]
let i=1 to centreno
   radiologists=radiologists+centres[i]
generate random number=random (3)
x=int(random*maxrad)
reader1=x+radiologists
set counter if reader1 prompted, reader1_P (4)
set counter if reader1 unprompted, reader1_U
do (5)
   generate random number=random
   x=int(random*maxrad)
   reader2=x+radiologists
   set counter if reader2 prompted, reader2_P
   set counter if reader2 unprompted, reader2_U
while (reader1==reader2)
let first=sum if prompting first (6)
let second=sum if prompting second
if first==second, generate random number, random
   if random<=0.5 then prompt=1
   else prompt=2
elseif (first<second) then prompt=1
else prompt=2
```

```
increment counters
generate probabilities for TN & TP for unprompted and prompted (7)
readers, TN_U, TP_U, TN_P, TP_P
```

Section 3: A pseudo-subject is created and assigned to have a cancer or to be normal. A random number is generated for each reader to substitute for their ability to detect a cancer/return a normal. The results are recorded, with 0 for 'not a cancer' and 1 for 'cancer'

```
Let j=1 to batchsize
    generate random number=random
    if random <= hit then batch[j]=1 (cancer)
       generate random number=random
                if random <= TP_U then rad_recall[j]=1 (cancer)
                else rad_recall[j]=0 (normal)
       generate random number=random
                if random <= TP_P then PROMAM_recall[j]=1 (cancer)
                else PROMAM_recall[j]=0 (normal)
    else batch[j]=0 (normal)
        generate random number=random
                 if random <= TN_U then rad_recall[j]=0 (normal)
                 else rad_recall[j]=1 (cancer)
        generate random number=random
                 if random <= TN_P then PROMAM_recall[j]=0 (normal)
                  else PROMAM_recall[j]=1 (cancer)
    print results
next j
```

The following stops the program once the runsize has been exceeded, while allowing a batch to be completed

```
if number of total subjects < runsize go to next batch
else stop
```

5.7 Presenting the data

It was decided to compare two methods of presenting the data to the various analyses.

1. from the summary results of the simulation, where each reader has a set of values comprised of the prompted TP and TN rates and unprompted TP and TN rates for that reader over all women seen by that reader

2. using the data as generated by the simulation, on a woman by woman basis

This was done to examine whether the loss of information that the summary results represented would influence the ability of the analyses to detect a significant difference between the prompted and unprompted readers. However, the data recorded in a true clinical setting would be on an individual basis, and so it is likely that only the latter method would ever be used in a true analysis.

5.8 Methods of analysis

Several methods of analysis will be considered and compared - from the simple McNemar's test, to more complex mixed models analysis, in order to determine whether a more fully specified analysis will give more precise results.

In order to compare these methods, the simulation will be run a number of times and the p-value of the hypothesis that there is no difference between the prompted and unprompted readers will be calculated. The proportion of these p-values that fall beneath 5% will then be plotted against the hyperparameter difference, as determined prior to the simulations. This will generate a selection of power curves.

5.8.1 Simple model

The most simple method of analysing these data is McNemar's test (see also page 36) [84]. The TP (or FP) rates for prompted and unprompted readers will be compared without reference to other factors that may influence their ability to detect a cancer. Only the second method of presenting the data (woman by woman) may be analysed this way. The SAS code to generate the result of the McNemar's test is given below, where promam and rad are the prompted and unprompted reading conditions. The option *agree* in the table statement produces the actual test.

```
proc freq;
table promam*rad / nocol norow nopercent agree;
output out=canc mcnem;
```

5.8.2 Fixed effects models

The next level of complexity was a fixed effects, generalised linear model (see page 49 for detailed explanation). As before, the model is composed of a response variable, explanatory variables and a link function, which in this case is the logit function. reader identifies the individual readers (0–19), paper defines whether a prompt was present (1) or absent (0), order is the reading order - whether the reader was first (1) or second (2), and paper|reader is the paper/reader interaction, whether the radiologists responded differently to the prompting.

PROC GENMOD was used to analyse the results in both of the following cases.

1. analysis of summary statistics (TP or FP rates)

```
ods output ParameterEstimates=test;
proc genmod data=summary;
class reader paper order;
model count/total = paper reader order paper|reader/ dist=b type3 wald;
```

```
lsmeans paper / diff;
```

where *count* was the number of correctly detected cancers (correctly returned normals) and *total* was the total number of cancers available to detect (total number of subjects).

2. analysis of individual women

```
n=1;
ods output geeemppest=test;
proc genmod data=complete;
class reader paper order run;
model recall/n = paper reader order paper|reader/ dist=b type3 wald;
repeated subject=run / type=CS;
lsmeans paper / diff;
```

`run` is the identification number of each woman, which allows us to treat the data as repeated measures. `recall` defines whether a woman was recalled (1) or not (0).

5.8.3 Random effects models

Mixed models allow us to model not only the fixed effects, but their underlying covariance structure by specifying random effects other than the residual. For example, centre random variation in a multi-centre trial. This is a relatively new field, coming to prominence with the increasing computing power available [90]. In this application, we not are interested in the individual readers' cancer detection ability, but it does need to be modelled in order to establish whether any perceived differences between the prompted and unprompted results are due solely to the effect of the prompts. And so, in the code below, we have `paper` (the prompting system) and `order` (the order in which readers saw the 'films', i.e. whether a reader read first or second) as fixed effects, and `reader` and `paper*reader` as random effects. The interaction term will invariably lead to larger confidence intervals around the estimates of treatment difference.

1. analysis of summary statistics

   ```
   %glimmix(data=summary,
   stmts=%str(class reader paper order;
   model count/total = paper order / ddfm=satterth;
   random reader paper*reader;
   lsmeans paper/ diff pdiff;
   title 'GLIMMIX - summary';),
   error=b);
   ```

2. analysis of individual women

   ```
   n=1;
   %glimmix(data=complete,
   stmts=%str(class reader paper order run;
   model recall/n = paper order /
   ddfm=satterth;
   random reader paper*reader;
   repeated order / subject=run type=cs;
   lsmeans paper/ diff pdiff;
   title 'GLIMMIX - full';),
   error=b);
   ```

The variables have the same definition as those in the fixed effects model. %glimmix is a macro written by Russ Wolfinger and Jason Brown of the SAS Institute Inc. for fitting generalised linear mixed models using PROC MIXED and the Output Delivery System, and may be obtained from the SAS website [97].

5.8.4 Running the simulation

A shell script was provided by Dr Rob Blake to run the set of commands to create first the set of approximately 100,000 simulated women, followed by the analysis of these data by the above SAS commands. The repeated analyses of these simulations was time consuming, requiring multiple calls of the PROC MIXED command within the %glimmix command. The script is included in Appendix L.

The analyses were submitted to two machines available to me; verum (the Medical Statistics Unit's main server) and waverley (actually a farm of machines, dedicated to heavy processing programs). Unfortunately, it took on average 40 hours on verum and 59 on waverley to simulate the 100,000 women and then analyse the approximately 600 cancers this generated, 1000 times. When examining the recalls, this took considerably longer, with just one of the 1000 required runs of the simulation/analysis taking approximately two hours to complete.

5.9 Hyperparameter selection

In the first instance, only the difference between the prompted and unprompted readers will be examined, where we assume no difference between readers. Later, we will look at the effect of differences between readers, then at the interaction of reader and prompting effects.

5.9.1 The Null model

For this model, which is used for testing purposes only, the hyperparameters are all set to 90% and 2.5%, i.e. each individual radiologist's TP and TN rate is generated from a Normal distribution with mean 0.9 and standard deviation 0.025. From this, the analysis should show no difference between prompted and non-prompted states.

5.9.2 Changing the accuracy of reading methods

The simplest comparison is to keep the TP and FP hyperparameters fixed over readers, but vary the difference between the prompted and unprompted hyperparameters. As this is merely an illustration, and that to examine all the possible combinations of TP and FP rates would take a long time, the unprompted reader hyperparameters will remain fixed at 90% and the prompted reader hyperparameters will be fixed at an alternative level in any simulation. The standard deviations of each individual reader's hyperparameter will remain at 2.5%.

5.9.3 Changing the accuracy of the readers

This next model allows the unprompted TP rates to vary across readers with $N \sim (0.9, 0.025^2)$, with the prompted TP rates a fixed increase on the unprompted rates. This will allow the unprompted readers to have different TP rates, but the difference between prompted and unprompted rates within readers will be fixed and will be the same for all readers within a particular simulation.

5.9.4 Changing the responses of the readers

In reality, it is unlikely that every reader will respond to the prompting system in the same way. It is more plausible that some readers will find the prompts to be a great help, while others will find them distracting, which will ultimately affect their cancer detection ability. Therefore, the unprompted rates will vary with $N \sim (0.9, 0.025^2)$ as before, while a difference will vary $N \sim (x, 0.03^2)$, where $x = 0$ to 6%. The prompted rate will then be the sum of the unprompted rate and the difference for each reader.

5.10 Results - part 1: a single analysis

In order to describe the statistical analyses used and the output generated by the SAS procedures for these analyses, an example data set has been generated. This will be analysed with each of the methods under consideration, and the results discussed in this section. For the illustration, the more realistic simulation model has been used to generate the example data, where TPu has a mean of 90% and standard deviation of 2.5% for cancer detection, and the difference between TPu and TPp is 4% with a standard deviation of 3%. In other words, TPp = TPu + difference.

PROC GENMOD had great difficulty with these data, and was unable to estimate the GEE parameters when there were uniform effect categories - in other words, when a radiologist correctly identified all the 'cancers'[1] to which they were exposed. Naturally, this event occurred more frequently as the individual probability of detecting a cancer increased. In this example, where the average probability of detecting a prompting-assisted cancer was 94%, a large number of simulations failed to resolve the additive repeated measures analysis. The data analysed in the following sections are from an example where the analysis did converge with PROC GENMOD.

The full output from the analyses may be found in Appendix M.1 and M.2.

5.10.1 McNemar's Test

		Unprompted		
		Not recalled	Recalled	Total
Prompted	Not recalled	0	30	30
	Recalled	50	497	547
	Total	50	527	577

Table 5.4: Results of the McNemar's test (cancers)

The results of this simulation example are summarised in table 5.4. The value of TP prompted was 547/577 (94.8%) and TP unprompted was 527/577 (91.3%), an actual difference of 3.5%. The p-value for the Null Hypothesis of no difference was 0.025 (95% CI for the difference of 0.4%, 6.5%). The not-recalled/not-recalled cell is empty as, in practice, it would not be possible to know which cancers had been missed by both readers.

5.10.2 Generalised Linear Model

Below are the results of the additive analysis model (no interaction term) as produced by PROC GENMOD. The p-value of the difference between the prompted and unprompted radiologists is significant at 0.02 (as can be seen from both the Wald statistics and the parameter estimates).

[1] Only cancers that were detected by one or more radiologist

```
            Wald Statistics For Type 3 GEE Analysis

                                        Chi-
          Source              DF      Square     Pr > ChiSq

          paper                1        5.72         0.0168
          reader              19       17.59         0.5503
          order                1        0.59         0.4407
```

```
                 Analysis Of GEE Parameter Estimates
                 Empirical Standard Error Estimates

                              Standard   95% Confidence
Parameter       Estimate       Error          Limits              Z Pr > |Z|

Intercept         2.2489     0.4159    1.4336    3.0641     5.41    <.0001
paper       0    -0.6006     0.2512   -1.0930   -0.1082    -2.39    0.0168
paper       1     0.0000     0.0000    0.0000    0.0000      .        .
reader      0     0.7153     0.6350   -0.5293    1.9599     1.13    0.2600
reader      1     0.2369     0.5793   -0.8985    1.3723     0.41    0.6826
reader      2     0.3317     0.6083   -0.8605    1.5239     0.55    0.5855
reader      3     0.4089     0.6555   -0.8760    1.6937     0.62    0.5328
reader      4     1.8789     1.0244   -0.1289    3.8866     1.83    0.0666
reader      5     1.2961     0.8119   -0.2952    2.8873     1.60    0.1104
reader      6     0.7391     0.7294   -0.6904    2.1687     1.01    0.3109
reader      7     0.2412     0.7221   -1.1740    1.6564     0.33    0.7383
reader      8     1.6269     1.0982   -0.5256    3.7794     1.48    0.1385
reader      9     1.9439     1.1110   -0.2336    4.1214     1.75    0.0802
reader     10     0.9850     0.7675   -0.5192    2.4892     1.28    0.1993
reader     11     0.2610     0.6586   -1.0298    1.5517     0.40    0.6919
reader     12     1.0620     0.6297   -0.1721    2.2961     1.69    0.0917
reader     13     0.4295     0.5545   -0.6573    1.5163     0.77    0.4386
reader     14     0.0493     0.5729   -1.0736    1.1723     0.09    0.9314
reader     15     0.0500     0.5559   -1.0395    1.1395     0.09    0.9284
reader     16     0.1045     0.5585   -0.9902    1.1992     0.19    0.8515
reader     17     1.0811     0.6993   -0.2895    2.4516     1.55    0.1221
reader     18     1.5054     0.8291   -0.1196    3.1304     1.82    0.0694
reader     19     0.0000     0.0000    0.0000    0.0000      .        .
order       1     0.1901     0.2466   -0.2932    0.6734     0.77    0.4407
order       2     0.0000     0.0000    0.0000    0.0000      .        .
```

As mentioned earlier, this particular data set was chosen in order to produce an analysis that would converge. Had the analyses included the cancers that were detected by neither reader, it is possible that fewer simulations would fail to converge. However, as it is unlikely that these cancers that were missed by both radiologists would be known during a 'real' analysis of clinic-produced data, the simulation and subsequent analysis emulates this situation.

Estimates of the overall cancer detection rates may be drawn from the Least Square Means output, where the LS estimate may be converted by the inverse application of the logit link. To be precise:

$$\mu = \frac{e^{LSE}}{1 + e^{LSE}} \tag{5.9}$$

where LSE is the Least Squares Means estimate.

```
                         Least Squares Means

                                 Standard               Chi-
Effect     paper  Estimate          Error       DF     Square     Pr > ChiSq

paper      0        2.4906         0.1723        1     208.86         <.0001
paper      1        3.0913         0.2030        1     231.90         <.0001

                  Differences of Least Squares Means

                                    Standard            Chi-
Effect   paper   _paper   Estimate      Error      DF   Square   Pr > ChiSq

paper    0       1         -0.6006     0.2512       1     5.72       0.0168

                     Differences of Least Squares Means

                  Effect   paper   _paper    Confidence Limits

                  paper    0       1          -1.0930     -0.1082
```

In this case, the estimate for the detection rate of the unprompted radiologists ($\texttt{paper}_0$) is 92.35%, and for the prompted radiologists ($\texttt{paper}_1$) is 95.65%, a difference of 3.3%. These estimates are different to those from table 5.4, due to the method of standardisation used in the Least Squares Means estimation.

Although the results for the additive model analysis did converge, the analysis did not converge for the individual woman data where the interaction term had been included.

Analysis of the summary data was also performed. The simple additive model did converge (p = 0.013), but the analysis of the summary data including the interaction term was suspect; for these data, the p-value for no difference between prompted and unprompted conditions is given as 0.67. Examination of the log file clearly indicated that the validity of the model fit was questionable at best. A result, however, was generated, which is extremely misleading. Had there not been other results with which to compare this one, it may have easily been accepted as a true result.

5.10.3 Mixed models

The mixed models analysis proved to be more robust when dealing with uniform effect categories. The output from fitting a model with fixed effects of `paper` and `order`, and random effects of `reader` and `reader*paper` is shown below.

```
                 Covariance Parameter Estimates

              Cov Parm          Subject    Estimate

              reader                        0.02047
              reader*paper                        0
              CS                run        -0.07183
              Residual                       1.0591
```

```
                        Solution for Fixed Effects

                                        Standard
Effect      paper  order  Estimate        Error      DF    t Value   Pr > |t|

Intercept                   2.8224       0.2174     171      12.98     <.0001
paper       0              -0.5564       0.2460     592      -2.26     0.0240
paper       1                    0            .       .          .          .
order              1        0.1803       0.2398     596       0.75     0.4524
order              2             0            .       .          .          .
```

```
                   Type 3 Tests of Fixed Effects
                      Num       Den
            Effect       DF        DF     F Value     Pr > F

            paper         1       592        5.12     0.0240
            order         1       596        0.57     0.4524
```

In this example, the reader*paper covariance parameter in the mixed model was estimated to be zero, and therefore had no effect on the estimates of standard error. Hence, there is no difference between the main results of the additive and interaction models, and so only the interaction model is illustrated. It is not always the case that the reader*paper covariance parameter is zero, but as the selection of the dataset was constrained by the ability of PROC GENMOD to converge, it was allowed to stand as an example. A zero (or negative) variance component may be caused by there being less variability between readers than would be expected by chance. Since a negative variance component is not permitted by the underlying model, it will usually be fixed at zero.

```
                          Least Squares Means

                                       Standard
Effect     paper    Estimate            Error         DF     t Value  Pr > |t|

paper      0          2.3561           0.1508       45.2       15.62    <.0001
paper      1          2.9125           0.1898        106       15.35    <.0001

                    Differences of Least Squares Means

                                        Standard
Effect  paper  _paper  Estimate     Error    DF    t Value   Pr > |t|

paper   0      1        -0.5564   0.2460   592      -2.26   0.0240

             Effect  paper  _paper       Lower        Upper

             paper   0      1          -1.0395     -0.07340
```

Examining the Least Squares Means again, we can see that the standard error for the difference between prompted and unprompted radiologists (on the linear scale) is smaller for the mixed model than for the generalised linear model, while the estimate of the difference (calculated from the inverse transformed absolute values) is larger.

5.10.4 Summary of Results

Although only the additive model for PROC GENMOD and the interaction model for the mixed model have been illustrated above, table 5.5 contains the results for the interaction model in PROC GENMOD and the additive mixed model. As mentioned previously, PROC GENMOD failed to converge with the interaction model, and the p-value for the summary data is misleading, coming as it does from the initial parameter estimates.

Analysis	Model	Data	Difference	p-value
McNemar	Simple	Full	3.47%	0.0253
Generalised Linear Model	Additive	Full	3.30%	0.0168
		Summary	3.32%	0.0130
	Interaction	Full	DNC	DNC
		Summary	DNC	0.6696
Mixed Model	Additive	Full	3.51%	0.02040
		Summary	3.53%	0.01060
	Interaction	Full	3.51%	0.0240
		Summary	3.53%	0.0106

Table 5.5: Summary of results of analysis of sample simulation, where 'Full' was the set of data from individual women and 'Summary' was the summary of the results by radiologist*prompting*order (DNC = Did Not Converge)

5.11 Results - part 2: Power curves

The following section will illustrate the results when the data set of simulated films are repeatedly generated and analysed, for increasingly complex simulation models.

Given the processor intensive nature of these simulations, only the 'detected' cancers will be analysed in this way. Since the PROC GENMOD analysis consistently failed to converge when there were uniform effect categories in the interaction terms, it will not be used in the simulation. Although the additive model also had problems with uniform effect categories, not all analyses failed to converge, and so it will remain as one of the methods of analysis. A note will be made to indicate the number of failures at each simulation.

With only limited time available for the multiple simulation runs, the interaction term for the mixed models analysis was dropped from the first two simulation models, as it increased the computing time required to complete each set of 1000 simulations, and added little information to the results. It was included in the third model as a limited amount of interaction was built into the model, whereas the previous models had been designed as additive models only.

5.11.1 Changing the accuracy of the reading method

For this model, the hyperparameters for both prompted and unprompted readers were fixed, and only the difference between them was allowed to change between each set. In other words, TPu was fixed at 90% for all readers and TPp increased with respect to TPu, but was the same for each reader.

On examination of the tabulated results, it appeared that the majority of the results from the GENMOD procedure were the same for the full analysis and summary analysis. This led to the discovery that the p-value generated and output to the simulation record by SAS was the incorrect value. It was, in fact, the value of the initial parameter estimate, and not the GEE parameter estimate as was claimed. Attempts to re-analyse the data using SAS v6.12 failed, although the analysis would run under SAS v8.2. However, only verum (the local machine) hosted SAS v8.2, and was, at the time, processing considerably slower than waverley. Hence, it would not prove possible to re-simulate the entire set again, and so only the analysis in question has been re-done. Due to insufficient processing time, some analyses have been left undone, although the values of interest (power around 80%) have been covered. Thus the values for the full (individual women) analysis with PROC GENMOD are not from the same simulation runs as the other four analyses.

Figure 5.3 illustrates the typical sinusoidal shape of the power function. This curve was generated by simulating the set of 100,000 women and analysing the 'detected' cancers by the five methods discussed, 1000 times for differences of 0% to 6% (absolute) in steps of 0.5%. As mentioned earlier, this was very heavy in computing time, even using two mainframes simultaneously.

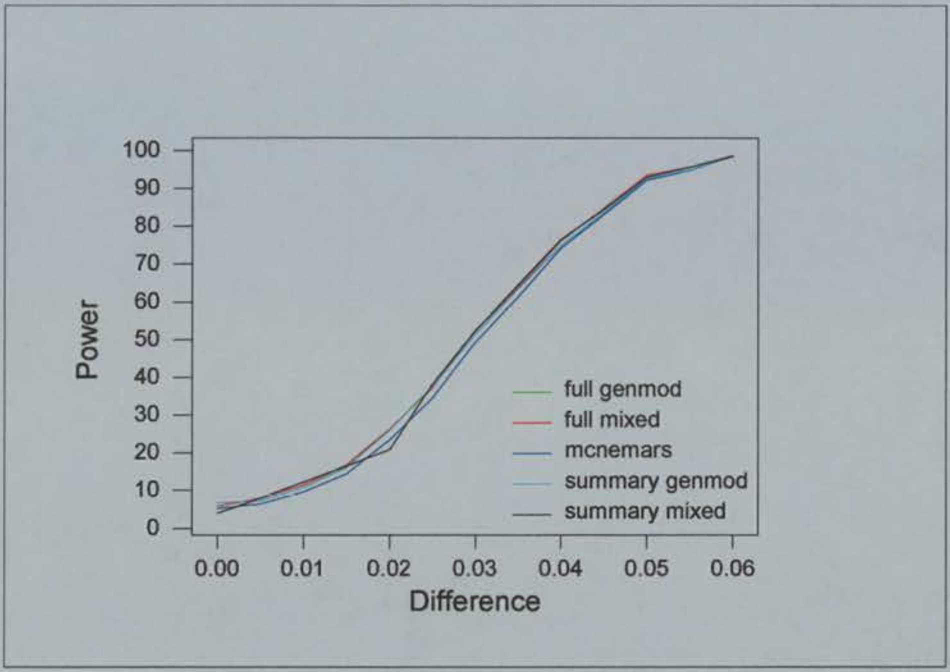


Figure 5.3: Power curve for the simple model (section 5.11.1)

Since the power curves are very closely plotted, the data have been tabulated (table 5.6). From this, it can be seen that the power exceeds 80% between a prompted increase in detection rate of 4% and 4.5%, as would be expected from the sample size calculations (see Appendix C). Included in the table are the percentages of results where the analysis did not converge (column 'Missing').

Given that the generation of the simulated data implies that the probability of a significant result should be 5% with a standard deviation of 0.475% when there is no difference, the 6.8% value that appears in the first line of the GENMOD:summary results seems extremely high ($p < 0.0001$), indicating that this approach may yield an inappropriately high proportion of significant results. However, further simulations on the same settings gave results more in line with the expected values, which suggests that this result was due to chance.

Absolute		GENMOD			MIXED	
difference	McNemar's	Missing	Full	Summary	Full	Summary
0.0%	5.2	23.4	3.8	6.8	5.9	4.0
0.5%	6.5	21.0	5.7	7.3	8.0	8.0
1.0%	9.7	24.3	10.4	10.8	11.5	12.3
1.5%	14.4	25.8	15.8	16.0	16.8	16.5
2.0%	23.5	28.1	24.2	26.1	26.4	20.9
2.5%	34.5	30.8	32.4	37.5	37.2	38.2
3.0%	49.4	32.6	47.9	51.5	52.6	52.4
3.5%	61.3	36.8	63.1	63.5	63.8	64.5
4.0%	74.3	39.0	75.1	75.0	76.3	76.6
4.5%	83.2	38.2	81.4	83.7	84.8	84.5
5.0%	92.3	42.6	92.5	92.7	93.7	93.1
5.5%	95.0	46.1	96.1	95.1	95.7	95.7
6.0%	98.7	48.4	97.7	98.6	98.8	98.6

Table 5.6: Results of the simple model (figure 5.3) - percentage of solutions that attained significance at the 5% level

5.11.2 Changing the baseline accuracy of the readers

In this simulation, the unprompted reader parameters were allowed to vary around a fixed hyperparameter (90% for TP and 89% for TN), and the difference between the prompted and unprompted conditions was the same for each simulated radiologist. As before, simulating the changes in recall rate would have been extremely time-consuming and so only the changes in cancer detection have been examined. The difference between TPu and TPp increased from 0 to 6% (absolute) in steps of 0.5% (see figure 5.4).

Again, the table of results is included to clarify the detail (table 5.7). As with the previous model, the GENMOD full analysis results have been calculated independently from the other four methods, with the percentage of failures to converge given in the table. It can be seen that the power exceeds 80% between 4% and 4.5% as before, although the proportions that are significant at the 5% level are lower in this table than in the previous table (table 5.6).

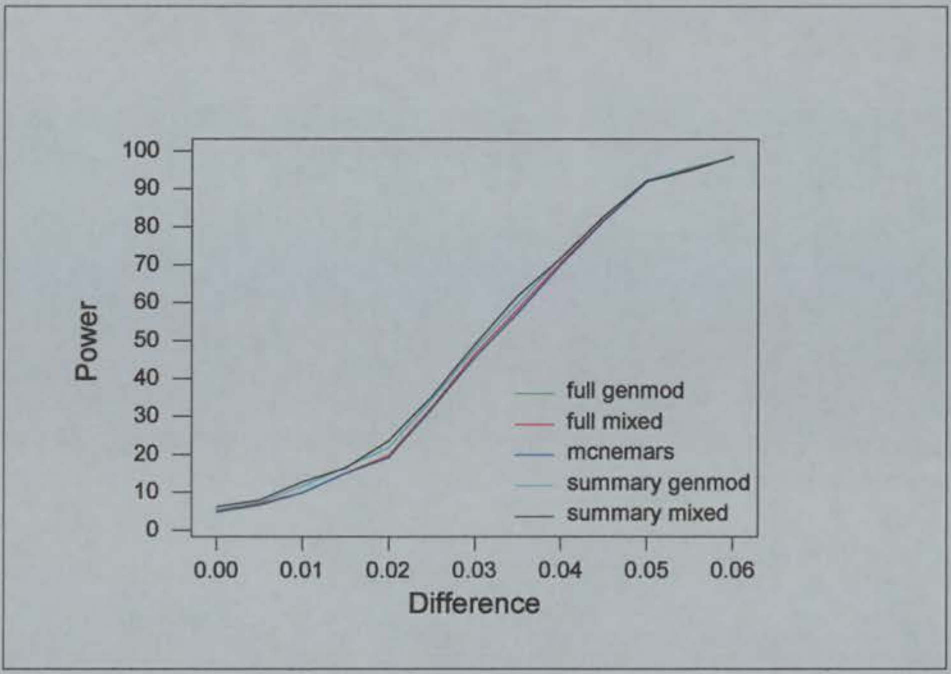


Figure 5.4: Power curve for the random unprompted reader and fixed difference model (section 5.11.2)

Absolute		GENMOD			MIXED	
difference	McNemar's	Missing	Full	Summary	Full	Summary
0.0%	4.8	26.4	6.7	5.3	5.0	6.1
0.5%	6.6		*	7.2	6.9	7.8
1.0%	9.9	26.0	9.7	11.6	9.9	12.7
1.5%	14.9		*	16.5	15.0	16.3
2.0%	19.2	36.2	19.8	21.8	19.7	23.6
2.5%	32.1		*	34.4	32.5	35.2
3.0%	45.7	38.0	46.0	48.1	46.6	49.1
3.5%	57.3		*	59.8	58.3	61.9
4.0%	70.0	51.6	70.0	71.8	70.6	71.7
4.5%	81.7	43.1	82.3	82.7	81.9	82.8
5.0%	92.0	54.4	92.1	92.4	92.0	92.2
5.5%	95.0	53.8	95.7	95.7	95.3	95.0
6.0%	98.5	56.8	99.54	98.5	98.6	98.5

Table 5.7: Results from the random unprompted reader and fixed difference (figure 5.4) - percentage of solutions that attained significance at the 5% level

5.11.3 Changing the responses of the readers

Here, not only are readers allowed to vary at the hyperparameter level, they are also allowed to vary in their accuracy when faced with prompting. As before, TPu varies Normally about 90% (sd=2.5%), with the difference between prompted and unprompted condition allowed to vary $N \sim (x, 0.03^2)$, where $x = 0$ to 6%. This difference is then added to TPu to calculate TPp. And so, although the average difference between prompted and unprompted readers may increase, an individual reader may have a poorer performance when prompted than when not prompted. As mentioned earlier, interactions have been included in the analysis using the mixed model; but the GENMOD procedure has been abandoned, due to its problems with uniform effects categories.

This simulation model is closer to a real situation than the previous two designs. In the previous chapter (chapter 4), one radiologist did indeed perform worse with the system than when unprompted.

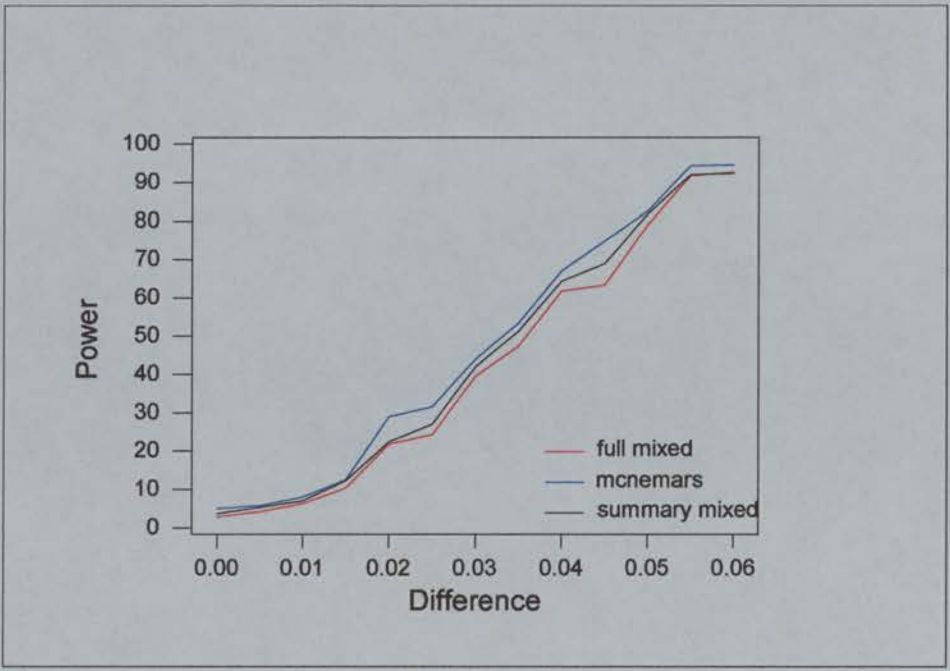


Figure 5.5: Power curve for the random unprompted reader and random difference model (section 5.11.3)

Figure 5.5 illustrates the results of this simulation, with the detail of the analysis in table 5.8. As can be seen from this figure, the power curve has become flatter, rising less steeply at the point at which we are interested - 80% power. Here, 80% power is

Absolute		MIXED	
difference	McNemar's	Full	Summary
0.000	4.9	2.8	3.6
0.005	5.7	4.1	5.2
0.010	7.9	6.3	6.9
0.015	12.4	10.3	12.2
0.020	28.9	21.8	22.4
0.025	31.5	24.2	26.9
0.030	44.0	39.6	42.0
0.035	53.3	47.4	51.2
0.040	67.0	61.8	64.4
0.045	74.9	63.3	68.9
0.050	82.6	79.0	81.8
0.055	94.4	91.8	92.1
0.060	94.6	92.7	92.4

Table 5.8: Results from the random unprompted reader and random difference (figure 5.5) - percentage of solutions that attained significance at the 5% level

passed at a difference of approximately 5%, as the addition of the random difference in the simulation model and the interaction term in the analysis model increases the standard errors. Better power has been achieved by the McNemar's test than for the mixed models analysis for the first time.

5.12 Correlation between readers

All the earlier models have assumed heterogeneity of cancers, where the readers, when faced with a cancer, have the same chance of detecting it whether their counterpart has detected it or not. No allowance has been made for the fact that 'easy' cancers are more likely to be detected by both readers than 'difficult' cancers.

In order to simulate this effect, the following code was entered into the simulation program:

Into the main declaration:

```
double hard, diff;
```

Following the declaration of the value of `hit`:

```
hard=0.3;
```

```
diff=0.05;
```

In the generation of individual 'cancers', just after a subject has been rated a cancer:

```
if (myrandom () <= hard)
  {
    rad_TP=rad_TP-diff;
    PROMAM_TP=PROMAM_TP-diff;
   }
```

The first two sections of code define and set the variables **hard** and **diff** to be the probability that a cancer is difficult to detect (30%) and the difference this makes to the detection ability of the readers (5%) respectively. In this simulation, it will be a simple fixed fall in the detection probability (third section of code, above), but could just as easily be a relative fall or other, perhaps more complex, relationship. As before, the reader/prompting interaction was set at the hyperparameter level.

Due to computer time constraints, fewer points on the power curve have been simulated (see figure 5.6), and only McNemar's test and mixed models analyses have been used to analyse the results. PROC GENMOD has not been applied because of its inability to deal with uniform effect categories in the previous simulations.

Absolute		Full		Summary	
difference	McNemar's	Additive	Interaction	Additive	Interaction
0.0%	4.2	4.2	2.7	5.5	3.4
1.0%	8.2	6.6	7.4	7.6	6.0
2.0%	27.6	21.0	21.2	30.2	22.6
3.0%	43.8	38.2	41.0	44.2	41.0
4.0%	67.1	66.4	62.4	68.7	65.6
5.0%	84.0	93.8	81.5	85.5	83.3
6.0%	95.4	95.3	94.7	96.0	95.3

Table 5.9: Results from the correlation simulation model (figure 5.6) - percentage of solutions that attained significance at the 5% level

In order to examine whether the inclusion of the interaction term has any influence on the power in this simulation, both the additive and interaction analysis models have been used to analyse the data. From table 5.9 above, it is clear that the inclusion of an interaction term decreases the power of the data. This is due to the inflation of the

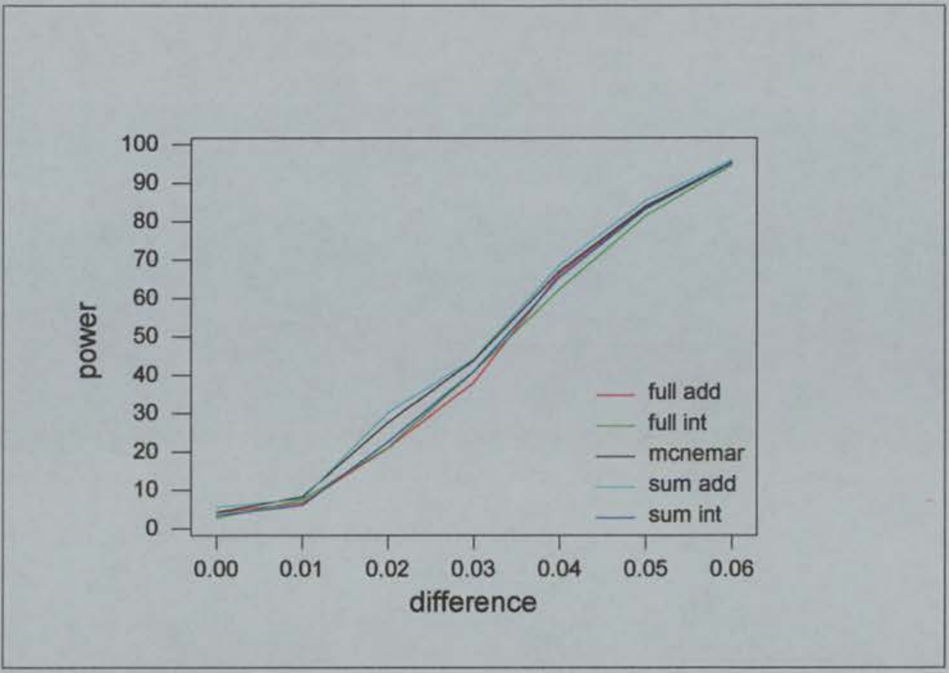


Figure 5.6: Power curve for the random unprompted reader and random difference model with correlation between decisions on cases (section 5.12)

standard error of the treatment effect (the prompting system, in this case). Treatment standard errors are calculated from the reader.prompting variation and, unlike the additive model, will be more generalisable to a wider population of readers than only those sampled.

5.13 Conclusions

Under normal clinical conditions, sets of films can be said to have three states; cancer, suspicious feature and normal. This simulation only dealt with the binary outcome (cancer/not cancer), but the third state, suspicious feature, would be likely to influence real data, as readers are more likely to recall the same suspicious cases, leading to greater correlation between the readers' responses to the recalled non-cancers than was evident in this model.

PROC GENMOD's limitations in its ability to deal with uniform effect categories has been clearly demonstrated in this chapter. Since the principles behind the simulations are based on real data, that radiologists can and do detect all the cancers in a batch some of the time, PROC GENMOD has shown itself to be entirely unsuitable for use in a real trial of a system such as PROMAM.

Mixed models take into account the potential correlations within the data, whereas conventional analyses treat all observations as independent. It also has the advantage of being more robust to missing data, a simulation that would have been performed had time allowed. For a set of real experimental data, a random effects model should be fitted, with fixed treatment effect (prompted or unprompted) and random reader and reader.prompting effects. There is, of course, a penalty in terms of power, as the standard errors are increased due to the inclusion of the reader.prompting interaction, but, in my view, this is a price that should be paid for the use of a more appropriate model, and the wider generalisability of the findings.

Chapter 6

Conclusions

6.1 Introduction

Early work illustrated that there is much scope for improvement in the cancer detection rate in breast screening, without a corresponding loss of specificity. Recent events in the Beatson Oncology Centre, Glasgow, have shown that the cancer service is massively over-stretched and under-resourced. With the increasing shortage of radiologists and radiographers, and the rising demand placed on the Breast Screening Programme, it is not difficult to imagine that technology is likely to play a greater role in the detection of breast cancer in the coming years. A system such as PROMAM could lower the load of the radiologists by potentially taking the place of the first reader.

PROMAM was, and is, just one of many applications of computer assistance in medical applications. With the rapid development of algorithms and processing power, the field is ever widening, as more applications are imagined and conceived.

6.2 Experimental results

The most surprising result to emerge from the "subjective reaction to prompting" experiment (Chapter 3) was the tolerance that radiologists had for the prompts. Conventional wisdom had decreed that high rates of prompts would be unacceptable, distracting and ultimately ignored. Our findings, however, showed a high tolerance for the false prompts that took the team completely by surprise. Upon questioning the radiologists involved, it was discovered that, provided a false prompt highlighted a feature that was real but benign, they could be rationalised, and then ignored. It was prompts that could not be rationalised that would cause problems, which may have been the cause of the low acceptable TP:FP ratio as seen in the Manchester experiments. Indeed, it was found that if the prompting rate was too low, it would have exact opposite effect, by being unable to sustain the radiologists' interest, thereby causing them to ignore the prompting system altogether.

A major concern of the radiologists in screening centres had been that the prompting system might cause the recall rate to increase. Recalls in the "subjective" experiment were shown to be unrelated to increasing prompting levels. The one significant result was only of concern when the hazards of multiple testing were ignored. Even then, the highest recall rate was at an intermediate prompting level (i.e. *medium*), suggesting more strongly that this might well be due to chance. Even if this effect was not due to random chance, it would still not support the argument of increasing prompting levels producing a correspondingly elevated recall rate. Therefore a higher prompting rate would still be acceptable in a clinical environment. There was an increase in the time taken to read a set of films as the prompting rate increased, but this was only of the order of 26% and would be expected to decrease as radiologists became accustomed to the system.

At the conclusion of this experiment, there were two major improvements to the system suggested by the radiologists:

- the ill-defined lesions algorithm needed to be improved
- the microcalcification algorithm should be capable of dismissing vascular calcification

It was agreed that, with these two improvements, the system would be acceptable in a clinical setting.

The next step in the process, the pre-clinical trial, proved to be a success. Although there were technical problems with the scanner, the software proved capable of dealing with the volume of films, while still generating acceptable TP and FP rates. Despite a slightly lower overall cancer detection rate by the prompted readers compared to the unprompted readers, four of the five radiologists had higher detection rates when prompted.

The only radiologist who had lower detection rates when prompted was in fact found to have dismissed four correctly prompted cancers of the seven s/he missed. This, combined with the fact that that when unprompted they managed to achieve a 100% detection rate, makes it very difficult to draw any meaningful conclusions in this instance. However, of the unprompted cases that were seen by this radiologist, all 26 were also detected by the prompted paired reader, while of the 36 that s/he saw prompted, only 30 were detected by the unprompted readers. Had s/he recalled the four cases that were correctly prompted, this would have given a 91.7% detection rate over 83.3% for the unprompted readers.

The fall in the recall rates for the prompted radiologists, although a surprise, was pleasing, as it had been a concern of radiologists that the prompting system would elevate the recall numbers above that with which they were able to cope. Time taken to complete a session had also been a concern, and, although it remained higher than the unprompted sessions, it was falling as the radiologists became more accustomed to the system. As radiologists become more familiar with the system, it is possible that more efficient ways of utilising the prompts may come into practice, bringing the difference in reading times between prompted and unprompted reading even lower.

6.3 Design of an evaluation experiment

Any worthwhile experiment to evaluate such a system as PROMAM would need to accept several design aspects in order to produce evaluable results. Mammograms would need to be multi-read; at the very least each mammogram in the experiment should be evaluated by one prompted and one unprompted reader; more if possible, although this is unlikely to happen in the UK Breast Screening Programme. Women must be recalled on a 'worst case' basis - i.e. if either the prompted or unprompted reader believes a case contains a suspicious feature, then that woman must be recalled for further examination. Anything less would make the comparison between prompted and unprompted readers impossible to quantify. Readers should also be balanced between prompted and not prompted, as well as reading order, to minimise potential biases.

With plausible assumptions of the degree of correlation between readers and a realistic effect size, such an experiment would require on the order of 100,000 women.

6.4 Simulating the clinical trial

The main conclusion from this work is that PROC GENMOD is ill-equipped to deal with the type of data we would have been dealing with had the PROMAM system gone to trial. Although it is unlikely that radiologists are 100% accurate, it may be possible that they are 'detectably 100%' either when prompted or unprompted - in other words of the ones that are detected by the pairings of which they were a part, they may have detected all of them. Other cancers may have been missed by both. This will lead to a uniform category effect, and eventually to the failure of PROC GENMOD to converge.

As would be expected, McNemar's test proved insensitive to underlying covariance structures, being unable to take into account the effects of the various factors at work in the model. As an initial examination of the data it is useful, as it will give an estimate of the difference, but it should always be followed by a more detailed analysis of the data.

The Mixed Models procedure, on the other hand, performed well under these data. Since the estimates of standard errors are drawn from all strata, the instability that PROC GENMOD has shown with uniform effect categories is not repeated with this analysis method.

Interaction terms should be included in any analysis of data of this type, despite the corresponding loss of power, as this will give results that are more generalisable. Mixed models are also considerably more robust to missing data, a situation that would easily arise in an over-stretched screening unit, under conditions that may require that some mammograms be single read.

6.5 The future of PROMAM

In this thesis, I have detailed the work done on the PROMAM project. This was a collaboration, the culmination of the hard work of many, some of whom have been mentioned. Had circumstances favoured the project, much more work could have been done from the basis of the digitisation of mammograms.

Currently, direct digitisation of mammographic images is an important area of development, paralleling the emergence of digital photography. It is anticipated that within 10–15 years, the use of film will have been completely superseded. PROMAM's format was ideally placed to take advantage of this methodological leap forward, as well as being able to provide a direct link to the proposed National Data Collection System.

Another potential use for the imaging system was as a proposed training tool, using the vast database of annotated images to teach radiologists and radiographers to read mammograms, enabling them to compare their decisions with the recorded decision of a senior radiologist. This had been greeted with enthusiasm by the radiologists at the South East Scotland Breast Screening Centre in Edinburgh, and has recently been funded to develop further.

6.6 Beyond PROMAM

In the years since the dissolution of the PROMAM team, and the cessation of progress on the algorithms, there has been an explosion of algorithms designed to detect a multitude of cancerous signs. Few, however, make it as far as a fully-integrated detection system, capable of handling the sheer volume and variety of a clinical set-up. The leaders in this field are R2 Technology's ImageChecker M1000 System, the CADx Second Look system, and to a lesser extent, the Fuji Computed Radiography system.

1. **ImageChecker M1000** - as with the PROMAM system, this system works by digitising the mammogram and analysing the resulting digital image for clusters of bright spots (microcalcification) and dense regions (masses). Additionally, it is also claimed to detect dense regions with radiating lines, which the PROMAM team did not achieve. A low resolution image on a monitor is produced to direct the radiologist's attention to the region of interest.

   R2, the proprietary company, offers a complete package for their system, which incorporates a single or continuous case loader for the scanner, an ethernet connection, and four types of mammogram display unit. From this, it may be inferred that R2 believe that this field has the potential to be extremely lucrative.

   In April 2002, R2 announced that the USA Food and Drug Administration (FDA) had granted clearance for use of the ImageChecker with full field digital mammography - the direct digitisation of the breast image, by-passing the need for film. Clearance for use with film-based screening had been granted in 1998. By April 2002, more than 300 ImageChecker systems had been installed worldwide.

   In 1998, a study of 300 patients with 1100 mammograms [98] showed that the R2 system had an detection rate of 82.1% (32 accurately detected from 39 histologically proven cancers), with 1797 prompts, 94.3% of these false positives (1695). Hence, although the detection rate is good, the false positive rate leaves a lot to be desired, being either 1.54 per image or 5.65 per woman.

A further experiment in 2000 [99, 100] of 1083 biopsy proven cancers indicated that the R2 system had a 60% detection rate on the 286 cases where a retrospective review had determined that there were visible signs of cancer on the prior mammograms. When further examined by a blinded review panel, 115 of these were considered to warrant recall. The R2 system had a 77% detection rate on these cancers. Although the authors report no significant increase between the radiologists' recall rates before and after the introduction of the CAD system, there is a difference; 8.3% before and 7.6% after. There is a slight improvement in the generation of false prompts; these have fallen to an average of one false prompt per film.

Since PROMAM, at the time of the preclinical experiment (Chapter 4), had detection rates of 90% for the microcalcification algorithm and 80% for the mass algorithm, with 1 in 4 and 1 in 2 women prompted respectively, the performance of the PROMAM system was considerably better.

Recently, the R2 system has been undergoing UK trials in the Canterbury Assessment Centre, in order to evaluate its usefulness within the UK Breast Screening Programme [101]. A set of 104 interval cancers (104 film pairs; one cancer, one non-cancer) generated 134 prompts on the mammograms containing the cancers, and 109 on the mammograms that did not contain a cancer.

Twenty-nine of these cancers had been classified as false negative or minimal signs on the previous mammograms. The R2 system detected 15 of these 29 (52%) when examining these earlier mammograms.

At least one reader rejected the correct prompt in five cases, which would highlight the earlier conclusion that even the most sensitive system may be over-ruled by the radiologist.

A multi-centre trial of the system, using a variety of category of film reader, is currently assessing the impact of the R2 system in screening [102]. This trial aims to assess the potential of computer aids as a solution to the manpower crisis in the screening programme. Two different reading protocols are being used in this trial;

(a) double reading by radiologists compared to double reading by radiologist and computer-supported non-radiologist, and

(b) double reading by radiologist and radiographer compared to computer-supported radiologist.

The sensitivity and specificity of the categories of reader will be assessed, as will the the economic implications for the screening programme. This trial is not expected to be published until late 2003.

2. **CADx Second Look** - Not as much is known about this system. As with PROMAM and R2, this system is designed to be an aid to a reader, and not a replacement. The mammogram is first read by the radiologist, then the Mammagraph^{TM} (the CADx name for a laser-printed image of the digitised mammogram, with the areas of suspicion highlighted) is examined and any regions of suspicion are re-checked. Second Look have also entered into the market with a fully comprehensive package, although they have chosen to print the 'Mammagraph' on paper, rather than display it on a screen. Approval from the FDA was awarded in February 2002, for both screening and diagnostic use.

   Publicity material released by the company claims that 26.2% of cancers missed by a radiologist would be detected by the Second Look system. In other words, the false negative interval cancers. However, even as long ago as 1998, PROMAM produced a 44% detection rate with FN interval cancers, so this is not as impressive as it may first appear.

An experiment in 2001 [103] demonstrated that the system was 90% (135/150) sensitive with the false positive rate of approximately 1.3 false prompts per image. Sensitivity on masses was 88.7% and on microcalcification was 98.2%. This is an improvement on PROMAM's performance at the time of the preclinical experiment, although the false prompt rate is still higher than that of PROMAM.

CADx are also actively involved in the fields of computer-aided detection in lung and colon cancer, as well as cardiovascular disease.

3. Fuji Systems - Computed Radiography. Although this is not, in itself a detection system, it has been used by several CAD systems as the method of capturing the digital image prior to analysis [66, 104]. Introduced in the early 1980s by Fuji Photo Film Japan, more than 1200 Fuji CR systems have now been installed in clinics in the USA, and 12,000 worldwide, making it the leader in the field of mammographic digital imaging. From their publicity:

   > Computed Radiography (CR) using photostimulable luminescent technology is a digital image acquisition and processing system for static projection radiography.

   The Fuji Computed Radiography System uses a phosphor screen with energy storage capability as an X-ray image receptor. As the radiation falls on this screen, the phosphor is activated in much the same way as the radiographic film is, recording an image.

   After exposure, the cassettes are transferred to a reader system. Here the imaging plate is scanned with a laser beam which stimulates luminescence proportional to the local X-ray exposure. The luminescence signal is converted to an electrical signal and is digitised.

   The data representing the image is subjected to digital signal processing to optimise the diagnostic content of the visualised data. The image can be recorded on laser printed film or transmitted and stored digitally. Experiments have shown [105] that an image adequate for screening purposes can be obtained using CR, even though the x-ray dosage is only 5% of that required for conventional radiography.

Fuji Medical Imaging claim that their computed radiography is the most dependable and fastest CR system in the world, and comparison with other systems would appear to support this [106].

Despite all the claims of accuracy by the various manufacturers, there is still no definitive proof that CAD will improve the sensitivity of the reader with no compromise in specificity. In fact, Brem and Schoonjans [107] believe (for microcalcifications, at least), that there is no significant changes in sensitivity when the mammograms were viewed by experienced radiologists.

Since the introduction of two views into the screening programme improved the cancer detection rates, advances are being made into digitally fusing the two views in order to reduce the number of false prompts generated by most CAD systems [108]. It is hoped that the fusion of information from more than one view will improve the performance by correlating objects using the geographical location, morphological and textural features, then using discriminant analysis to classify the object pairs as a true or false mass.

Recently, a new algorithm [109] was announced which used the information from prior mammograms to classify masses as malignant or benign. Early tests of this algorithm are promising, with the information on the prior mammogram significantly improving the accuracy of classification of the masses. As this information is often what is used by the radiologists in assisting them in their classification of the mass, this is an important step forward. However, in the experiments conducted by the PROMAM team, it was discovered that the radiologists preferred that the system *did not* classify the masses and microcalcifications for them. It will be interesting to follow this algorithm into tests of the system in conjunction with screening radiologists.

6.7 Full field digital mammography

As mentioned earlier, full field digital mammography is expected to replace the need for photographic film in the detection of early breast cancer in the very near future. The developers of the R2 system [110] believe that this process can only be exploited fully by the unique advantages that only CAD can provide. The ability to perform the analysis of the breast image, without the laborious processing of the film and digitisation, would mean that women attending a clinic for screening could receive their results and any further investigations in the one visit. As the delay between attendance and receiving the results can be up to four weeks, this would ease the anxiety of many, and ensure that the few who needed it received further treatment as quickly as possible. Obviously, reducing the time to diagnosis will improve the prognosis of the cancer, saving lives, money and valuable treatment time.

Digital mammography will also eliminate the need to recall women for technical recalls, as the image may be displayed within seconds, rather than the hours required to process the film. Repeats may be done 'on the spot', eliminating the need to recall the woman. Also, other techniques such as magnification mammograms [111] and ultrasound scans may be performed.

Digital images, direct or scanned, may also be attached to a woman's clinic records, making the retrieval of images considerably easier than it often is currently. Telemammography [112, 113], the transmission of high resolution copies of mammograms over some form of digital data connection, is an ideal use of this technology, with the ability to consult an expert remotely and discuss the same image in real time.

Mammography is not the only field of medicine to benefit from computed radiography. As the radiation dose is lower than for conventional x-ray photography, it is ideal for those tissues which are particularly sensitive to irradiation; for example, the brain [114], the spine [115], and the heart [116]. Additionally, a system that delivers a lower dose will allow more images to be taken with the same or lower risk to the patient.

Obviously, a lower dose is beneficial to all tissues, as any exposure to radiation can increase the risk of treatment related morbidity. Studies have shown [117, 118] that the dose delivered during regular breast cancer screening, particularly for younger women, can have a detrimental effect and may lead to the induction of cancer. This risk, although small, is cumulative. Despite this, however, the benefits of screening outweigh the risk, and a reduction of the radiation dose can be only beneficial.

As processing power becomes cheaper and more accessible, digital imaging and analysis will become more feasible. An imaging system is already under consideration for the UK screening programme, although much testing lies ahead of it before it is ready for mass implementation. This is almost certainly how breast screening will be conducted in the future, as the technology is bound to continue to improve.

Breast cancer is a field in which a great deal of scientific research is being conducted. Many journals are dedicated to the topic, a large number of which are mentioned in the bibliography. A current search of the MIMAS/ISI database for "breast cancer" and "detection" and "screening" yielded 145 articles from 2001 alone; from genetic clinics for women with a family history of breast cancer [119] to new theories of breast cancer markers [120], to new advances in pathological detection [121].

6.8 Beyond Breast Cancer

Currently, mammography appears to be the focus of much of the work of computer-aided detection/diagnosis systems. Mammography appears to be particularly suited to this form of analysis, as it is basically the detection of cancer 'signal' within the 'noise' of breast tissue. Microcalcification in particular is well suited to such an approach, with most good algorithms achieving detection rates in excess of 90%.

A search of the MIMAS/ISI (wos.mimas.ac.uk) database with the topic "computer aided detection" yielding 33 results from 2001, 26 of which were related to breast cancer. The remainder involved colonography and the detection of lung nodules.

CAD in medicine is not new, the earliest recorded abstract in the MIMAS database is from July 1986 and deals with the computer-aided detection of lung nodules [122]. The earliest mention of mammograms is in 1988 [123], when the detection of microcalcifications on mammograms is examined. However, one of the most widely used CAD systems in medicine is PAPNET; a computer-assisted Pap cervical smear test.

Detected early, cervical cancer has an almost 100% chance of cure, a claim supported by the 70% reduction in cervical cancer in countries where smear tests are regularly performed. 80% of all new cases each year are in developing countries, where only 5% of the female population have access to a screening programme, compared with 40-50% in developed countries [124].

Traditionally, Pap (Papanicolaou, named for the inventor of the cervical smear test) smear testing relies on the human eye to look for abnormal cells under a microscope. Since a patient with a serious abnormality can have less than 12 abnormal cells in the 30,000 from a typical sample, it is, unsurprisingly, very difficult to detect all cases of early cancer. In fact, false negatives of between 14% and 33% of positive smears have been reported [125].

PAPNET uses neural networks [126], and assists the human 'diagnostician' by identifying the 128 most suspicious cells for rescreening with light microscopy. This, the developers (Neuromedical Systems Inc.) claim, reduces the work by 98%, thereby reducing human fatigue and improving accuracy ten-fold. Claims for an improvement in malignancy detection rate of up to 30% have been made, although some studies have shown little or no difference being made to the sensitivity of skilled cytographers [127, 128].

The most common use for this system is the detection of false negatives from samples that have been classified as normal by manual screening. One such example was documented in an Australian pathology centre, over the course of a year [129]. During this time, 54,658 samples classified as normal were examined by the PAPNET system and resulted in 266 samples being reclassified as abnormal. In this instance, the PAPNET system detected an additional 7% as abnormal.

Although PAPNET is one of the oldest on the market, it is not the only system available in the field of computer-aided detection of pre-cancerous signs of cervical cancer. Both AutoCyte and AutoPap are relative newcomers to the market, claiming improvements in detection without concurrent loss of specificity [130, 131]. AutoPap recently received approval from the FDA for use in screening and non-"high risk" smears. A clinical trial [132] comparing the system with the standard practice of manual screening plus 10% random quality control rescreening showed that a significantly higher number of malignancies were being detected by the AutoPap system.

With the advent of ever more powerful computers, many more applications for computer-aided detection are being discovered. Other forms of X-ray have already been examined, with lung cancer being a major beneficiary [133]. Other modalities are also being explored; for example, ultrasound [134] and computed tomography (CT) scans [135]. Most of these new applications are in their infancy, but many build on the successes demonstrated by the application of the technology in mammography. In essence, any data which forms detectable patterns and may be digitised without significant loss of definition could be subject to the same type of scrutiny which has benefited mammography over the past two decades. For example, pre-processing of the scalp electroencephalogram (EEC) signal from infants at risk of seizures [136] allows a drastic reduction in the number of false alarms.

6.9 In Conclusion

Computer-aided detection is rapidly becoming a widely accepted part of medical research, as the variety of applicable fields broadens. The alternative definition of CAD, computer-aided diagnosis, is also broadening, as attempts are made to categorise detected features into 'benign' or 'malignant'.

As with most types of medical treatment, CAD (in both its forms) has advantages and disadvantages. As time, experience and expertise increase, it is hoped that the problems would begin to be heavily out-weighed by the improvements to detection/diagnosis that the ability to pre-process the data will provide.

6.9.1 The advantages

Computers are, at the very heart, simple creatures. If given the same data, they will return the same answer, time after time, ensuring that the reproducibility of the results is high [137]. Humans are not so obliging, which is why double reading has proven so effective. Algorithms do not suffer from 'blind spots', areas where, for example, some radiologists consistently fail to detect malignancies. They do not suffer from fatigue, which can be a major problem in the cervical screening programme, nor do they become distracted by their environment. Computers are the ideal employees; they do not take holidays, they do not fall ill and they are always available when needed.

By allowing a CAD system to take the place of the 'first reader', it has the potential to liberate a highly trained member of staff for other duties. This would clearly produce huge benefits for the overstretched and undermanned UK Breast Screening Programme.

As progress on the training of non-radiologists to read mammograms increases, a prompting system to assist them in their decision-making will be of immense use, both clinically and psychologically. Clearly, any decision made by an inexperienced radiographer that is backed by the prompting system will have a positive effect on their confidence in their own ability. This would then, hopefully, be reflected in a higher specificity, which is the greatest obstacle to wide-spread use of radiographer reading.

It is to be hoped that this will be demonstrated in the results of the UK R2 trial [102], as this will examine the sensitivity and specificity of radiologists and non-radiologists, supported by the prompting system. Unfortunately, the results from this trial will not be available until late 2003.

With the advent of direct digitisation, the ability for breast cancer screening to fully comply with the government's stated aim for 'one-stop clinics' could finally be realised. With the capability for images to be processed and analysed within minutes of the woman undergoing a mammogram, technical failures could be redone immediately, and further investigation or treatment could be initiated with considerably less delay. This would obviously reduce anxiety on the part of the patient, expedite treatment, which would improve prognosis, save lives and money.

The potential exists for CAD to dramatically improve the detection in several fields, particularly screening, where the low level of malignancies in amongst many normals makes 100% detection difficult.

6.9.2 The disadvantages

As with most things in life, computer-aided detection does have its disadvantages. As mentioned before, computers are very simple creatures. They cannot, to any great extent, learn from their mistakes, which will lead to the same errors being repeated time after time. Although these errors will eventually be automatically dismissed by readers accustomed to the system, it may be a long learning process, which each reader will have to suffer.

Computers are notoriously unreliable, and any system that depends on constant availability is liable to encounter problems. All hardware eventually develops problems, and will need to be replaced and repaired. With the ideal of the telemammography being the rapid movement of digital information, there is always the risk of external links being hacked, and sensitive information being accessed.

Any increase in sensitivity is nearly always paid for by a decrease in specificity and vice versa. In most screening situations, this is always the trade off that is made between enough recalls that cancers are not missed, yet not so many that the screening system is swamped. This was clearly illustrated in the PROMAM pre-clinical experiment discussed in Chapter 4, as the recall rate rose linearly with the proportion of detected cancers.

A trial of the R2 system in a community breast centre, interpreting 12,860 screening mammograms over a 12 month period [138], observed an increase in the recall rate from 6.5% to 7.7%. This was, however, balanced by an increase of 19.5% in cancer detection, and a higher proportion of early-stage malignancies (73% to 78%).

Another possible consequence of excessive false prompts is the temptation on the part of the reader to ignore the information provided. It is an entirely human reaction, as time is wasted on eliminating prompts that could have been more profitably spent. It may also prove difficult to demonstrate to readers that the prompting system has improved their particular sensitivity, as they may not believe that they would have missed the feature which was prompted.

While prompting remains in its infancy, the costs can be prohibitively expensive, as much new equipment must be purchased in order to scan, analyse and display the results. For example, the cost of detecting a false negative cancer using PAPNET was calculated to be $25,748 by Troni et al [128]. Naturally, should a system be adopted over the entire NHS, for example, the individual costs should reduce.

The introduction of any new system into a well established set-up can be disconcerting and disruptive. Changes, no matter how benign or well-intentioned, will often alter the way films are read, which will in turn, alter sensitivity, specificity or possibly both. Periods of transition are almost inevitable as readers become accustomed to the changes in practice and they develop an understanding of the strengths and weaknesses of the system. It would be hoped that this transition period is as short and as smooth as possible, and that a prompting would ultimately improve sensitivity and/or specificity.

No system will ever be 100% sensitive or specific, mainly because readers are not. If it is not possible for, say, a radiologist to detect all cancers, how can we expect a computer to detect all the subtleties that are signs of pre- or early cancer?

6.10 In Summary

This thesis has attempted to illustrate the detail of experiments and testing that comprised the development period of the PROMAM project. This included designing experiments that accommodated the problems of testing 'live' in a clinical setting, while ensuring that the data were unbiased and as well-balanced as possible. This allowed for efficient analysis, which gave us results that were generalisable to a wider population of radiologists.

The work done here has also shown that, despite all initial indications to the contrary, radiologists are far more tolerant of prompts than they were previously given credit for. We have demonstrated that radiologists are willing to accept a high level of false prompts, provided the prompts are 'sensible' and are easily rationalised.

Despite the disappointment felt by the entire team at the failure of the multi-centre trial, some of the variety of potential methods of analysis have been examined in detail. This highlighted clearly the shortcomings of the PROC GENMOD procedure, and explored the advantages of the mixed models method of analysis.

Work of this nature will, invariably, be a collaborative effort, but I hope it has been shown that much better use of resources can be achieved when there is integrated statistical support for a project.

Appendix A

Glossary

Annotation when a cancer is delineated on a high-resolution image by a radiologist

Assessment when a woman is recalled to the clinic due to a suspicious feature on her mammogram

Batch a set of cases that would all be seen by the same two radiologists

Cancer a pathology proven malignancy

Case the films from a particular lady. May be malignant or non-malignant

CHI number the personal identification for each woman in the Scottish Breast Screening system

Cranio-Caudal (CCs) the horizontal view films, with compression of the breast from top and bottom. Are routinely given to prevalent attendees

False Positive (FP) a recall made that was not a pathology proven malignancy

False Positive Rate (FP rate) the proportion of non-cancers that were recalled i.e. number of FP/number of women screened

Film a mammogram

Film Bag a cardboard folder containing the films and records of a woman

Incident the name given to a screening other than the first

Ill-defined lesion An area of increased density with a fuzzy edge on a mammogram. An indication of a cancer, especially in conjunction with microcalcification

Mass Short-hand for an ill-defined lesion

Mediolateral obliques the vertical view films, with compression of the breast from the sides, diagonally from the shoulder to the stomach. Often just called obliques

Microcalcification Small flecks of calcium within the breast tissue. Clusters are highly indicative of a cancer or pre-cancer, especially in conjunction with an ill-defined lesion

Minimisation a method of assigning radiologists to batches method designed to balance the combinations of radiologist and prompting

Nidus the central mass within a spiculated, tentacle or stellate lesion

Prevalent the name given to a woman's first time of screening

Previous films the films taken at a previous screening round. Are used to look for any changes that have occurred over time

Prompted the presence of a piece of paper containing a low-resolution image of the films, with suspicious areas highlighted

Recall rate the proportion of women recalled for further assessment at screening i.e. number of women recalled/number of women screened. Will not include tech recalls

Reporting form the form used at screening to record patient data and indicate whether the woman should be recalled

Screening the action of routinely calling a woman in for a mammogram once every three years

Screening conditions the usual working practices employed in a screening centre when reading normal screening throughput

Sensitivity a measure of how good a method of detection is at detecting the condition (number of cancers detected/number of cancers available to be detected)

Specificity a measure of how good the method is at excluding those without the condition (number of women declared free of disease and free of disease/number of women free of disease)

Tech recall a woman recalled for technical reasons, rather than for a suspicious region on a mammogram

True Positive (TP) a correctly identified cancer

True Positive Rate (TP rate) the proportion of cancers correctly identified i.e. number of TP/number of cancers

Two-view where both obliques and CCs are taken of the mammography breast

Appendix B

Proof for the hypothesis that $\mathrm{M(D)} \leq \mathrm{M(I)}$

Assumptions for comparison of blinded and non-blinded reading

1. The cancer detection rate of R_1 is fixed in both cases (independence and dependence), since R_1 is independent of R_2
2. The cancers discovered by R_2 that were missed by R_1 ($\mathrm{R}_1-\mathrm{R}_2+$) remain fixed, as these are not influenced by prior knowledge of R_1.

We propose that an unblinded reader, R_2, may report cases when prompted by R_1 that would not have been discovered had the readings been independent. Hence, some of the cancers discovered by both R_2 and R_1 would, in actual fact, only have been reported by R_1, had the reading been blinded. If the independent, blinded case is the standard by which we compare, then the table for the comparison to the non-blinded case is as follows:

	R_2+	R_2-	
R_1+	a+pb	(1-p)b	a+b
R_1-	c	d	

where p is the proportion of cases from cell $\mathrm{R}_1+\mathrm{R}_2-$ that would have been reported had the session been non-blinded. Hence, the mean second screener contribution from a non-blinded second screener [M(D)] is given by:

$$\frac{c+(1-p)b}{2a+(1+p)b+c}$$

Proof that M(I)$\geq$ M(D)

We know that: a, b, $c \geq 0$, with $0 \leq p \leq 1$. Let M(I) < M(D)

$$\begin{aligned}
\frac{b+c}{2a+b+c} &< \frac{c+(1-p)b}{2a+(1+p)b+c} \\
(b+c)(2a+(1+p)b+c) &< (2a+b+c)(c+(1-p)b) \\
2ab+(1+p)b^2+bc &< 2ac+2ab(1-p)+bc \\
+2ac+(1+p)bc+c^2 & \quad +(1-p)b^2+c^2+(1-p)bc \\
2ab+b^2+pb^2+bc+pbc &< 2ab-2abp+b^2-pb^2+bc-pbc \\
pb^2+pbc &< -2abp-pb^2-pbc \\
pb+pc &< -2ap-pb-pc \qquad \text{if p=0 then 0<0} \\
2ap+2bp+2cp &< 0 \\
2(a+b+c) &< 0
\end{aligned}$$

which is not true, since a, b, and c are all ≥ 0

Hence, M(I) $\not<$ M(D). Therefore M(I)$\geq$ M(D).

Appendix C

Calculation of sample size

Cancer detection rate = 0.5%

	Agreement									
Increase	0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94	0.96	0.98
−5%	221192	195912	170632	145352	120072	94792	69512	44232	.	.
−4%	355492	315992	276492	236992	197492	157992	118492	78992	39492	.
−3%	649548	579325	509103	438881	368659	298437	228214	157992	87770	.
−2%	1500992	1342992	1184992	1026992	868992	710992	552992	394992	236992	78992
−1%	6161992	5529992	4897992	4265992	3633992	3001992	2369992	1737992	1105992	473992
+1%	6477992	5845992	5213992	4581992	3949992	3317992	2685992	2053992	1421992	789992
+2%	1658992	1500992	1342992	1184992	1026992	868992	710992	552992	394992	236992
+3%	754881	684659	614437	544214	473992	403770	333548	263325	193103	122881
+4%	434492	394992	355492	315992	276492	236992	197492	157992	118492	78992
+5%	284392	259112	233832	208552	183272	157992	132712	107432	82152	56872
+6%	201881	184325	166770	149214	131659	114103	96548	78992	61437	43881
+7%	151543	138645	125747	112849	99951	87053	74155	61257	48359	35461
+8%	118492	108617	98742	88867	78992	69117	59242	49367	39492	29617
+9%	95572	87770	79967	72165	64362	56560	48758	40955	33153	25350
+10%	78992	72672	66352	60032	53712	47392	41072	34752	28432	22112

Table C.1: Sample size calculations when cancer detection rate is 0.5%

Cancer detection rate = 0.6%

	Agreement									
	0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94	0.96	0.98
Increase										
−5%	184325	163259	142192	121125	100059	78992	57925	36859	.	.
−4%	296242	263325	230409	197492	164575	131659	98742	65825	32909	.
−3%	541288	482770	424251	365733	307214	248696	190177	131659	73140	.
−2%	1250825	1119159	987492	855825	724159	592492	460825	329159	197492	65825
−1%	5134992	4608325	4081659	3554992	3028325	2501659	1974992	1448325	921659	394992
+1%	5398325	4871659	4344992	3818325	3291659	2764992	2238325	1711659	1184992	658325
+2%	1382492	1250825	1119159	987492	855825	724159	592492	460825	329159	197492
+3%	629066	570548	512029	453511	394992	336474	277955	219437	160918	102400
+4%	362075	329159	296242	263325	230409	197492	164575	131659	98742	65825
+5%	236992	215925	194859	173792	152725	131659	110592	89525	68459	47392
+6%	168233	153603	138974	124344	109714	95085	80455	65825	51196	36566
+7%	126285	115536	104788	94040	83291	72543	61795	51047	40298	29550
+8%	98742	90513	82284	74055	65825	57596	49367	41138	32909	24680
+9%	79642	73140	66638	60136	53634	47132	40630	34128	27626	21124
+10%	65825	60559	55292	50025	44759	39492	34225	28959	23692	18425

Table C.2: Sample size calculations when cancer detection rate is 0.6%

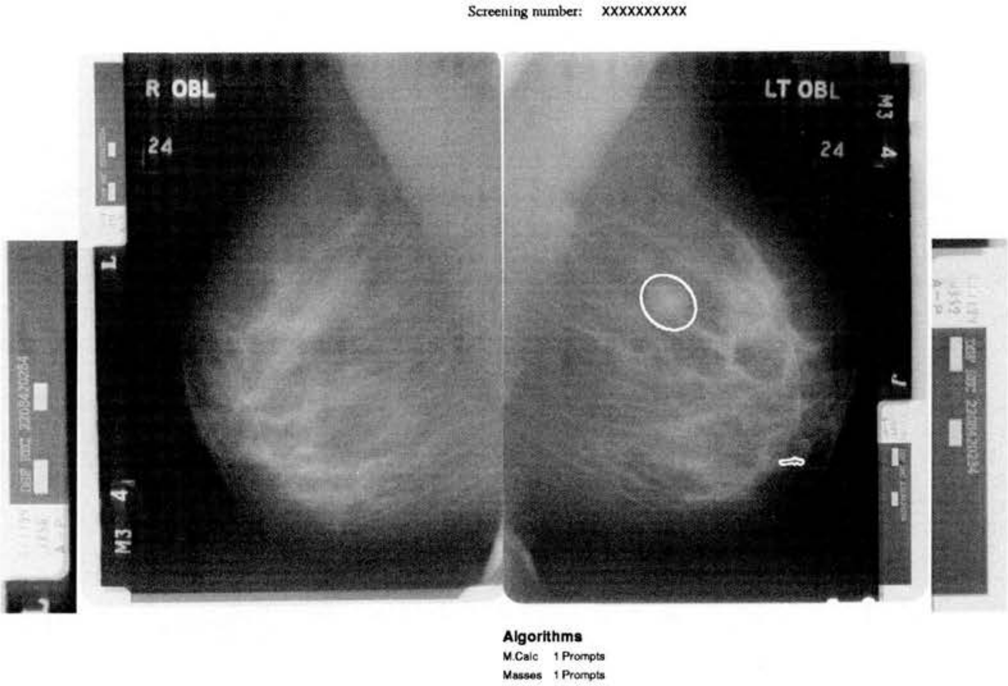
Cancer detection rate = 0.7%

	Agreement									
	0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94	0.96	0.98
Increase										
−5%	157992	139935	121878	103821	85764	67706	49649	31592	.	.
−4%	253921	225706	197492	169278	141064	112849	84635	56421	28206	.
−3%	463960	413802	363643	313484	263325	213167	163008	112849	62691	.
−2%	1072135	959278	846421	733564	620706	507849	394992	282135	169278	56421
−1%	4401421	3949992	3498564	3047135	2595706	2144278	1692849	1241421	789992	338564
+1%	4627135	4175706	3724278	3272849	2821421	2369992	1918564	1467135	1015706	564278
+2%	1184992	1072135	959278	846421	733564	620706	507849	394992	282135	169278
+3%	539198	489040	438881	388722	338564	288405	238246	188087	137929	87770
+4%	310349	282135	253921	225706	197492	169278	141064	112849	84635	56421
+5%	203135	185078	167021	148964	130906	112849	94792	76735	58678	40621
+6%	144198	131659	119119	106579	94040	81500	68960	56421	43881	31341
+7%	108243	99030	89817	80604	71392	62179	52966	43753	34540	25327
+8%	84635	77581	70528	63474	56421	49367	42314	35260	28206	21153
+9%	68264	62691	57117	51544	45971	40398	34825	29251	23678	18105
+10%	56421	51906	47392	42878	38364	33849	29335	24821	20306	15792

Table C.3: Sample size calculations when cancer detection rate is 0.7%

Appendix D

An example of a prompt sheet



Screening number: XXXXXXXXXX

Algorithms

M.Calc 1 Prompts

Masses 1 Prompts

Figure D.1: Produced by the PROMAM team, with thanks to Mark Hartswood

Appendix E

Instructions for radiologists in the subjective reaction to prompting experiment

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 microcalcification clusters, and masses. Prompts for potential masses will always appear as circles or ellipses, prompts for microcalcifications 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.

Unprompted 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.

Appendix F

SAS code

F.1 Subjective reaction to prompting experiment

The following variables are defined as:

rad	=	radiologist identifier
set	=	set identifier
session	=	session identifier (first session, second session etc)
cond	=	condition identifier (null, low, medium, high)
recall	=	number of recalls made in each radiologist*session combination
rep	=	whether a particular woman during a particular session was 'recalled'
chi	=	the CHI number that uniquely identifiers each woman
times	=	the time taken to complete a reading session in seconds
prompt	=	cond recoded into 0, 1, 2, 3

F.1.1 Recalls made

F.1.1.1 The SAS code for the Wald Type 3 results, n=111

```
proc genmod;
class rad set session cond;
model recall/n=rad cond set session / dist=b type3 wald;
```

F.1.1.2 The SAS code for the repeated measures results, n=1

```
proc genmod;
class chi set session rad cond;
model rep/n=rad cond set session / dist=b type3 wald;
repeated sub=chi / type=CS corrw;
```

F.1.1.3 The SAS code for the case where *prompt* is a covariate

```
if cond=‘null’ then prompt=0;
if cond=‘low’ then prompt=1;
if cond=‘medium’ then prompt=2;
if cond=‘high’ then prompt=3;

proc genmod;
class rad set session;
model recall/n=prompt rad set session / dist=b type3 wald;
```

F.1.2 Time taken to complete both groups

F.1.2.1 Generalised Linear Model

```
proc glm;
class rad set session cond;
model times=cond rad set session;
lsmeans cond / pdiff cl;
      OR
lsmeans cond / adjust=bon pdiff cl;
```

F.1.2.2 Generalised Linear Model, with *prompt* as a co-variate

```
if cond=‘null’ then prompt=0;
if cond=‘low’ then prompt=1;
if cond=‘medium’ then prompt=2;
if cond=‘high’ then prompt=3;

proc glm;
class rad set session;
model times=prompt rad set session;
```

F.1.3 Questionnaire responses

F.1.3.1 Likert scores

```
proc glm;
class rad cond session;
model likert=rad cond session / dist=b type3 wald;
```

Appendix G

The questionnaires used in the subjective reaction experiment

The following questionnaires (Appendix G and Appendix H) are copyright Mark Hartswood and are reproduced by kind permission

G.1 Pre-experiment questionnaire

Q1 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. □

Q2 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)						
...		□	□	□	□	□
...		□	□	□	□	□

Q3 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					
Q4 The presence of a prompt will make you more inclined to recommend recall	□	□	□	□	□
Q5 The absence of a prompt makes you less likely to recommend recall	□	□	□	□	□

Q6 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						Least useful
		1	2	3	4	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 microcalcification 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

G.2 Post-experiment questionnaire

Questions 1 to 6 are duplicates of the pre-experiment questions (see pages 201 - 205)

Q7 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)

	High	Medium	Low	No prompts
Mass Prompt Rate	□	□	□	□
Calcification prompt rate	□	□	□	□
Sensitivity	□	□	□	□

Q8 What is the highest FP rate you would be willing to accept? (Tick one box only for each)

	High	Medium	Low	No prompts
Mass Prompt Rate	□	□	□	□
Calcification prompt rate	□	□	□	□
Sensitivity	□	□	□	□

Q9 What is the lowest sensitivity you would find useful in a screening context? (Tick one box only for each)

	High	Medium	Low	No prompts
Mass Prompt Rate	□	□	□	□
Calcification prompt rate	□	□	□	□
Sensitivity	□	□	□	□

If you have any further comments with respect to any aspect of the experiment, please write them below:

Thank-you for completing this questionnaire

G.3 Post-session 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.

Q10 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 be 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.	□	□	□	□	□

	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.	□	□	□	□	□

Q11 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) □

Q12 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 microcalcification detection component of this system would be useful to you as it currently stands?	□	□

Q13 Please rate the system components: (Tick one box for each)

	Too sensitive	Just right	Not sensitive enough
Overall	□	□	□
Masses	□	□	□
Microcalcification	□	□	□

Q14 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 microcalcification clusters being detected) Please tick one box per sensitivity setting.

	Very Useful	Useful	Doubtful	Of no use
95%	□	□	□	□
90%	□	□	□	□
85%	□	□	□	□
80%	□	□	□	□

Q15 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

Appendix H

The questionnaires used in the pre-clinical experiment

Due to the repetitive nature of the questionnaires, where a question has been repeated from the questionnaire from Chapter 3, a note will be made indicating the location and number of the relevant question.

H.1 Pre-experiment questionnaire

Q1 as Q2 of the previous questionnaire (page 202) Q2 as Q3 of the previous questionnaire (page 203)

Q3 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 microcalcification clusters	□	□	□	□	□
Detection of ill defined lesions	□	□	□	□	□
Detection of architectural distortions	□	□	□	□	□
Detection of asymmetries	□	□	□	□	□
Classification of microcalcifications	□	□	□	□	□
Classification of ill defined lesions	□	□	□	□	□
Others					
...	□	□	□	□	□
...	□	□	□	□	□

(Please tick one box per statement. Feel free to add additional features, and rate them accordingly)

Q4 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 microcalcification clusters	□	□	□	□
Prompting for ill defined lesions	□	□	□	□
Prompting for architectural distortions	□	□	□	□
Prompting for asymmetries	□	□	□	□
Classification of prompted microcalcifications 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)

Q5 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 microcalcifications □

Prompting for ill defined lesions □

Prompting for architectural distortions □

Prompting for asymmetries □

Classification of prompted microcalcifications as benign to malignant. □

Classification of prompted masses as benign or malignant □

Other

... □

... □

(Feel free to add additional properties, and number them accordingly)

Q6 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)

Q7 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:

Q8a as Q4 of the previous questionnaire (page 204)

Q8b as Q5 of the previous questionnaire (page 204)

Q9 as Q6 of the previous questionnaire (page 205)

Thank you for completing this questionnaire

H.2 Post-experiment questionnaire

Questions 1 to 9 are duplicates of the pre-experiment questions (see pages 211 - 216)

Q10 At the outset of this experiment we gave you an estimate 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?

Microcalcifications .....%

Ill-defined lesions .....%

Overall (sensitivity for detecting all feature types) .....%

Q11 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)		□	□	□	□	□	

Q12 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)	□	□	□

Q13 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

H.3 Post-session questionnaire

Q1 as Q10 in the previous questionnaire (see page 207)

Q2 as Q11 in the previous questionnaire (see page 209)

Q3 as Q12 in the previous questionnaire (see page 209)

Q4 Please rate the system's sensitivity: (Tick one box for each)

	Too sensitive	Just right	Not sensitive enough
Overall	□	□	□
Masses	□	□	□
Microcalcification	□	□	□

Q5 Please rate the system's specificity: (Tick one box for each)

	Too specific	Just right	Not specific enough
Overall	□	□	□
Masses	□	□	□
Microcalcification	□	□	□

Q6 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:

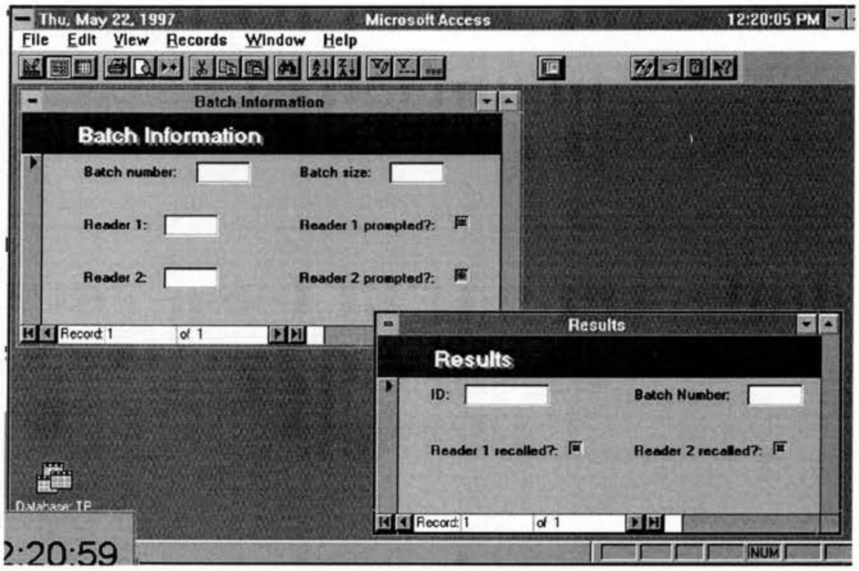
Q7 General Impressions

As Q15 in the previous questionnaire (see page 210)

Thank you for completing this questionnaire

Appendix I

Forms used in the pre-clinical experiment



Batch Information

Batch Information

Batch number: Batch size:

Reader 1: Reader 1 prompted?: ☒

Reader 2: Reader 2 prompted?: ☒

Results

Results

ID: Batch Number:

Reader 1 recalled?: ☒ Reader 2 recalled?: ☒

Figure I.1: The Access forms used in the pre-clinical experiment

Appendix J

The batch results of the pre-clinical experiment

				Prompted			Unprompted		
Batch	Batch size	Cancers in batch	Prompted	reader	cancers	recalls	reader	cancers	recalls
1	98	5	Second	A	5	6	E	5	17
2	102	5	Second	C	5	12	E	5	6
3	100	6	Second	B	5	5	C	5	13
4	101	4	First	B	4	7	C	4	8
5	99	2	First	E	1	7	B	1	5
6	100	8	First	E	7	8	D	6	17
7	101	5	First	C	5	14	E	5	9
8	99	4	First	D	4	6	A	3	5
9	100	5	Second	E	4	7	A	4	2
10	99	7	Second	A	5	9	C	7	6
11	100	6	First	C	6	6	E	6	5
12	101	7	Second	E	7	5	B	7	7
13	99	5	First	D	5	6	E	5	3
14	99	4	First	D	3	5	A	3	9
15	101	6	Second	C	6	10	B	5	7
16	100	7	First	A	7	4	C	7	13
17	102	6	Second	E	3	2	B	6	8
18	100	5	First	E	5	5	D	5	6
19	103	2	Second	B	2	6	D	2	9
20	99	3	Second	E	2	5	B	1	9

Table J.1: Batch results from the pre-clinical experiment

Appendix K

Cancers not recalled by both radiologists

Cases where cancers were recalled by the prompted radiologist, but not by the unprompted radiologist

ID	Batch	Prompted	Prompted Reader	Unprompted Reader	Algorithm result	Type
710411200	6	First	E	D	Cancer prompted	mass
3008469744	6	First	E	D	Cancer prompted	mass
2705340386	8	First	D	A	Cancer prompted	mass
2402329726	15	Second	C	B	Cancer prompted	calc
212451189	20	Second	E	B	Cancer prompted	calc

Cases where cancers were recalled by the unprompted radiologist, but not by the prompted radiologist

ID	Batch	Prompted	Prompted Reader	Unprompted Reader	Algorithm result	Type
1004329709	6	First	E	D	Cancer missed	both
1101441208	10	Second	A	C	Cancer missed	mass
1307381464	10	Second	A	C	Cancer prompted	both
307449726	17	Second	E	B	Cancer prompted	mass
2904261109	17	Second	E	B	Cancer prompted	calc
202431207	17	Second	E	B	Cancer prompted	calc

Cases where cancers were missed by both radiologists

ID	Batch	Prompted	Prompted Reader	Unprompted Reader	Algorithm result	Type
1903431166	3	Second	B	C	Cancer prompted	mass
2512441200	5	First	E	B	Cancer missed	both
2812431024	9	Second	E	A	Cancer prompted	mass
2811381163	14	First	D	A	Cancer missed	mass
1207441023	20	Second	E	B	Cancer missed	calc

Appendix L

The Simulation

L.1 The program

```
#include <stdio.h>
#include <math.h>

/* Definitions for random number generator */
#define IM1 2147483563
#define IM2 2147483399
#define AM (1.0/IM1)
#define IMM1 (IM1-1)
#define IA1 40014
#define IA2 40692
#define IQ1 53668
#define IQ2 52774
#define IR1 12211
#define IR2 3791
#define NTAB 32
#define NDIV (1+IMM1/NTAB)
#define EPS 1.2e-7
#define RNMX (1.0-EPS)

long R, S, box[NTAB], iy;

void initran();
float myrandom();
float gasdev();

main(int argc, char *argv[])
{
  int promam1, promam2, reader1_P, reader1_U, reader2_P, reader2_U;
  int counter[20][4], centres[6]={3,4,5,2,3,3}, centreno, maxrad, radiologist;
  int first, second, prompt;
  int reader1, reader2, k;
  int x, i, j, batch_no, run=0, col1, col2, col3, batchsize;
  int batch[100], rad_recall[100], PROMAM_recall[100];
  float random, size;
  double rad_TP, rad_TN, PROMAM_TP, PROMAM_TN, hit;
  double prob[20][8];

  FILE *data;

  promam1=promam2=reader1_P=reader1_U=reader2_P=reader2_U=0;
  first=second=prompt=k=0;
  reader1=reader2=-1;

/* Initialise the random number generator (prompts for seed) */
  if (argc < 2)
  {
    R = 0;
  } else {
    R = atoi(argv[1]);
  }
  initran();

  for (i=0;i<20;i++) {
     for (j=0;j<4;j++) {
       counter[i][j]=0;
       }
     }

  data = fopen("simulation.dat","r");
  if (!data) {
    printf("Error opening file\n");
    exit(1);
  }

  for (i=0;i<20;i++) {
     for (j=0;j<8;j++) {
        fscanf(data, "%lf", &prob[i][j]);
        }
     }

  hit=0.006; /* Probability of a cancer */

do
  {
    random=gasdev();
    size=random*2.843+6.97;
    batchsize=size*size;
    radiologist=0;
    centreno=myrandom()*6;
    maxrad=centres[centreno];
    for(i=0;i<centreno;i++)
      {
radiologist=radiologist+centres[i];
      }

    x = myrandom()*maxrad;
    reader1 = x+radiologist;
    reader1_P =counter[reader1][0]+batchsize;
    reader1_U =counter[reader1][1]+batchsize;

    do
      {
x=myrandom()*maxrad;
reader2=x+radiologist;
reader2_P=counter[reader2][2]+batchsize;
reader2_U=counter[reader2][3]+batchsize;
      }
    while (reader1==reader2);

      first=abs((promam1 + batchsize) - promam2) + abs(reader1_P -
      (reader1_U - batchsize)) + abs(reader2_U - (reader2_P - batchsize));
      second=abs(promam1 - (promam2 + batchsize)) + abs(reader1_U -
    d  (reader1_P - batchsize)) + abs(reader2_P - (reader2_U - batchsize));
      if (first==second)
{
  if (myrandom() <=0.5)
    prompt=1;
  else prompt=2;
}
      else if (first<second)
prompt=1;
      else prompt=2;
      if (prompt==1)
{
  promam1=promam1+batchsize;
  counter[reader1][0]=counter[reader1][0]+batchsize;
  PROMAM_TN=gasdev()*prob[reader1][1]+prob[reader1][0];
  do
    {
      PROMAM_TN=gasdev()*prob[reader1][1]+prob[reader1][0];
    }
  while (PROMAM_TN<prob[reader1][0]-3*prob[reader1][1] ||
 PROMAM_TN>prob[reader1][0]+3*prob[reader1][1]);
  PROMAM_TP=gasdev()*prob[reader1][3]+prob[reader1][2];
  do
    {
      PROMAM_TP=gasdev()*prob[reader1][3]+prob[reader1][2];
    }
          while (PROMAM_TP<prob[reader1][2]-3*prob[reader1][3] ||
 PROMAM_TP>prob[reader1][2]+3*prob[reader1][3]);

  counter[reader2][3]=counter[reader2][3]+batchsize;
  rad_TN=gasdev()*prob[reader2][5]+prob[reader2][4];
  do
    {
      rad_TN=gasdev()*prob[reader2][5]+prob[reader2][4];
    }
          while (rad_TN<prob[reader1][4]-3*prob[reader1][5] ||
 rad_TN>prob[reader1][4]+3*prob[reader1][5]);
  rad_TP=gasdev()*prob[reader2][7]+prob[reader2][6];
  do
    {
      rad_TP=gasdev()*prob[reader2][7]+prob[reader2][6];
    }
          while (rad_TP<prob[reader1][6]-3*prob[reader1][7] ||
 rad_TP>prob[reader1][6]+3*prob[reader1][7]);
}
else
{
  promam2=promam2+batchsize;
  counter[reader1][1]=counter[reader1][1]+batchsize;
  rad_TN=gasdev()*prob[reader1][5]+prob[reader1][4];
  do
    {
      rad_TN=gasdev()*prob[reader1][5]+prob[reader1][4];
    }
          while (rad_TN<prob[reader1][4]-3*prob[reader1][5] ||
 rad_TN>prob[reader1][4]+3*prob[reader1][5]);
  rad_TP=gasdev()*prob[reader1][7]+prob[reader1][6];
  do
    {
      rad_TP=gasdev()*prob[reader1][7]+prob[reader1][6];
    }
  while (rad_TP<prob[reader1][6]-3*prob[reader1][7] ||
 rad_TP>prob[reader1][6]+3*prob[reader1][7]);


  counter[reader2][2]=counter[reader2][2]+batchsize;
  PROMAM_TN=gasdev()*prob[reader2][1]+prob[reader2][0];
  do
    {
        PROMAM_TN=gasdev()*prob[reader2][1]+prob[reader2][0];
      }
    while (PROMAM_TN<prob[reader1][0]-3*prob[reader1][1] ||
  PROMAM_TN>prob[reader1][0]+3*prob[reader1][1]);

    PROMAM_TP=gasdev()*prob[reader2][3]+prob[reader2][2];
    do
      {
        PROMAM_TP=gasdev()*prob[reader2][3]+prob[reader2][2];
      }
    while (PROMAM_TP<prob[reader1][2]-3*prob[reader1][3] ||
  PROMAM_TP>prob[reader1][2]+3*prob[reader1][3]);

}
        for(j=0; j<batchsize; j++) /* Number of cancers in batch */
{
    if (myrandom() <= hit)
      {
        batch[j] = 1;
        if (myrandom() <= rad_TP)
rad_recall[j] = 1;
        else
rad_recall[j] = 0;
        if (myrandom() <= PROMAM_TP)
PROMAM_recall[j] = 1;
        else
PROMAM_recall[j] = 0;
      }
    else
      {
        batch[j] = 0;
        if (myrandom() <= rad_TN)
rad_recall[j] = 0;
        else
rad_recall[j] = 1;
        if (myrandom() <= PROMAM_TN)
PROMAM_recall[j] = 0;
        else
PROMAM_recall[j] = 1;
      }
    run++;
    printf("%d, %d, %d, %d, %d, %d, %d, %d\n", run, batch[j],
        rad_recall[j], PROMAM_recall[j], centreno, reader1, reader2, prompt);
}
    }
    while (run<100000);
```

```
}

void initran()
{
  int j;
  long k;

  if (R == 0)
  {
    fprintf(stderr, "\nPlease input seed for random number generator\n");
    scanf("%ld", &R);
  }
  if(R < 1)
    R = 1;

  S = R;

  for(j = NTAB+50 ; j >= 0 ; j--)
    {
      k = R / IQ1;
      R = IA1 * (R - k * IQ1) - k * IR1;
      if(R < 0)
R += IM1;

      if(j < NTAB)
box[j] = R;
    }
  iy = box[0];
}


float myrandom()
{
  int j;
  long k;
  float temp;

  k = R / IQ1;
  R = IA1 * (R - k * IQ1) - k * IR1;
  if(R < 0)
    R += IM1;

  k = S / IQ2;
  S = IA2 * (S - k * IQ2) - k * IR2;
  if (S < 0)
    S += IM2;
  j = iy / NDIV;

  iy = box[j] - S;
  box[j] = R;

  if(iy < 1)
    iy += IMM1;

  if((temp=AM*iy) > RNMX)
    return RNMX;
  else
    return temp;
}

float gasdev()
{
static int iset=0;
static float gset;
float fac, rsq, v1, v2;
 if (iset==0) {
   do {
     v1=2.0*myrandom()-1.0;
     v2=2.0*myrandom()-1.0;
     rsq=v1*v1+v2*v2;
   }
   while(rsq>=1.0 || rsq==0.0);
   fac=sqrt(-2.0*log(rsq)/rsq);
   gset=v1*fac;
   iset=1;
   return v2*fac;
 }
 else {
   iset=0;
   return gset;
 }
}
```

L.2 The SAS Program

```
libname prime '~lindaw/Thesis/Simulation/Testsite';
options ls=80 nodate nonumber;

%inc '~lindaw/Thesis/Simulation/Testsite/glmm800.sas';
data centres; infile '~lindaw/Thesis/Simulation/Testsite/centres.dat' dlm=',';
input reader centre;

data work; infile 'test.out' dlm=',';
input run cancer rad promam centre reader1 reader2 prompted;
detect=0;
if cancer=1 and (rad=1 or promam=1) then detect=1;
/* converts cancers into detected-cancers */
data detect; set work; /* calculates McNemars test */
if detect=1;
proc freq;
table promam*rad / nocol norow nopercent agree;
output out=canc mcnem;
title 'Cancer';


data canc; set canc; /* saves p-value in data file called canc */
cancer=p_mcnem;
Obs=obs;
keep Obs cancer;

data rad; set work;    /* stacks data so that each reader is*/
if prompted=1 then reader=reader2;/* displayed in one record */
if prompted=1 then order=2;
if prompted=2 then reader=reader1;
if prompted=2 then order=1;
recall=rad;
paper=0;
keep run detect recall reader order paper centre;

data prompt; set work;
if prompted=1 then reader=reader1;
if prompted=1 then order=1;
if prompted=2 then reader=reader2;
if prompted=2 then order=2;
recall=promam;
paper=1;
keep run detect recall reader order paper centre;

data stack; set rad prompt;
data complete; set stack;
if detect=1;
n=1;
  /*fixed effects model, full spec */
ods output geeemppest=test;
proc genmod data=complete;
class reader paper order run;
model recall/n = paper reader order paper|reader/ dist=b type3 wald;
repeated subject=run / type=CS;
title 'GENMOD - full';
/* saves p-value in dataset called gen1 */

data gen1; set test;
if Parm='paper' and Level1=0;
genmod1=ProbZ;
Obs=obs;
keep Obs genmod1;
/* mixed model, full spec */
%glimmix(data=complete,
stmts=%str(class reader paper order centre run;
model recall/n = paper order/
ddfm=satterth;
random reader reader*paper;
repeated order / subject=run type=cs;
lsmeans paper/ diff pdiff;
title 'GLIMMIX - full';),
error=b);

data mix1; set _diff;     /* saves p-value in dataset called mix1 */
mixed1=probt;
Obs=obs;
keep mixed1 Obs;


proc sort data=stack; by run;

proc freq data=stack;
table reader*paper*order*detect / noprint out=total;

proc freq data=stack; /* summarises data into TP & FN scores */
table reader*paper*order*detect*recall / noprint out=correct;

data temp1; set total;
total=count;
drop count;

data temp2; set correct;if recall=detect;
```

```
proc sort data=temp1; by reader paper order detect; /*merges correct with */
proc sort data=temp2; by reader paper order detect; /*total values */
data all; merge temp1 temp2; by reader paper order detect;
drop recall percent;
proc sort data=all; by reader;
proc sort data=centres; by reader;
data full; merge all centres; by reader;
data summary; set full;
if detect=1;
/* fixed effects model - summary results */
ods output ParameterEstimates=test;
proc genmod data=summary;
class reader paper order;
model count/total = paper reader order paper|reader / dist=b type3 wald;
title 'GENMOD - summary';
  /* saves p-value in dataset called gen2 */
data gen2; set test;
if Parameter='paper' and Level1=0;
genmod2=ProbChisq;
Obs=obs;
keep obs genmod2;
/* mixed model - summary results */
%glimmix(data=summary,
stmts=%str(class reader paper order;
model count/total = paper order  / ddfm=satterth;
random reader reader*paper;
lsmeans paper/ diff pdiff;
title 'GLIMMIX - summary';),
error=b);

data mix2; set _diff;    /* saves p-value in dataset called mix2 */
mixed2=probt;
Obs=obs;
keep mixed2 Obs;

proc sort data=canc; by obs; /* merges all results */
proc sort data=gen1; by obs;
proc sort data=gen2; by obs;
proc sort data=mix1; by obs;
proc sort data=mix2; by obs;

data all; merge canc gen1 gen2 mix1 mix2; by obs;

data sims; set all;
drop obs;
```

```
proc print noobs; /* prints results */
```

L.3 The Shell script

```
#!/usr/local/GNU/bin/bash.new

dir=$1

i=1

while [ $i -lt 1001 ]
do
   rand=‘date +%u%S%M‘

   cd $dir

   ../testing $rand > test.out

   /usr/bin/nice -2 sas ../cancer

   tail -1 cancer.lst >> beta.dat

   i=$(($i + 1))
done
```

Appendix M

Results of the simulation

M.1 The additive model

```
                 The FREQ Procedure
               Table of promam by rad

            promam     rad

            Frequency|       0|       1|  Total
            ---------+--------+--------+
                   0 |      0 |     30 |     30
            ---------+--------+--------+
                   1 |     50 |    497 |    547
            ---------+--------+--------+
            Total          50      527      577

           Statistics for Table of promam by rad

                    McNemar's Test
               -----------------------
               Statistic (S)    5.0000
               DF                    1
               Pr > S           0.0253

               Simple Kappa Coefficient
             --------------------------------
             Kappa                    -0.0695
             ASE                       0.0088
             95% Lower Conf Limit     -0.0868
             95% Upper Conf Limit     -0.0522

                  Sample Size = 577
```

```
                          GENMOD - full

                      The GENMOD Procedure

                       Model Information

         Data Set                         WORK.COMPLETE
         Distribution                          Binomial
         Link Function                            Logit
         Response Variable (Events)              recall
         Response Variable (Trials)                   n
         Observations Used                         1154
         Number Of Events                          1074
         Number Of Trials                          1154


                    Class Level Information

Class     Levels    Values

reader        20    0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
paper          2    0 1
order          2    1 2
run          577    164 423 1026 1247 1268 1589 1910 2387 2663 2680
                    2768 2920 3043 3105 3283 3311 3330 3910 4333 4436
                    4495 4619 4713 4787 4792 4923 5041 5304 5391 5430
                    5642 5717 5779 6031 6135 6173 6312 6338 6376 7050
                    7152 7154 7365 7838 7984 8169 8228 8410 8492 8598
                    8772 ...


                        Parameter Information

         Parameter      Effect       reader    paper    order

         Prm1           Intercept
         Prm2           paper                  0
         Prm3           paper                  1
         Prm4           reader       0
         Prm5           reader       1
         Prm6           reader       2
         Prm7           reader       3
         Prm8           reader       4
         Prm9           reader       5
         Prm10          reader       6
         Prm11          reader       7
         Prm12          reader       8
```

```
          Prm13          reader       9
          Prm14          reader       10
          Prm15          reader       11
          Prm16          reader       12
          Prm17          reader       13
          Prm18          reader       14
          Prm19          reader       15
          Prm20          reader       16
          Prm21          reader       17
          Prm22          reader       18
          Prm23          reader       19
          Prm24          order                          1
          Prm25          order                          2

                              GENMOD - full

                         The GENMOD Procedure

                  Criteria For Assessing Goodness Of Fit

     Criterion                 DF          Value        Value/DF

      Deviance                1132       554.3184         0.4897
      Scaled Deviance         1132       554.3184         0.4897
      Pearson Chi-Square      1132      1143.2811         1.0100
      Scaled Pearson X2       1132      1143.2811         1.0100
      Log Likelihood                    -277.1592


Algorithm converged.


                 Analysis Of Initial Parameter Estimates

                            Standard     Wald 95%         Chi-
Parameter     DF Estimate    Error   Confidence Limits  Square Pr > ChiSq
Intercept      1   2.2495   0.4138   1.4385    3.0604   29.56      <.0001
paper      0   1  -0.6036   0.2431  -1.0801   -0.1271    6.16      0.0130
paper      1   0   0.0000   0.0000   0.0000    0.0000     .          .
reader     0   1   0.7311   0.6400  -0.5233    1.9854    1.30      0.2533
reader     1   1   0.2372   0.5758  -0.8915    1.3658    0.17      0.6805
reader     2   1   0.3307   0.6036  -0.8522    1.5137    0.30      0.5837
reader     3   1   0.4151   0.6466  -0.8523    1.6824    0.41      0.5209
reader     4   1   1.9414   1.0795  -0.1744    4.0572    3.23      0.0721
reader     5   1   1.2945   0.8143  -0.3016    2.8906    2.53      0.1119
reader     6   1   0.6930   0.7094  -0.6975    2.0835    0.95      0.3287
reader     7   1   0.2281   0.7164  -1.1760    1.6322    0.10      0.7502
reader    8   1   1.6048    1.0859   -0.5236    3.7332     2.18       0.1395
reader    9   1   1.9301    1.0799   -0.1865    4.0467     3.19       0.0739
reader    10  1   1.0688    0.8171   -0.5326    2.6702     1.71       0.1908
reader    11  1   0.2403    0.6495   -1.0326    1.5132     0.14       0.7113
reader    12  1   1.0713    0.6375   -0.1782    2.3208     2.82       0.0929
reader    13  1   0.4233    0.5468   -0.6484    1.4951     0.60       0.4388
reader    14  1   0.0480    0.5767   -1.0823    1.1782     0.01       0.9337
reader    15  1   0.0520    0.5522   -1.0304    1.1343     0.01       0.9250
reader    16  1   0.1162    0.5523   -0.9663    1.1986     0.04       0.8334
reader    17  1   1.0917    0.7034   -0.2869    2.4703     2.41       0.1206
reader    18  1   1.5084    0.8123   -0.0837    3.1004     3.45       0.0633
reader    19  0   0.0000    0.0000    0.0000    0.0000      .           .
order     1   1   0.1855    0.2374   -0.2798    0.6509     0.61       0.4345
order     2   0   0.0000    0.0000    0.0000    0.0000      .           .
Scale         0   1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.


                              GEE Model Information

          Correlation Structure                       Exchangeable
          Subject Effect                          run (577 levels)
          Number of Clusters                                   577
          Correlation Matrix Dimension                           2
          Maximum Cluster Size                                   2

                                  GENMOD - full

                              The GENMOD Procedure

                              GEE Model Information

           Minimum Cluster Size                                   2


Algorithm converged.


                     Analysis Of GEE Parameter Estimates
                     Empirical Standard Error Estimates

                             Standard   95% Confidence
Parameter     Estimate     Error         Limits            Z Pr > |Z|

Intercept       2.2489    0.4159   1.4336    3.0641     5.41   <.0001
paper      0   -0.6006    0.2512  -1.0930   -0.1082    -2.39   0.0168
paper     1   0.0000   0.0000   0.0000   0.0000      .       .
reader    0   0.7153   0.6350  -0.5293   1.9599    1.13   0.2600
reader    1   0.2369   0.5793  -0.8985   1.3723    0.41   0.6826
reader    2   0.3317   0.6083  -0.8605   1.5239    0.55   0.5855
reader    3   0.4089   0.6555  -0.8760   1.6937    0.62   0.5328
reader    4   1.8789   1.0244  -0.1289   3.8866    1.83   0.0666
reader    5   1.2961   0.8119  -0.2952   2.8873    1.60   0.1104
reader    6   0.7391   0.7294  -0.6904   2.1687    1.01   0.3109
reader    7   0.2412   0.7221  -1.1740   1.6564    0.33   0.7383
reader    8   1.6269   1.0982  -0.5256   3.7794    1.48   0.1385
reader    9   1.9439   1.1110  -0.2336   4.1214    1.75   0.0802
reader   10   0.9850   0.7675  -0.5192   2.4892    1.28   0.1993
reader   11   0.2610   0.6586  -1.0298   1.5517    0.40   0.6919
reader   12   1.0620   0.6297  -0.1721   2.2961    1.69   0.0917
reader   13   0.4295   0.5545  -0.6573   1.5163    0.77   0.4386
reader   14   0.0493   0.5729  -1.0736   1.1723    0.09   0.9314
reader   15   0.0500   0.5559  -1.0395   1.1395    0.09   0.9284
reader   16   0.1045   0.5585  -0.9902   1.1992    0.19   0.8515
reader   17   1.0811   0.6993  -0.2895   2.4516    1.55   0.1221
reader   18   1.5054   0.8291  -0.1196   3.1304    1.82   0.0694
reader   19   0.0000   0.0000   0.0000   0.0000      .       .
order     1   0.1901   0.2466  -0.2932   0.6734    0.77   0.4407
order     2   0.0000   0.0000   0.0000   0.0000      .       .


               Wald Statistics For Type 3 GEE Analysis

                                      Chi-
                  Source       DF    Square    Pr > ChiSq

                  paper         1      5.72        0.0168
                  reader       19     17.59        0.5503
                  order         1      0.59        0.4407

                             GLIMMIX - FULL

                          The Mixed Procedure
                           Model Information

              Data Set                     WORK._DS
              Dependent Variable           _z
              Weight Variable              _w
              Covariance Structures        Variance Components,
                                           Compound Symmetry
              Subject Effect               run
              Estimation Method            REML
              Residual Variance Method     Profile
```

```
      Fixed Effects SE Method       Model-Based
      Degrees of Freedom Method     Satterthwaite


                   Class Level Information

        Class     Levels    Values

        reader        20    0 1 2 3 4 5 6 7 8 9 10 11 12
                            13 14 15 16 17 18 19
        paper          2    0 1
        order          2    1 2
        centre         6    0 1 2 3 4 5

                       GLIMMIX - FULL

                    The Mixed Procedure

                   Class Level Information

       Class     Levels    Values

       run          577    164 423 1026 1247 1268 1589
                                1910 2387 2663 2680 2768 2920
                                3043 3105 3283 3311 3330 3910
                                4333 4436 4495 4619 4713 4787
                                4792 4923 5041 5304 5391 5430
                                5642 5717 5779 6031 6135 6173
                                6312 6338 6376 7050 7152 7154
                                7365 7838 7984 8169 8228 8410
                                8492....(truncated for space)

                          Dimensions

               Covariance Parameters              3
               Columns in X                       5
               Columns in Z                      20
               Subjects                           1
               Max Obs Per Subject             1154
               Observations Used               1154
               Observations Not Used              0
               Total Observations              1154


                       Parameter Search

CovP1     CovP2    CovP3 Variance   Res Log Like  -2 Res Log Like
```

```
0.02047 -0.07183  1.0591   1.0591     -3232.8047          6465.6094


                           Iteration History

    Iteration    Evaluations     -2 Res Log Like        Criterion

            1              1      6465.60935097       0.00000000


                       Convergence criteria met.


                    Covariance Parameter Estimates

                    Cov Parm      Subject    Estimate

                    reader                    0.02047
                    CS            run        -0.07183
                    Residual                   1.0591


                           Fit Statistics

                -2 Res Log Likelihood          6465.6
                 AIC (smaller is better)        6471.6
                 AICC (smaller is better)       6471.6
                 BIC (smaller is better)        6474.6

                             GLIMMIX - FULL

                          The Mixed Procedure

                  PARMS Model Likelihood Ratio Test

                      DF     Chi-Square        Pr > ChiSq

                       2           0.00            1.0000

                     Solution for Fixed Effects

                                           Standard
Effect      paper  order  Estimate      Error    DF  t Value   Pr > |t|
Intercept                   2.8224     0.2174   171    12.98     <.0001
paper       0              -0.5564     0.2460   592    -2.26     0.0240
paper       1                    0          .     .        .          .
```

```
order             1        0.1803   0.2398  596     0.75     0.4524
order             2             0        .    .      .          .


                       Type 3 Tests of Fixed Effects

                              Num     Den
           Effect             DF      DF     F Value     Pr > F

           paper               1     592        5.12     0.0240
           order               1     596        0.57     0.4524


                           Least Squares Means

                           Standard
Effect  paper  Estimate    Error     DF    t Value    Pr > |t|

paper   0        2.3561   0.1508   45.2      15.62      <.0001
paper   1        2.9125   0.1898    106      15.35      <.0001


                     Differences of Least Squares Means

                                   Standard
Effect  paper  _paper  Estimate    Error     DF  t Value  Pr > |t|

paper   0      1        -0.5564   0.2460    592    -2.26    0.0240


                               GLIMMIX - FULL
                           GLIMMIX Model Statistics

              Description                         Value

              Deviance                         572.5856
              Scaled Deviance                  540.6488
              Pearson Chi-Square              1134.9172
              Scaled Pearson Chi-Square       1071.6155
              Extra-Dispersion Scale             1.0591

                            GENMOD - summary

                          The GENMOD Procedure

                            Model Information
```

```
Data Set                          WORK.SUMMARY
Distribution                          Binomial
Link Function                            Logit
Response Variable (Events)               COUNT    Frequency Count
Response Variable (Trials)               total
Observations Used                           80
Number Of Events                          1074
Number Of Trials                          1154


                     Class Level Information

Class  Levels   Values

reader     20   0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
paper       2   0 1
order       2   1 2


                Criteria For Assessing Goodness Of Fit

     Criterion                 DF           Value        Value/DF

     Deviance                  58         57.4894          0.9912
     Scaled Deviance           58         57.4894          0.9912
     Pearson Chi-Square        58         51.4676          0.8874
     Scaled Pearson X2         58         51.4676          0.8874
     Log Likelihood                     -277.1592


Algorithm converged.


                  Analysis Of Parameter Estimates

                              Standard      Wald 95%        Chi-
Parameter      DF Estimate      Error  Confidence Limits  Square Pr > ChiSq

Intercept       1   2.2495    0.4138   1.4385    3.0604   29.56      <.0001
paper       0   1  -0.6036    0.2431  -1.0801   -0.1271    6.16      0.0130
paper       1   0   0.0000    0.0000   0.0000    0.0000     .          .
reader      0   1   0.7311    0.6400  -0.5233    1.9854    1.30      0.2533
reader      1   1   0.2372    0.5758  -0.8915    1.3658    0.17      0.6805
reader      2   1   0.3307    0.6036  -0.8522    1.5137    0.30      0.5837
reader      3   1   0.4151    0.6466  -0.8523    1.6824    0.41      0.5209
reader      4   1   1.9414    1.0795  -0.1744    4.0572    3.23      0.0721
reader      5   1   1.2945    0.8143  -0.3016    2.8906    2.53      0.1119
reader     6   1   0.6930   0.7094  -0.6975   2.0835   0.95      0.3287
reader     7   1   0.2281   0.7164  -1.1760   1.6322   0.10      0.7502
reader     8   1   1.6048   1.0859  -0.5236   3.7332   2.18      0.1395
reader     9   1   1.9301   1.0799  -0.1865   4.0467   3.19      0.0739
reader    10   1   1.0688   0.8171  -0.5326   2.6702   1.71      0.1908
reader    11   1   0.2403   0.6495  -1.0326   1.5132   0.14      0.7113
reader    12   1   1.0713   0.6375  -0.1782   2.3208   2.82      0.0929
reader    13   1   0.4233   0.5468  -0.6484   1.4951   0.60      0.4388
reader    14   1   0.0480   0.5767  -1.0823   1.1782   0.01      0.9337
reader    15   1   0.0520   0.5522  -1.0304   1.1343   0.01      0.9250
reader    16   1   0.1162   0.5523  -0.9663   1.1986   0.04      0.8334
reader    17   1   1.0917   0.7034  -0.2869   2.4703   2.41      0.1206
reader    18   1   1.5084   0.8123  -0.0837   3.1004   3.45      0.0633
reader    19   0   0.0000   0.0000   0.0000   0.0000    .          .
order      1   1   0.1855   0.2374  -0.2798   0.6509   0.61      0.4345
order      2   0   0.0000   0.0000   0.0000   0.0000    .          .
Scale          0   1.0000   0.0000   1.0000   1.0000

NOTE: The scale parameter was held fixed.


                  Wald Statistics For Type 3 Analysis

                                          Chi-
                  Source          DF    Square    Pr > ChiSq

                  paper            1      6.16        0.0130
                  reader          19     17.45        0.5593
                  order            1      0.61        0.4345


                           The Mixed Procedure

                            Model Information

              Data Set                     WORK._DS
              Dependent Variable           _z
              Weight Variable              _w
              Covariance Structure         Variance Components
              Estimation Method            REML
              Residual Variance Method     Profile
              Fixed Effects SE Method      Model-Based
              Degrees of Freedom Method    Satterthwaite


                        Class Level Information
```

```
          Class     Levels    Values

          reader        20    0 1 2 3 4 5 6 7 8 9 10 11 12
                              13 14 15 16 17 18 19
          paper          2    0 1
          order          2    1 2


                            Dimensions

                 Covariance Parameters              2
                 Columns in X                       5
                 Columns in Z                      20
                 Subjects                           1
                 Max Obs Per Subject               80
                 Observations Used                 80
                 Observations Not Used              0
                 Total Observations                80


                        Parameter Search

  CovP1     CovP2  Variance     Res Log Like  -2 Res Log Like

0.06873    0.7951    0.7951        -115.1392         230.2785


                         Iteration History

    Iteration    Evaluations    -2 Res Log Like       Criterion

            1              1       230.27845497      0.00000000


                     Convergence criteria met.


                        The Mixed Procedure

                        Covariance Parameter
                              Estimates

                       Cov Parm      Estimate

                       reader         0.06873
                       Residual        0.7951
```

```
                         Fit Statistics

              -2 Res Log Likelihood           230.3
              AIC (smaller is better)         234.3
              AICC (smaller is better)        234.4
              BIC (smaller is better)         236.3


                PARMS Model Likelihood Ratio Test

                 DF     Chi-Square       Pr > ChiSq

                  1           0.00           1.0000


                   Solution for Fixed Effects

                                         Standard
Effect     paper order Estimate     Error     DF  t Value  Pr > |t|

Intercept               2.8321     0.2003  70.3     14.14    <.0001
paper      0           -0.5636     0.2137  59.9     -2.64    0.0106
paper      1                 0          .     .         .         .
order            1      0.1815     0.2084  60.1      0.87    0.3874
order            2           0          .     .         .         .


                  Type 3 Tests of Fixed Effects

                          Num       Den
           Effect          DF        DF    F Value     Pr > F

           paper            1      59.9       6.95     0.0106
           order            1      60.1       0.76     0.3874


                      Least Squares Means

                                       Standard
Effect     paper    Estimate              Error        DF     t Value      Pr > |t|

paper      0          2.3592             0.1452      36.2       16.24        <.0001
paper      1          2.9228             0.1785        60       16.37        <.0001
                                          GLIMMIX - SUMMARY

                      The Mixed Procedure
                Differences of Least Squares Means

                              Standard
Effect  paper  _paper  Estimate   Error    DF  t Value Pr > |t|

paper   0      1        -0.5636  0.2137  59.9    -2.64   0.0106

                     GLIMMIX Model Statistics

               Description                    Value

               Deviance                     70.1043
               Scaled Deviance              88.1676
               Pearson Chi-Square           57.6263
               Scaled Pearson Chi-Square    72.4744
               Extra-Dispersion Scale        0.7951
```

M.2 The interaction model

```
                    Cancer

              The FREQ Procedure

            Table of promam by rad

     promam     rad

     Frequency|       0|       1|  Total
     ---------+--------+--------+
            0 |      0 |     30 |     30
     ---------+--------+--------+
            1 |     50 |    497 |    547
     ---------+--------+--------+
     Total          50      527      577


    Statistics for Table of promam by rad

               McNemar's Test
           -----------------------
           Statistic (S)    5.0000
           DF                    1
           Pr > S           0.0253


           Simple Kappa Coefficient
       --------------------------------
       Kappa                    -0.0695
       ASE                       0.0088
       95% Lower Conf Limit     -0.0868
       95% Upper Conf Limit     -0.0522

              Sample Size = 577
```

```
                         GENMOD - full

                     The GENMOD Procedure

                       Model Information

      Data Set                           WORK.COMPLETE
      Distribution                            Binomial
      Link Function                              Logit
      Response Variable (Events)                recall
      Response Variable (Trials)                     n
      Observations Used                           1154
      Number Of Events                            1074
      Number Of Trials                            1154


                    Class Level Information

Class    Levels  Values

reader       20  0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
paper         2  0 1
order         2  1 2
run         577  164 423 1026 1247 1268 1589 1910 2387 2663 2680
                 2768 2920 3043 3105 3283 3311 3330 3910 4333 4436
                 4495 4619 4713 4787 4792 4923 5041 5304 5391 5430
                 5642 5717 5779 6031 6135 6173 6312 6338 6376 7050
                 7152 7154 7365 7838 7984 8169 8228 8410 8492 8598
                 8772 ...


                     Parameter Information

    Parameter        Effect          reader     paper     order

    Prm1             Intercept
    Prm2             paper                      0
    Prm3             paper                      1
    Prm4             reader          0
    Prm5             reader          1
    Prm6             reader          2
    Prm7             reader          3
    Prm8             reader          4
    Prm9             reader          5
    Prm10            reader          6
    Prm11            reader          7
    Prm12            reader          8
```

Prm13	reader	9		
Prm14	reader	10		
Prm15	reader	11		
Prm16	reader	12		
Prm17	reader	13		
Prm18	reader	14		
Prm19	reader	15		
Prm20	reader	16		
Prm21	reader	17		
Prm22	reader	18		
Prm23	reader	19		
Prm24	order			1
Prm25	order			2
Prm26	reader*paper	0	0	
Prm27	reader*paper	0	1	
Prm28	reader*paper	1	0	
Prm29	reader*paper	1	1	
Prm30	reader*paper	2	0	
Prm31	reader*paper	2	1	
Prm32	reader*paper	3	0	
Prm33	reader*paper	3	1	
Prm34	reader*paper	4	0	
Prm35	reader*paper	4	1	
Prm36	reader*paper	5	0	
Prm37	reader*paper	5	1	
Prm38	reader*paper	6	0	
Prm39	reader*paper	6	1	
Prm40	reader*paper	7	0	
Prm41	reader*paper	7	1	
Prm42	reader*paper	8	0	
Prm43	reader*paper	8	1	
Prm44	reader*paper	9	0	
Prm45	reader*paper	9	1	
Prm46	reader*paper	10	0	
Prm47	reader*paper	10	1	
Prm48	reader*paper	11	0	
Prm49	reader*paper	11	1	
Prm50	reader*paper	12	0	
Prm51	reader*paper	12	1	
Prm52	reader*paper	13	0	
Prm53	reader*paper	13	1	
Prm54	reader*paper	14	0	
Prm55	reader*paper	14	1	
Prm56	reader*paper	15	0	
Prm57	reader*paper	15	1	
Prm58	reader*paper	16	0	
Prm59	reader*paper	16	1	

```
     Prm60          reader*paper    17        0
     Prm61          reader*paper    17        1
     Prm62          reader*paper    18        0
     Prm63          reader*paper    18        1
     Prm64          reader*paper    19        0
     Prm65          reader*paper    19        1


                  Criteria For Assessing Goodness Of Fit

    Criterion                   DF           Value         Value/DF

    Deviance                  1113        535.4126           0.4811
    Scaled Deviance           1113        535.4126           0.4811
    Pearson Chi-Square        1113        965.8606           0.8678
    Scaled Pearson X2         1113        965.8606           0.8678
    Log Likelihood                       -267.7063

WARNING: Negative of Hessian not positive definite.


                  Analysis Of Initial Parameter Estimates

                                  Standard   Wald 95% Confidence     Chi-
Parameter      DF    Estimate       Error          Limits          Square

Intercept       1      2.1078      0.5401      1.0492      3.1664    15.23
paper      0    1     -0.3213      0.7528     -1.7967      1.1542     0.18
paper      1    0      0.0000      0.0000      0.0000      0.0000      .
reader     0    1      0.8117      0.8964     -0.9452      2.5685     0.82
reader     1    1      0.1316      0.9093     -1.6507      1.9138     0.02
reader     2    1      0.3428      0.9048     -1.4305      2.1162     0.14
reader     3    1      0.3557      0.9053     -1.4187      2.1301     0.15
reader     4    1     24.1426      1.1531     21.8824     26.4027   438.32
reader     5    1     24.1873      0.9105     22.4027     25.9719   705.69
reader     6    1      0.7025      1.1550     -1.5613      2.9664     0.37
reader     7    1      0.6839      1.1550     -1.5799      2.9478     0.35
reader     8    1     24.1722      1.1566     21.9053     26.4392   436.76
reader     9    1      0.9025      1.1506     -1.3527      3.1576     0.62
reader    10    1     24.1380      0.9175     22.3398     25.9362   692.18
reader    11    1      0.1926      0.9102     -1.5914      1.9766     0.04
reader    12    1     24.1829      0.7498     22.7134     25.6524  1040.33
reader    13    1      0.1841      0.7426     -1.2714      1.6396     0.06
reader    14    1      1.1284      1.1476     -1.1208      3.3775     0.97
reader    15    1      0.1178      0.8032     -1.4565      1.6921     0.02
reader    16    1     -0.3454      0.7542     -1.8237      1.1329     0.21
```

```
                    Analysis Of Initial
                    Parameter Estimates

               Parameter                Pr > ChiSq

               Intercept                    <.0001
               paper            0           0.6696
               paper            1            .
               reader           0           0.3652
               reader           1           0.8849
               reader           2           0.7047
               reader           3           0.6944
               reader           4           <.0001
               reader           5           <.0001
               reader           6           0.5430
               reader           7           0.5538
               reader           8           <.0001
               reader           9           0.4328
               reader           10          <.0001
               reader           11          0.8324
               reader           12          <.0001
               reader           13          0.8042
               reader           14          0.3255
               reader           15          0.8834
               reader           16          0.6470


                Analysis Of Initial Parameter Estimates

                                    Standard  Wald 95% Confidence    Chi-
Parameter           DF  Estimate     Error        Limits           Square

reader        17     1   24.1692    0.8074   22.5866   25.7517   895.99
reader        18     1    1.2708    1.1443   -0.9720    3.5136     1.23
reader        19     0    0.0000    0.0000    0.0000    0.0000      .
order         1      1    0.1762    0.2396   -0.2933    0.6458     0.54
order         2      0    0.0000    0.0000    0.0000    0.0000      .
reader*paper  0   0  1   -0.1540    1.2776   -2.6581    2.3501     0.01
reader*paper  0   1  0    0.0000    0.0000    0.0000    0.0000      .
reader*paper  1   0  1    0.0811    1.1822   -2.2359    2.3981     0.00
reader*paper  1   1  0    0.0000    0.0000    0.0000    0.0000      .
reader*paper  2   0  1   -0.0641    1.2166   -2.4485    2.3203     0.00
reader*paper  2   1  0    0.0000    0.0000    0.0000    0.0000      .
reader*paper  3   0  1    0.0992    1.2874   -2.4240    2.6223     0.01
reader*paper  3   1  0    0.0000    0.0000    0.0000    0.0000      .
reader*paper  4   0  0  -22.7920    0.0000  -22.7920  -22.7920      .
```

```
reader*paper  4  1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  5  0  0  -23.5362   0.0000  -23.5362  -23.5362       .
reader*paper  5  1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  6  0  1   -0.0823   1.4713   -2.9659    2.8014    0.00
reader*paper  6  1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  7  0  1   -0.8023   1.4860   -3.7148    2.1101    0.29
reader*paper  7  1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  8  0  0  -22.9751   0.0000  -22.9751  -22.9751       .
reader*paper  8  1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  9  0  1   23.6008 102152.9   -200192  200239.6    0.00
reader*paper  9  1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  10 0  0  -23.7624   0.0000  -23.7624  -23.7624       .
reader*paper  10 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  11 0  1    0.0805   1.2945   -2.4567    2.6177    0.00
reader*paper  11 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  12 0  0  -23.6813   0.0000  -23.6813  -23.6813       .
reader*paper  12 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  13 0  1    0.4728   1.0956   -1.6746    2.6202    0.19
reader*paper  13 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  14 0  1   -1.6179   1.3625   -4.2884    1.0526    1.41
reader*paper  14 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  15 0  1   -0.1415   1.1067   -2.3105    2.0276    0.02
reader*paper  15 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  16 0  1    0.8234   1.1064   -1.3451    2.9919    0.55
reader*paper  16 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  17 0  0  -23.6010   0.0000  -23.6010  -23.6010       .
reader*paper  17 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  18 0  1    0.3821   1.6211   -2.7952    3.5593    0.06
reader*paper  18 1  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  19 0  0    0.0000   0.0000    0.0000    0.0000       .
reader*paper  19 1  0    0.0000   0.0000    0.0000    0.0000       .
Scale               0    1.0000   0.0000    1.0000    1.0000


                          Analysis Of Initial
                          Parameter Estimates

                    Parameter              Pr > ChiSq

                    reader         17          <.0001
                    reader         18          0.2668
                    reader         19               .
                    order          1           0.4620
                    order          2                .
                    reader*paper   0     0     0.9040
                    reader*paper   0     1          .
                    reader*paper   1     0     0.9453
                    reader*paper   1    1          .
                    reader*paper   2    0     0.9580
                    reader*paper   2    1          .
                    reader*paper   3    0     0.9386
                    reader*paper   3    1          .
                    reader*paper   4    0          .
                    reader*paper   4    1          .
                    reader*paper   5    0          .
                    reader*paper   5    1          .
                    reader*paper   6    0     0.9554
                    reader*paper   6    1          .
                    reader*paper   7    0     0.5892
                    reader*paper   7    1          .
                    reader*paper   8    0          .
                    reader*paper   8    1          .
                    reader*paper   9    0     0.9998
                    reader*paper   9    1          .
                    reader*paper   10   0          .
                    reader*paper   10   1          .
                    reader*paper   11   0     0.9504
                    reader*paper   11   1          .
                    reader*paper   12   0          .
                    reader*paper   12   1          .
                    reader*paper   13   0     0.6661
                    reader*paper   13   1          .
                    reader*paper   14   0     0.2351
                    reader*paper   14   1          .
                    reader*paper   15   0     0.8983
                    reader*paper   15   1          .
                    reader*paper   16   0     0.4568
                    reader*paper   16   1          .
                    reader*paper   17   0          .
                    reader*paper   17   1          .
                    reader*paper   18   0     0.8137
                    reader*paper   18   1          .
                    reader*paper   19   0          .
                    reader*paper   19   1          .
                    Scale

NOTE: The scale parameter was held fixed.
```

```
                      GEE Model Information

           Correlation Structure                    Exchangeable
           Subject Effect                       run (577 levels)
           Number of Clusters                                577
           Correlation Matrix Dimension                        2
           Maximum Cluster Size                                2
           Minimum Cluster Size                                2
WARNING: The generalized Hessian matrix is not positive definite.
         Iteration will be terminated.


ERROR: Error in estimation routine.


                  Analysis Of GEE Parameter Estimates
                   Empirical Standard Error Estimates

                               Standard   95% Confidence
Parameter            Estimate     Error       Limits            Z Pr > |Z|

Intercept              2.1078      .          .         .         .      .
paper          0      -0.3213      .          .         .         .      .
paper          1       0.0000   0.0000     0.0000    0.0000       .      .
reader         0       0.8117      .          .         .         .      .
reader         1       0.1316      .          .         .         .      .
reader         2       0.3428      .          .         .         .      .
reader         3       0.3557      .          .         .         .      .
reader         4      24.1426      .          .         .         .      .
reader         5      24.1873      .          .         .         .      .
reader         6       0.7025      .          .         .         .      .
reader         7       0.6839      .          .         .         .      .
reader         8      24.1722      .          .         .         .      .
reader         9       0.9025      .          .         .         .      .
reader         10     24.1380      .          .         .         .      .
reader         11      0.1926      .          .         .         .      .
reader         12     24.1829      .          .         .         .      .
reader         13      0.1841      .          .         .         .      .
reader         14      1.1284      .          .         .         .      .
reader         15      0.1178      .          .         .         .      .
reader         16     -0.3454      .          .         .         .      .
reader         17     24.1692      .          .         .         .      .
reader         18      1.2708      .          .         .         .      .
reader         19      0.0000   0.0000     0.0000    0.0000       .      .
order          1       0.1762      .          .         .         .      .
order          2       0.0000   0.0000     0.0000    0.0000       .      .
```

```
reader*paper 0  0   -0.1540     .         .         .        .      .
reader*paper 0  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 1  0    0.0811     .         .         .        .      .
reader*paper 1  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 2  0   -0.0641     .         .         .        .      .
reader*paper 2  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 3  0    0.0992     .         .         .        .      .
reader*paper 3  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 4  0 -22.7920      .         .         .        .      .
reader*paper 4  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 5  0 -23.5362      .         .         .        .      .
reader*paper 5  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 6  0   -0.0823     .         .         .        .      .
reader*paper 6  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 7  0   -0.8023     .         .         .        .      .
reader*paper 7  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 8  0 -22.9751      .         .         .        .      .
reader*paper 8  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 9  0   23.6008     .         .         .        .      .
reader*paper 9  1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 10 0 -23.7624      .         .         .        .      .
reader*paper 10 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 11 0    0.0805     .         .         .        .      .
reader*paper 11 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 12 0 -23.6813      .         .         .        .      .
reader*paper 12 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 13 0    0.4728     .         .         .        .      .
reader*paper 13 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 14 0   -1.6179     .         .         .        .      .
reader*paper 14 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 15 0   -0.1415     .         .         .        .      .
reader*paper 15 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 16 0    0.8234     .         .         .        .      .
reader*paper 16 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 17 0 -23.6010      .         .         .        .      .
reader*paper 17 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 18 0    0.3821     .         .         .        .      .
reader*paper 18 1    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 19 0    0.0000   0.0000    0.0000    0.0000     .      .
reader*paper 19 1    0.0000   0.0000    0.0000    0.0000     .      .
```

```
                          GLIMMIX - FULL

                       The Mixed Procedure

                        Model Information

          Data Set                     WORK._DS
          Dependent Variable           _z
          Weight Variable              _w
          Covariance Structures        Variance Components,
                                       Compound Symmetry
          Subject Effect               run
          Estimation Method            REML
          Residual Variance Method     Profile
          Fixed Effects SE Method      Model-Based
          Degrees of Freedom Method    Satterthwaite

                     Class Level Information

            Class     Levels    Values

            reader        20    0 1 2 3 4 5 6 7 8 9 10 11 12
                                13 14 15 16 17 18 19
            paper          2    0 1
            order          2    1 2
            centre         6    0 1 2 3 4 5
            run          577    164 423 1026 1247 1268 1589
                                1910 2387 2663 2680 2768 2920
                                3043 3105 3283 3311 3330 3910
                                .... (truncated for space)

                         Dimensions

             Covariance Parameters              4
             Columns in X                       5
             Columns in Z                      60
             Subjects                           1
             Max Obs Per Subject             1154
             Observations Used               1154
             Observations Not Used              0
             Total Observations              1154

                       Parameter Search

  CovP1   CovP2     CovP3    CovP4 Variance  Res Log Like -2 Res Log Like

0.02047       0  -0.07183   1.0591   1.0591    -3232.8047       6465.6093
```

```
                     Iteration History

  Iteration    Evaluations    -2 Res Log Like       Criterion

          1              1     6465.60934859      0.00000000

                  Convergence criteria met.


               Covariance Parameter Estimates

              Cov Parm          Subject    Estimate

              reader                        0.02047
              reader*paper                        0
              CS                run        -0.07183
              Residual                       1.0591


                      Fit Statistics

             -2 Res Log Likelihood          6465.6
             AIC (smaller is better)        6471.6
             AICC (smaller is better)       6471.6
             BIC (smaller is better)        6474.6


              PARMS Model Likelihood Ratio Test

                  DF     Chi-Square      Pr > ChiSq
                   2           0.00          1.0000


                Solution for Fixed Effects

                                        Standard
Effect       paper  order  Estimate      Error    DF   t Value   Pr > |t|

Intercept                    2.8224     0.2174   171     12.98     <.0001
paper        0              -0.5564     0.2460   592     -2.26     0.0240
paper        1                    0          .     .       .         .
order               1        0.1803     0.2398   596      0.75     0.4524
order               2             0          .     .       .         .
```

```
                        Type 3 Tests of Fixed Effects

                            Num      Den
          Effect             DF       DF    F Value     Pr > F

          paper               1      592       5.12     0.0240
          order               1      596       0.57     0.4524

                             Least Squares Means

                                        Standard
Effect    paper    Estimate            Error       DF     t Value      Pr > |t|

paper     0          2.3561           0.1508     45.2       15.62        <.0001
paper     1          2.9125           0.1898      106       15.35        <.0001

                      Differences of Least Squares Means

                                          Standard
Effect  paper  _paper  Estimate          Error    DF  t Value  Pr > |t|

paper   0      1         -0.5564        0.2460   592    -2.26    0.0240


               Description                         Value

               Deviance                         572.5856
               Scaled Deviance                  540.6488
               Pearson Chi-Square              1134.9172
               Scaled Pearson Chi-Square       1071.6155
               Extra-Dispersion Scale             1.0591


                             The GENMOD Procedure

                              Model Information

Data Set                           WORK.SUMMARY
Distribution                           Binomial
Link Function                             Logit
Response Variable (Events)                COUNT    Frequency Count
Response Variable (Trials)                total
Observations Used                            80
Number Of Events                           1074
Number Of Trials                           1154
```

```
                    Class Level Information

Class      Levels     Values

reader         20     0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
paper           2     0 1
order           2     1 2


             Criteria For Assessing Goodness Of Fit

     Criterion                 DF          Value        Value/DF

     Deviance                  39        38.5836          0.9893
     Scaled Deviance           39        38.5836          0.9893
     Pearson Chi-Square        39        31.0955          0.7973
     Scaled Pearson X2         39        31.0955          0.7973
     Log Likelihood                     -267.7063


WARNING: Negative of Hessian not positive definite.


                Analysis Of Parameter Estimates

                             Standard   Wald 95% Confidence      Chi-
Parameter      DF  Estimate     Error          Limits          Square

Intercept       1    2.1078    0.5401     1.0492     3.1664     15.23
paper      0    1   -0.3213    0.7528    -1.7967     1.1542      0.18
paper      1    0    0.0000    0.0000     0.0000     0.0000       .


                Analysis Of Parameter Estimates

               Parameter                Pr > ChiSq

               Intercept                    <.0001
               paper          0             0.6696
               paper          1               .
```

```
                     Analysis Of Parameter Estimates

                                      Standard  Wald 95% Confidence    Chi-
Parameter            DF  Estimate       Error       Limits            Square

reader        0       1    0.8117     0.8964   -0.9452    2.5685       0.82
reader        1       1    0.1316     0.9093   -1.6507    1.9138       0.02
reader        2       1    0.3428     0.9048   -1.4305    2.1162       0.14
reader        3       1    0.3557     0.9053   -1.4187    2.1301       0.15
reader        4       1   24.6451     1.1531   22.3849   26.9052     456.76
reader        5       1   24.7784     0.9105   22.9938   26.5629     740.60
reader        6       1    0.7025     1.1550   -1.5613    2.9664       0.37
reader        7       1    0.6839     1.1550   -1.5799    2.9478       0.35
reader        8       1   23.9104     1.1566   21.6435   26.1774     427.36
reader        9       1    0.9025     1.1506   -1.3527    3.1576       0.62
reader        10      1   24.5846     0.9175   22.7864   26.3828     718.04
reader        11      1    0.1926     0.9102   -1.5914    1.9766       0.04
reader        12      1   25.1291     0.7498   23.6596   26.5986    1123.33
reader        13      1    0.1841     0.7426   -1.2714    1.6396       0.06
reader        14      1    1.1284     1.1476   -1.1208    3.3775       0.97
reader        15      1    0.1178     0.8032   -1.4565    1.6921       0.02
reader        16      1   -0.3454     0.7542   -1.8237    1.1329       0.21
reader        17      1   24.7851     0.8074   23.2025   26.3676     942.24
reader        18      1    1.2708     1.1443   -0.9720    3.5136       1.23
reader        19      0    0.0000     0.0000    0.0000    0.0000        .
order         1       1    0.1762     0.2396   -0.2933    0.6458       0.54
order         2       0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  0   0   1   -0.1540     1.2776   -2.6581    2.3501       0.01
reader*paper  0   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  1   0   1    0.0811     1.1822   -2.2359    2.3981       0.00
reader*paper  1   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  2   0   1   -0.0641     1.2166   -2.4485    2.3203       0.00
reader*paper  2   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  3   0   1    0.0992     1.2874   -2.4240    2.6223       0.01
reader*paper  3   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  4   0   0  -23.2944     0.0000  -23.2944  -23.2944        .
reader*paper  4   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  5   0   0  -24.1273     0.0000  -24.1273  -24.1273        .
reader*paper  5   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  6   0   1   -0.0823     1.4713   -2.9659    2.8014       0.00
reader*paper  6   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  7   0   1   -0.8023     1.4860   -3.7148    2.1101       0.29
reader*paper  7   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  8   0   0  -22.7133     0.0000  -22.7133  -22.7133        .
reader*paper  8   1   0    0.0000     0.0000    0.0000    0.0000        .
reader*paper  9   0   1   24.1326   133270.8   -261182  261230.1       0.00
reader*paper  9   1   0    0.0000     0.0000    0.0000    0.0000        .
```

```
reader*paper  10  0   0  -24.2090     0.0000  -24.2090  -24.2090         .
reader*paper  10  1   0    0.0000     0.0000    0.0000    0.0000         .
reader*paper  11  0   1    0.0805     1.2945   -2.4567    2.6177      0.00
reader*paper  11  1   0    0.0000     0.0000    0.0000    0.0000         .
reader*paper  12  0   0  -24.6275     0.0000  -24.6275  -24.6275         .
reader*paper  12  1   0    0.0000     0.0000    0.0000    0.0000         .
reader*paper  13  0   1    0.4728     1.0956   -1.6746    2.6202      0.19
reader*paper  13  1   0    0.0000     0.0000    0.0000    0.0000         .


                     Analysis Of Parameter Estimates

                    Parameter                Pr > ChiSq

                    reader         0             0.3652
                    reader         1             0.8849
                    reader         2             0.7047
                    reader         3             0.6944
                    reader         4             <.0001
                    reader         5             <.0001
                    reader         6             0.5430
                    reader         7             0.5538
                    reader         8             <.0001
                    reader         9             0.4328
                    reader         10            <.0001
                    reader         11            0.8324
                    reader         12            <.0001
                    reader         13            0.8042
                    reader         14            0.3255
                    reader         15            0.8834
                    reader         16            0.6470
                    reader         17            <.0001
                    reader         18            0.2668
                    reader         19             .
                    order          1             0.4620
                    order          2              .
                    reader*paper   0    0        0.9040
                    reader*paper   0    1         .
                    reader*paper   1    0        0.9453
                    reader*paper   1    1         .
                    reader*paper   2    0        0.9580
                    reader*paper   2    1         .
                    reader*paper   3    0        0.9386
                    reader*paper   3    1         .
                    reader*paper   4    0         .
                    reader*paper   4    1         .
                    reader*paper   5    0         .
```

```
            reader*paper   5    1          .
            reader*paper   6    0        0.9554
            reader*paper   6    1          .
            reader*paper   7    0        0.5892
            reader*paper   7    1          .
            reader*paper   8    0          .
            reader*paper   8    1          .
            reader*paper   9    0        0.9999
            reader*paper   9    1          .
            reader*paper   10   0          .
            reader*paper   10   1          .
            reader*paper   11   0        0.9504
            reader*paper   11   1          .
            reader*paper   12   0          .
            reader*paper   12   1          .
            reader*paper   13   0        0.6661
            reader*paper   13   1          .


                     Analysis Of Parameter Estimates

                                     Standard   Wald 95% Confidence   Chi-
Parameter            DF   Estimate    Error          Limits          Square

reader*paper  14  0   1    -1.6179    1.3625   -4.2884    1.0526     1.41
reader*paper  14  1   0     0.0000    0.0000    0.0000    0.0000      .
reader*paper  15  0   1    -0.1415    1.1067   -2.3105    2.0276     0.02
reader*paper  15  1   0     0.0000    0.0000    0.0000    0.0000      .
reader*paper  16  0   1     0.8234    1.1064   -1.3451    2.9919     0.55
reader*paper  16  1   0     0.0000    0.0000    0.0000    0.0000      .
reader*paper  17  0   0   -24.2169    0.0000  -24.2169  -24.2169      .
reader*paper  17  1   0     0.0000    0.0000    0.0000    0.0000      .
reader*paper  18  0   1     0.3821    1.6211   -2.7952    3.5593     0.06
reader*paper  18  1   0     0.0000    0.0000    0.0000    0.0000      .
reader*paper  19  0   0     0.0000    0.0000    0.0000    0.0000      .
reader*paper  19  1   0     0.0000    0.0000    0.0000    0.0000      .
Scale                 0     1.0000    0.0000    1.0000    1.0000


                     Analysis Of Parameter Estimates

           Parameter                 Pr > ChiSq

           reader*paper   14   0        0.2351
           reader*paper   14   1          .
           reader*paper   15   0        0.8983
           reader*paper   15   1          .
```

```
          reader*paper   16   0        0.4568
          reader*paper   16   1          .
          reader*paper   17   0          .
          reader*paper   17   1          .
          reader*paper   18   0        0.8137
          reader*paper   18   1          .
          reader*paper   19   0          .
          reader*paper   19   1          .
          Scale

NOTE: The scale parameter was held fixed.


                        GLIMMIX - SUMMARY

                       The Mixed Procedure

                             Model Information

        Data Set                     WORK._DS
        Dependent Variable           _z
        Weight Variable              _w
        Covariance Structure         Variance Components
        Estimation Method            REML
        Residual Variance Method     Profile
        Fixed Effects SE Method      Model-Based
        Degrees of Freedom Method    Satterthwaite


                          Class Level Information

          Class     Levels    Values

          reader        20    0 1 2 3 4 5 6 7 8 9 10 11 12
                              13 14 15 16 17 18 19
          paper          2    0 1
          order          2    1 2


                               Dimensions

                 Covariance Parameters              3
                 Columns in X                       5
                 Columns in Z                      60
                 Subjects                           1
                 Max Obs Per Subject               80
                 Observations Used                 80
                 Observations Not Used               0
                 Total Observations                 80


                           Parameter Search

  CovP1   CovP2     CovP3  Variance       Res Log Like  -2 Res Log Like

0.06873       0    0.7951    0.7951          -115.1392         230.2785


                          Iteration History

    Iteration    Evaluations    -2 Res Log Like       Criterion

            1              1       230.27845497      0.00000000


                     Convergence criteria met.


                          GLIMMIX - SUMMARY

                        The Mixed Procedure

                       Covariance Parameter
                             Estimates

                     Cov Parm          Estimate

                     reader             0.06873
                     reader*paper             0
                     Residual            0.7951

                          Fit Statistics

                 -2 Res Log Likelihood          230.3
                 AIC (smaller is better)        234.3
                 AICC (smaller is better)       234.4
                 BIC (smaller is better)        236.3

                  PARMS Model Likelihood Ratio Test

                    DF    Chi-Square        Pr > ChiSq

                     1          0.00            1.0000
```

```
                      Solution for Fixed Effects

                                          Standard
Effect      paper  order  Estimate         Error      DF  t Value  Pr > |t|

Intercept                   2.8321        0.2003    70.3    14.14    <.0001
paper       0              -0.5636        0.2137    59.9    -2.64    0.0106
paper       1                    0             .       .        .         .
order              1        0.1815        0.2084    60.1     0.87    0.3874
order              2             0             .       .        .         .


                     Type 3 Tests of Fixed Effects

                              Num       Den
            Effect             DF        DF    F Value     Pr > F

            paper               1      59.9       6.95     0.0106
            order               1      60.1       0.76     0.3874


                          Least Squares Means

                                      Standard
Effect     paper    Estimate             Error      DF      t Value    Pr > |t|

paper      0          2.3592            0.1452    36.2        16.24      <.0001
paper      1          2.9228            0.1785      60        16.37      <.0001

                   Differences of Least Squares Means

                                        Standard
Effect   paper   _paper   Estimate         Error     DF  t Value  Pr > |t|

paper    0       1         -0.5636        0.2137   59.9    -2.64    0.0106

                        GLIMMIX Model Statistics

               Description                           Value

               Deviance                            70.1043
               Scaled Deviance                     88.1676
               Pearson Chi-Square                  57.6263
               Scaled Pearson Chi-Square           72.4744
               Extra-Dispersion Scale               0.7951
```

Bibliography

[1] Jacqueline Dinnes, editor. *SHPIC Report - 1997 Breast Cancer*. Scottish Health Purchasing Information Centre, 1997.

[2] G.A. Bjarnason. Menstrual cycle chronobiology: is it important in breast cancer screening and therapy? *Lancet*, 347:345–6, 1996.

[3] V. James, J. Kearsley, T. Irving, Y. Amemiya, and D. Cookson. Using hair to screen for breast cancer. *Nature*, 398:33–4, 1999.

[4] W.J. Crisp, M.J. Higgs, W.K. Cowan, W.J. Cunliffe, J. Liston, L.G. Lunt, D.J. Peakman, and J.R. Young. Screening for breast cancer detects tumours at an earlier biological stage. *Br J Surg*, 80:863–865, 1993.

[5] Department of Health and Social Security. *Breast Cancer Screening. Report to the Health Ministers of England, Wales, Scotland and Northern Ireland by a Working Group chaired by Professor Sir Patrick Forrest.* HMSO, London, 1986.

[6] S. Shapiro. Evidence on screening for breast cancer from a randomised trial. *Cancer*, 39:2772–2782, 1977.

[7] L. Tabar, A. Gad, L.H. Holmberg, U. Ljunquist, G. Eklund, F. Pettersson, C.J.G. Fagerberg, L. Baldetorp, O. Grontoft, B. Lundstrom, Manson J.C., and N.E. Day. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet*, 1:829–832, 1985.

[8] A.L.M. Verbeek, J.H.C.L. Hendricks, R. Holland, M. Mravunac, F. Sturmans, and N.E. Day. Reduction of breast cancer mortality through mass screening with modern mammography. First results of the Nijmegen Project. *Lancet*, 1:1222–1224, 1984.

[9] H.J.A. Colette, N.E. Day, J.J. Rombach, and F. De Waard. Evaluation of screening for breast cancer in a non-randomised study (the DOM project) by means of a case control study. *Lancet*, 1:1224–1226, 1984.

[10] M.M. Roberts, F.E. Alexander, T.J. Anderson, A.P.M. Forrest, W. Hepburn, A. Huggins, A.E. Kirkpatrick, J. Lamb, W. Lutz, and B.B. Muir. The Edinburgh randomised trial of screening for breast cancer: description of method. *Br J Cancer*, 50:1–6, 1984.

[11] UK Trial of Early Detection of Breast Cancer Group. Trial of early detection of breast cancer: description of method. *Br J Cancer*, 44:618–627, 1981.

[12] National Audit Office, editor. *Report by the Controller and Auditor General. Cervical and Breast Screening in England.* HMSO, London, 1992.

[13] S. Shapiro, E.A. Coleman, M. Broeders, M. Codd, H. de Koning, J. Fracheboud, S. Moss, E. Paci, S. Stachenko, and R. Ballard-Barbash. Breast cancer screening programmes in 22 countries: current policies, administration and guidelines. *Int J Epidemiol*, 27:735–742, 1998.

[14] D. Horton Taylor, K. McPherson, S. Parbhoo, and N. Perry. Response of women aged 65-74 to invitation for screening for breast cancer by mammography: a pilot study in London, UK. *J Epidemiol Community Health*, 50:77–80, 1996.

[15] Julietta Patnick, editor. *National Health Service Breast Screening Programme Review.* HMSO, 1999.

[16] G. Rubin, L. Garvican, and S. Moss. Routine invitation of women aged 65-69 for breast cancer screening: results of first year of pilot study. *BMJ*, 317:388–9, 1998.

[17] Julietta Patnick, editor. *National Health Service Breast Screening Programme Review.* HMSO, 2000.

[18] R. Boer, H. de Koning, A. Threfall, P. Warmerdam, A. Street, E. Friedman, and C. Woodman. Cost effectiveness of shortening screening interval or extending age range of NHS breast screening programme: computer simulation study. *BMJ*, 317:224–226, 1998.

[19] P.G. Peer, A.L. Verbeek, H. Straatman, J.H. Hendriks, and R. Holland. Age-specific sensitivities of mammographic screening for breast cancer. *Breast Cancer Res Treat*, 38:153–60, 1996.

[20] A.M. Kavanagh, H. Michell, H. Farrugia, and G.G. Giles. Monitoring interval cancers in an Australian mammographic screening programme. *J Med Screen*, 6:139–43, 1999.

[21] K. Kerlikowske, D. Grady, S.M. Rubin, C. Sandrock, and V. Ernster. Efficacy of screening mammography: A meta-analysis. *JAMA*, 273:149–154, 1995.

[22] Organising Committee and Collaborators. Breast-cancer screening with mammography in women aged 40-49 years. *Int J Cancer*, 68:693–699, 1996.

[23] C.B. Woodman, A.G. Threlfall, C.R. Boggis, and P. Prior. Is the three year breast screening interval too long? Occurance of interval cancers in NHS breast screening programme's north western region. *BMJ*, 310:224–226, 1995.

[24] F.E. Alexander, T.J. Anderson, H.K. Brown, A.P.M. Forrest, W. Hepburn, A.E. Kirkpatrick, C. McDonald, B.B. Muir, R.J. Prescott, S.M. Shepherd, A. Smith, and J. Warner. The Edinburgh randomised trial of breast cancer screening: results after 10 year follow-up. *Br J Cancer*, 70:542–548, 1994.

[25] M. H. van den Akker-van, H. de Koning, and P. van der Maas. Reduction in breast cancer mortality due to the introduction of mass screening in The Netherlands: comparison with the United Kingdom. *J Med Screen*, 6:30–4, 1999.

[26] P.C. Gøtzsche and O. Olsen. Is screening for breast cancer with mammography justifiable? *Lancet*, 355:129–34, 2000.

[27] S.W. Duffy. Interpretation of the breast screening trials: a commentary on the recent paper by Gøtzsche and Olsen. *The Breast*, 10:209–212, 2001.

[28] H.J. de Koning. Commentary. *Lancet*, 355:80, 2000.

[29] Age Concern. Breast Cancer Awareness Month: Older women unaware of breast cancer risk (11.10.00). Press release, published on the Age Concern website; www.ageconcern.org.uk, 2000.

[30] J.C. Wells and J. Cooke. Film reading practice of UK breast screening units. *The Breast*, 5:404–409, 1996.

[31] L.J. Williams, M. Hartswood, and R.J. Prescott. Methodological issues in mammography double reading studies. *J Med Screen*, 5:202–206, 1998.

[32] E.R.E. Denton and S. Field. Just how valuable is double reporting in screening mammography. *Clin Radiol*, 52:466–468, 1997.

[33] E.L. Thurfjell, K.A. Lernevall, and A.A.S. Taube. Benefit of independent double reading in a population-based mammography screening programme. *Radiology*, 191:241–244, 1994.

[34] I. Anttinen, M. Pamilo, M. Soiva, and M. Roiha. Double reading of mammography screening films - one radiologist or two? *Clin Radiol*, 48:414–421, 1993.

[35] H.E. Deans, D. Everington, C. Cordiner, A.E. Kirkpatrick, and E. Lindsay. Scottish experience of double reading in the national breast screening programme. *The Breast*, 7:75–79, 1998.

[36] E. Thurfjell. Mammographic screening: one versus two view and independent double reading. *Acta Radiol*, 35:345–350, 1994.

[37] E.D.C. Anderson, B.B. Muir, J.S. Walsh, and A.E. Kirkpatrick. The efficacy of double reading mammograms in breast screening. *Clin Radiol*, 49:248–251, 1994.

[38] J. Feldman, R.A. Smith, R. Giusti, B. DeBuono, J.P. Fulton, and H.D. Scott. Peer review of mammography interpretations in a breast cancer screening program. *Am J Public Health*, 85:837–839, 1995.

[39] R.M.L. Warren and S.W. Duffy. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *Br J Radiol*, 68:958–962, 1995.

[40] S. Ciatto, M.R. Del Turco, D. Morrone, S. Catarzi, D. Ambrogetti, A. Cariddi, and M. Zappa. Independent double reading of screening mammograms. *J Med Screen*, 2:99–101, 1995.

[41] 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. *BMJ*, 311:1189–1193, 1995.

[42] R.G. Blanks, S.M. Moss, and M.G. Wallis. A comparison of two view and one view mammography in the detection of small invasive cancers: results from the National Health Service breast screening programme. *J Med Screen*, 3:200–203, 1996.

[43] R.G. Blanks, M.G. Wallis, and S.M. Moss. A comparison of cancer detection rates achieved by breast cancer screening programmed by number of readers, for one and two view mammography; results from the UK National Health Service breast screening programme. *J Med Screen*, 5:195–201, 1998.

[44] R.G. Blanks, R.M. Given-Wilson, and S.M. Moss. Efficiency of cancer detection during routine repeat (incident) mammographic screening: two versus one view mammography. *J Med Screen*, 5:141–5, 1998.

[45] K.C. Young, M.G. Wallis, R.G. Blanks, and S.M. Moss. Influence of number of views and mammographic film density on the detection of invasive cancers: results from the NHS Breast Screening Programme. *Br J Radiol*, 70:482–488, 1997.

[46] R.G. Blanks, S.M. Moss, and M.G. Wallis. Use of two view mammography compared with one view in the detection of small invasive cancers: further results from the National Health Service breast screening programme. *J Med Screen*, 4:98–101, 1997.

[47] S. Bryan, J. Brown, and R. Warren. Mammography screening: an incremental cost effectiveness analysis of two view versus one view procedures in London. *J Epidemiol Community Health*, 49:70–78, 1995.

[48] E.A. Sickles. Findings at mammographic screening on only one standard projection: outcomes analysis. *Radiology*, 208:471–475, 1998.

[49] R.M. Warren, S.W. Duffy, and S. Bashir. The value of the second view in screening mammography. *Br J Radiol*, 69:105–108, 1996.

[50] R. Pauli, S. Hammond, J. Cooke, and J. Ansell. Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening. *J Med Screen*, 3:18–22, 1996.

[51] L.W. Bassett, A.J. Hollatz-Brown, R. Bastani, J.G. Pearce, K. Hirji, and L. Chen. Effects of a program to train radiologic technologists to identify abnormalities on mammograms. *Radiology*, 194:189–192, 1995.

[52] D. Asbury, C.R.M. Boggis, D. Sheals, A.G. Threlfall, and C.B.J. Woodman. NHS breast screening programme: is the high incidence of interval cancers inevitable? *BMJ*, 313:1369–1370, 1996.

[53] P.A. Sylvester, M.N. Vipond, E. Kutt, J.D. Davies, A.J Webb, and J.R. Farndon. A comparative audit of prevalent, incident and interval cancers in the Avon breast screening programme. *Ann R Coll Surg Engl*, 79:272–5, 1997.

[54] J.S. Michaelson, E. Halpern, and D.B. Kopans. Breast cancer: computer simulation method for estimating optimal intervals for screening. *Radiology*, 212:551–560, 1999.

[55] A. Salomon. Beiträge zur pathologie und klinik der mammacarcinome. *Arch Klin Chir*, 101:573–668, 1913.

[56] J.G. Elmore, C.K. Wells, C.H. Lee, D.H. Howard, and A.R. Feinstein. Variability in radiologists' interpretation of mammograms. *N Engl J Med*, 331:1493–1499, 1994.

[57] S.M. Williams, T.C.A. Doyle, S. Chartres, A.K. Richardson, and J.M. Elwood. Impact of independent double reading of mammograms from the inception of a population-based breast cancer screening programme. *The Breast*, 4:282–288, 1995.

[58] D.M. Parham, M. Creagh-Barry, R.A. Hill, and D.S. Nicholas. Observer variation in the grading of screen detected mammographic abnormalities. 1. Assessment of reproducibility. *The Breast*, 5:422–424, 1996.

[59] C.A. Beam, D.C. Sullivan, and P.M. Layde. Effect of human variability on independent double reading in screening mammography. *Acad Radiol*, 3:891–897, 1996.

[60] D.B. Fogel, E.C. Wasson 3rd, E.M. Boughton, and V.W. Porto. Evolving artificial neural networks for screening features from mammograms. *Artif Intell Med*, 14:317–26, 1998.

[61] Z. Huo, M.L. Giger, and C.E. Metz. Effect of dominant features on neural network performance in the classification of mammgraphic lesions. *Phys Med Biol*, 44:2579–95, 1999.

[62] W. Qian, L. Li, L. Clarke, R.A. Clark, and J. Thomas. Digital mammography: comparison of adaptive and nonadaptive CAD methods for mass detection. *Acad Radiol*, 6:471–80, 1999.

[63] I. Leichter, R. Lederman, P. Bamberger, B. Novak, S. Fields, and S.S. Buchbinder. The use of an interactive software program for quantitative characterization of microcalcifications on digitized film-screen mammograms. *Invest Radiol*, 34:394–400, 1999.

[64] S. Kheddache and H. Kvist. Digital mammography using storage phosphor plate technique - optimizing image processing parameters for the visibility of lesions and anatomy. *Eur J Radiol*, 24:237–44, 1997.

[65] E. Thurfjell, M. Gelig Thurfjell, E. Egge, and N. Bjurstam. Sensitivity and specificity of computer-aided breast cancer detection in mammography screening. *Acta Radiol*, 39:384–388, 1998.

[66] S. Nawano, K. Murakami, N. Moriyama, H. Kobatake, H. Takeo, and K. Shimura. Computer-aided diagnosis in full digital mammography. *Invest Radiol*, 34:310–6, 1999.

[67] G.M. te Brake, N. Karssemeijer, and J.H. Hendricks. Automated detection of breast carcinomas not detected in a screening program. *Radiology*, 207:465–1, 1998.

[68] Y. Jiang, R.M. Nishikawa, R.A. Schmidt, C.E. Metz, M.L. Giger, and K. Doi. Improving breast cancer diagnosis with computer-aided diagnosis. *Acad Radiol*, 6:22–23, 1999.

[69] H.P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. *Radiology*, 212:817–27, 1999.

[70] N.C. Hambly, L. Miller, H.T. MacGillivray, J.T. Herd, and W.A. Cormack. Precision astrometry with SuperCOSMOS. *Monthly notices of the Royal Astronomical Society*, 298:897–904, 1998.

[71] L. Miller and N. Ramsay. The detection of malignant masses by non-linear multiscale analysis. In *Digital Mammography*, pages 335–340. Elsevier Science, 1996.

[72] A. Hume, P. Thanisch, M. Hartswood, and R. Procter. On the evaluation of microcalcification detection algorithms. In *Digital Mammography*. Elsevier Science, 1996.

[73] N. Karssemeijer and J.H.C.L. Hendricks. Computer-assisted reading of mammograms. *Eur Radiol*, 7:743–748, 1997.

[74] T. Parr, R. Zwiggelaar, S. Astley, C. Boggis, and C. Taylor. Comparison of methods for combining evidence for spiculated lesions. In *Digital Mammography*, pages 71–78. Kluwer Academic Publishers, 1998.

[75] S. Heddle, A.C. Hume, and A.E.K. Kirkpatrick. Evaluation of a prompting system using interval cancers. In *Digital Mammography*, pages 355–358. Kluwer Academic Publishers, 1998.

[76] A.A. Duncan and M.G. Wallis. Classifying interval cancers. *Clin Radiol*, 50:774–777, 1995.

[77] W. Simpson, F. Neilson, J.R. Young, and the Northern Region Breast Screening Radiology Audit Group. The identification of false negatives in a population of interval cancers: a method for audit of screening mammography. *The Breast*, 4:183–188, 1995.

[78] H.C. Burrell, D.M. Sibberington, A.R. Wilson, S.E. Pinder, A.J. Evans, L.J. Yeoman, et al. Screening interval breast cancers: mammographic features and prognosis factors. *Radiology*, 199:811–7, 1996.

[79] R.D. Jones, L. McLean, J.R. Young, W. Simpson, and F. Neilson. Proportion of cancers detected at the first incident screen which were false negative at the prevalent screen. *The Breast*, 5:339–343, 1996.

[80] P.A. Sylvester, E. Kutt, A. Baird, M.N. Vipond, A.J Webb, and J.R. Farndon. Rate and classification of interval cancers in the breast screening programme. *Ann R Coll Surg Engl*, 79:276–7, 1997.

[81] M. Hartswood, R. Procter, L. Williams, R. Prescott, and P. Dixon. Drawing the line between perception and interpretation in computer-aided mammography. In

First International Conference on Allocation of Functions, pages 275–291, Galway, 1997. IEA Press.

[82] S. Ciatto, M. Rosselli Del Turco, and M. Zappa. The detectability of breast cancer by screening mammography. *Br J Cancer*, 71:337–339, 1995.

[83] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[84] M. Bland. *An introduction to medical statistics.* Oxford Medical Publications, Oxford, first edition, 1987.

[85] C.E. Metz. ROC methodology in radiologic imaging. *Invest Radiol*, 21:720–733, 1986.

[86] A.N. Angelos Tosteson and C.B. Begg. A general regression methodology for ROC curve estimation. *Med Decis Making*, 8:204–215, 1988.

[87] Ian Hutt. *The Computer-Aided Detection of Abnormalities in Digital Mammograms.* PhD thesis, University of Manchester, 1996.

[88] M. Hartswood, R. Procter, L. Williams, and R. Prescott. Subjective responses to prompting in screening mammography. In *Medical Image Understanding and Analysis*, pages 205–208, Oxford, 1997.

[89] P. McCullagh and J.A. Nelder. *Generalised Linear Models.* Chapman and Hall, London, 1983.

[90] H. Brown and R. Prescott. *Applied Mixed Models in Medicine.* John Wiley and Sons, Chichester, 1999.

[91] Mark Hartswood. *Human Factors in Computer-Aided Mammography.* PhD thesis, University of Edinburgh, 2000.

[92] A.N. Oppenheim. *Questionnaire Design and Attitude Measurement.* Heinemann, London, 1973.

[93] M. Hartswood, R. Procter, and L.J. Williams. Prompting in mammography: Computer-aided detection or computer-aided diagnosis? In *Medical Image Understanding and Analysis*, pages 101–104, Leeds, 1998.

[94] L.J. Williams, R.J. Prescott, and M. Hartswood. Computer-aided cancer detection in the UK breast screening programme. In *Digital Mammography*, pages 359–362. Kluwer Academic Publishers, 1998.

[95] M. Hartswood, R. Proctor, and L.J. Williams. Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography. In *Digital Mammography*, pages 363–370. Kluwer Academic Publishers, 1998.

[96] W.H. Press et al. *Numerical recipes in C : the art of scientific computing.* Cambridge University Press, Cambridge, 1992. Ch. 7, p. 282.

[97] SAS Institute Inc. website: www.sas.co./service/techsup/faq/stat_macro/glimacr.html, 2001.

[98] H. Sittek, C. Perlet, R. Helmberger, E. Linsmeier, M. Kessler, and M. Reiser. Computer-aided diagnosis in routine mammography. *Radiologe*, 38:848–852, 1998.

[99] L.J.W. Burhenne, S.A. Wood, C.J. D'Orsi, S.A. Feig, D.B. Kopans, K.F. O'Shaughnessy, E.A. Sickles, L. Tabar, C.J. Vyborny, and R.A. Castellino. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology*, 215:554–562, 2000.

[100] R.L. Birdwell, D.M. Ikeda, K.F. O'Shaughnessy, and E.A. Sickles. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology*, 219:192–202, 2001.

[101] L. Garvican and S. Field. A pilot evaluation of the R2 image checker system and users' response in the detection of interval breast cancers on previous screening films. *Clin Radiol*, 56:833–837, 2001.

[102] Health Technology Assessment Programme. Impact of computer-placed prompts on sensitivity and specificity with different groups of mammographic film readers. On-going trial, detailed on the NCCHTA website (www.ncchta.org), 2002.

[103] A. Malich, C. Marx, M. Facius, T. Boehm, M. Fleck, and W.A. Kaiser. Tumour detection rate of a new commercially available computer-aided detection system. *Eur Radiol*, 11:2454–2459, 2001.

[104] B. Jouan. Digital mammography performed with computed radiography technology. *Eu J Radiol*, 31:18–24, 1999.

[105] T. Yamada and Y. Muramatsu. Computed radiography for breast-cancer. *Jpn J Clin Oncol*, 20:164–168, 1990.

[106] K.A. Fetterly and N.J. Hangiandreou. Image quality evaluation of a desktop computed radiography system. *Med Phys*, 27:2669–2679, 2000.

[107] R.F. Brem and J.M. Schoonjans. Radiologist detection of microcalcifications with and without computer-aided detectiom: a comparative study. *Clin Radiol*, 56:150–154, 2001.

[108] S. Paquerault, N. Petrick, H.P. Chan, B. Sahiner, and M.A. Helvie. Improvement of computerized mass detection on mammograms: fusion of two-view information. *Med Phys*, 29:238–247, 2002.

[109] L. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and M. Gurcan. Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses. *Med Phys*, 28:2309–2317, 2001.

[110] J. Roehrig and R.A. Castellino. The promise of computer-aided detection in digital mammography. *Eur J Radiol*, 31:35–39, 1999.

[111] K. Perisinakis, J. Damilakis, E. Kontogiannis, and N. Gourtsoyiannis. Film-screen magnification versus electronic magnification and enhancement of digitized contact mammograms in the assessment of subtle microcalcifications. *Invest Radiol*, 36:726–733, 2001.

[112] H.E. Reynolds. Advances in breast imaging. *Hematology - Oncology Clinics of North America*, 13:333–48, 1999.

[113] G.S. Maitz, T.S. Chang, J.H. Sumkin, P.W. Wintz, C.M. Johns, M. Ganott, et al. Preliminary clinical evaluation of a high-resolution telemammography system. *Invest Radiol*, 32:236–40, 1997.

[114] A.L. Rafanan, P. Kakulavar, J. Perl, J.C. Andrefsky, D.R. Nelson, and A.C. Arroliga. Head computed tomography in medical intensive care unit patients: clinical indications. *Crit Care Med*, 28:1306–1309, 2000.

[115] J.R. Heyne, J. Sehner, R. Neumann, B. Werner, R. Adler, M. Freesmeyer, and W.A. Kaiser. Reduction of radiation exposure by using storage phosphor radiography on pelvis and lumbar spine. *Rofo Fortschr Geb Rontgenstr Neuen Bildgeb Verfahr*, 174:104–111, 2002.

[116] S. Iwano, T. Ishigaki, K. Shimamoto, K. Inamura, T. Maeda, M. Ikeda, T. Ishiguchi, and T. Kozuka. Detection of subtle pulmonary disease on cr chest images: monochromatic crt monitor vs color crt monitor. *Eu Radiol*, 11:59–64, 2001.

[117] A. Leon, G. Verdu, M.D. Cuevas, M.D. Salas, J.I. Villaescusa, and F. Bueno. Study of radiation induced cancers in a breast screening programme. *Radiat Prot Dosim*, 93:19–30, 2001.

[118] H. Jung. Estimates of benefits versus radiation risks from mammographic screening. *Radiologe*, 41:385–395, 2001.

[119] J. Myles, S. Duffy, R. Nixon, and et al. Initial results of a study into the effectiveness of breast cancer screening in a population identified to be at high risk. *Rev Epidemiol Sante*, 49:471–475, 2001.

[120] L.G. Keith, J.J. Oleszczuk, and M. Laguens. Circadian rhythm chaos: a new breast cancer marker. *Int J Fertil Women M*, 46:238–247, 2001.

[121] K. Pachmann, P. Heiss, U. Demel, and G. Tilz. Detection and quantification of small numbers of circulating tumour cells in peripheral blood using laser scanning cytometry (LSC(R)). *Clin Chem Lab Med*, 39:811–817, 2001.

[122] M.L. Giger, K. Doi, and H. Macmahon. Computer-aided detection of lung nodules by use of a filtering technique. *Med Phys*, 13:596, 1986.

[123] H.P. Chan, K.N. Doi, C.J. Vyborny, K.L. Lam, and R.A. Schmidt. Computer-aided detection of microcalcifications in mammograms - methodology and preliminary clinical-study. *Invest Radiol*, 23:664–671, 1988.

[124] National Cervical Cancer Coalition. Worldwide cervical cancer issues. Website, www.nccc-online.org, 2002.

[125] K.E. Hartmann, K. Nanda, S. Hall, and E. Myers. Technological advances for evaluation of cervical cytology: is newer better? *Obstet Gynecol Surv*, 56:765–774, 2001.

[126] M.E. Boon and L.P. Kok. Neural-network processing can provide means to catch errors slip through human screening of Pap smears. *Diagn Cytopathol*, 9:411–416, 1993.

[127] H. Doornewaard, Y.T. van der Schuow, Y. van der Graaf, A.B. Bos, J.D.F. Habbema, and J.G. van der Tweel. The diagnostic value of computer-assisted primary cervical smear screening: a longitundinal cohort study. *Mod Pathol*, 12:995–1000, 1999.

[128] G.M. Troni, I. Cipparrone, M.P. Cariaggi, S. Ciatto, G. Miccinesi, M. Zappa, and M. Confortini. Detection of false-negative Pap smears using the PAPNET system. *Tumori*, 86:455–457, 2000.

[129] A. Farnsworth, F.M. Chambers, and C.S. Goldschmidt. Evaluation of the PAPNET system in a general pathology department. *Medical Journal of Australia*, 165:429–431, 1996.

[130] J.W. Bishop, R.H. Kaufman, and D.A. Taylor. Multicentre comparison of manual and automated screening of AutoCyte gynecologic preparations. *Acta Cytologica*, 43:34–38, 1999.

[131] T.J. Colgan, S.F. Patten, and J.S.J. Lee. A clinical trial of the AutoPap AP300 QC system for quality control of cervicovaginal cytology in the clinical laboratory. *Acta Cytologica*, 39:1191–1198, 1995.

[132] D.C. Wilbur and M.K. Norton. The primary screening clinical trials of the TriPath AutoPap(R) system. *Epidemiology*, 13:S30–S33, 2002.

[133] T. Shimada, N. Kodama, H. Satoh, K. Hiwatashi, T. Ishida, Y. Nishimura, and I. Fukumoto. Proposal of a nodule density-enhancing filter for plain chest radiographs on the basis of the thoracic wall outline detected by Hough transformation. *IEICE T Inf Syst*, E85D:88–95, 2002.

[134] Y.H. Chou, C.M. Tiu, G.S. Hung, S.C. Wu, T.Y. Chang, and H.K. Chiang. Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis. *Ultrasound Med Biol*, 27:1493–1498, 2001.

[135] R.M. Summers. Challenges for computer-aided diagnostics for ct colonography. *Abdom Imaging*, 27:268–274, 2002.

[136] P. Celka and P. Colditz. A computer-aided detection of eeg seizures in infants: a singular spectrum approach and performance comparison. *IEEE Trans Biomed Eng*, 49:455–462, 2002.

[137] H. Doornewaard, Y.T. van der Schuow, and Y. van der Graaf. Reproductibility in double scanning of cervical smears with the papnet system. *Acta Cytologica*, 44:604–610, 2000.

[138] T.W. Freer and M.J. Ulissey. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast centre. *Radiology*, 220:781–786, 2001.

Methodological issues in mammography double reading studies

Linda J Williams, Mark Hartswood, Robin J Prescott

Abstract

***Objectives*—An examination of the methods used in assessing cancer detection rates in double reading studies to obtain a clear interpretation of the disparate results from differing studies.**

***Setting*—National breast screening programmes.**

***Methods*—Critical appraisal of methodologies used in the comparison of cancer detection rates with single or double reading.**

***Results*—The reported improvement in cancer detection rate with double reading varies greatly between studies, depending upon whether the study is blinded and the statistic used. A method of calculating the increase in cancer detection rate due to a second reader is proposed.**

(*J Med Screen* 1998;5:202–206)

Keywords: mammography; double reading; statistical measures

Many studies indicate that the practice of double reading, within a breast screening programme, offers an opportunity for screening centres to improve their cancer detection rates. Beam and Sullivan[1] imply that we should know this from first principles—that double reading, in its simplest form, should be at least as sensitive as the most sensitive participating radiologist. What it is important to discover is the actual quantitative increase in the cancer detection rate due to double reading and the associated effect that this practice will have on specificity. Difficulties in achieving this arise from the relatively small sample size in some studies and the disparate methodologies used.

In England and Wales the Forrest report[2] recommended that single reading be adopted as the standard practice, although in Scotland double reading is the standard. Both of these "standards" have evidence to support their implementation as the actual improvement due to double reading quoted in the literature can vary dramatically from paper to paper. The actual improvements in cancer detection reported range from 1.5%[3] to 15%,[4] with reported changes in recall rate varying from a decrease of 45%[5] to an increase of 37%.[6] Although the lowest figure for cancer detection improvement would suggest that the small improvement in sensitivity due to a second reader is an inefficient use of finite resources, the higher figure could be used to justify double reading. The wide variety of reported gains makes an accurate evaluation of the possible improvement in detection rate difficult. It may be that some of this variation can be explained by the different double reading procedures used, which also vary from study to study. The picture is then further complicated by the fact that these evaluations are mostly done in clinical context by looking back over a period of time and analysing the results retrospectively. Hence, no experimental procedure is followed to control potential biases and errors.

This paper presents a critical analysis of the methodology used in the various studies, and tries to explain some of the sources of variation in the reported improvements due to double reading. In particular, we look at the use of the "mean second screener contribution" as the statistic used to make comparisons between studies with disparate methodologies, with a possible alternative suggested for such cases. Other phenomena, such as order effect, which are largely undiscussed in other papers, are also examined.

Methods

Two methods of estimating the relative improvement in cancer detection rate due to a second reader are presented. The first of these is a widely used, "standard" method. The second method is an alternative, which we argue is to be preferred. An example illustrating and contrasting the results for these methods is then presented. Finally, papers which evaluate double reading were extracted using the following search strategy, and the findings were tabulated. These references were selected from a reference search of the ISI database using the key phrase "double reading". The subsequent listing was further refined by excluding non-screening experiments, foreign language papers, and letters, and by focusing on double versus single reading, rather than including studies such as two view versus one view experiments. This provided papers by Thurfjell,[4] Deans,[6] Warren,[7] and Anderson.[8] Examination of the references in these articles led to the Antinnen[5] and Ciatto[9] papers. The Beam and Sullivan[1] letter was a response to the Thurfjell paper. A citation search was also performed on the above papers, which led to the Denton paper.[3] Although this may not be an exhaustive list, it is non-selective in that we have included every example that was reported by the search.

THE MEAN SECOND SCREENER CONTRIBUTION

One of the more commonly used statistics for measuring the improvement in cancer detection gained by the addition of a second reader is the mean second screener contribution (MSSC).[4 10] In effect, this is the average

Department of Public Health Sciences, University of Edinburgh
L J Williams, *research associate, statistician*
R J Prescott, *director of medical statistics unit, statistician*

Department of Computer Science, University of Edinburgh
M Hartswood, *research associate, human factors*

Correspondence to:
Linda J Williams, Medical Statistics Unit, The University of Edinburgh, Medical School, Teviot Place, Edinburgh EH8 9AG, UK.

Accepted for publication 7 October 1998

Table 1 Cancers reported by each radiologist

	R_2+	R_2−	
R_1+	*a*	*b*	*a+b*
R_1−	*c*	*d*	
			n

R_1 = first reader; R_2 = second reader; + = cancers detected; − = cancers missed.

Table 2 Second reader not blinded to first reader's decision

	R_2+	R_2−	
R_1+	170	2	172
R_1−	19	0	19
	189	2	191

MSSC = 5.82%.

Table 3 Second reader blinded to first reader's decision

	R_2+	R_2−	
R_1+	153	19	172
R_1−	19	0	19
	172	19	191

MSSC = 11.05%.

number of cancers detected by only one of the radiologists, divided by the average number of cancers found by each radiologist. The number of cancers reported by each radiologist is usually summarised as shown in table 1.

From this, the mean second screener contribution is defined as:

$$\text{MSSC} = \frac{(b+c)/2}{((a+b)+(a+c))/2}$$

which is generally presented as:

$$\text{MSSC} = \frac{(b+c)/2}{a+(b+c)/2}$$

This can be rearranged to give the slightly simpler form of:

$$\text{MSSC} = \frac{b+c}{2a+b+c}$$

THE EFFECT OF NON-BLINDING ON THE MSSC

The above statistic gives an efficient estimate of the second screener effect when the study is blinded and there is worst case recall. However, full blinding is unusual, and in its absence there is always the possibility that the second reader may be influenced by the decision of the first reader. In particular, in the absence of blinding, it is plausible that the second radiologist may "find" some cases which have been prompted by the first radiologist that s/he would not have discovered had the readings been independent. Hence, in the notation of table 1, *a* may be relatively increased and *b* decreased when compared with blind reading. The consequence of this is that failure to blind the second radiologist may result in a lower MSSC than had they been blinded, even though the number of cancers detected ($(a+b) + c$) remains the same.

AN ALTERNATIVE MEASURE

A measure which we believe to be preferable, and which is sometimes reported, is the proportional increase in cancer detection rate due to the second reader: $c/(a+b)$. This measure is not influenced by the relative sizes of *a* and *b* (R_1+R_2+ and R_1+R_2- respectively), just the number of cancers discovered by R_1 and the additional cancers discovered by R_2. Thus this avoids the problems due to blinding/not blinding posed by the MSSC.

The alternative measure also has the property that it is easier to calculate both the point estimate (the actual increase) and the standard error of the increase. Whereas the MSSC has a complex formula for the standard error,[4] the SE of $c/(a+b)$ is given by a logarithmic transformation, as:

$$\text{SE}\left(\log_e \frac{c}{a+b}\right) = \sqrt{\frac{1}{a+b}+\frac{1}{c}}$$

This measure and the MSSC have been presented in terms of determining the increase in cancer detection, but they also apply in precisely the same way to the determination of increases in recall rate.

EXAMPLE

If we assume that both readers have the same chance of detecting cancers missed by their counterpart when they read blind, then we potentially have the situation illustrated below. In the first case, the second reader is not blinded, and can therefore be influenced by the decisions of the first reader. In the second, the second reader has no information as to the first reader's decision (tables 2 and 3)

If we use the alternative measure $c/(a+b)$, then the improvement due to the second reader is 11.05% in both cases (the same as the

ble 4 Main procedural differences between the studies examined

	Sample size	*Age range*	*Number of radiologists*	*Number of cancers*	*Randomised trial*	*Recall criteria*	*Blinded*
nton[3]	62.5% of 36 320	>50	2	225	No	Worst case	Yes
urfjell[4 10]	11 343	40–74	2	76	Unknown	Discussion of flagged cases	Yes
tinnen[5]	15 547	50–59	4	68	No - 2 always first, 2 always second	Flagged cases reviewed by both	Yes
ans[6]	Not given	>50	35 over 4 years*	2473	No - varies across clinics	Worst case (except Glasgow, third reader)	No
rren[7]	33 734	>50	3	269	No - by chance	Consensus or review by senior radiologist	Yes (on initial reading)
derson[8]	28 170	>50	3	191	No - first in usually first reader	Worst case	No
itto[9]	18 817	50–70	4	125	Not clear	Worst case	Yes

ll Scottish breast screening service radiologists.

Table 5 Methods of calculating the improvement in cancer detection due to double reading

	Method of calculating improvement	*Stated improvement (%)*	*MSSC (%)*	*c/a+b (%)*	*95% CI of (c/a+b) (%)*
Denton[3]	Double reporting - single reporting (R_1 or R_2)	1.5–4.2	Not calculable*	Not calculable	Not calculable
Thurfjell[4 10]	MSSC	15	15.2	8.6	(3.7 to 19.7)
Antinnen[5]	MSSC	8.9	8.8	6.25	(2.3 to 17.2)
Deans[6]	*c/(a+b+c)*	10.5 (12.3)†	6.4 (7.6)	11.7 (14.1)	(10.3 to 13.3) ((12.0 to 16.6))
Warren[7]‡	*c/a+b*	14	7.1	13.75	(9.6 to 19.8)
Anderson[8]	*(b+c)/(a+b+c)*	10.4	5.8	Not calculable	Not calculable
Ciatto[9]	MSSC	4.6	4.6	7.7	(3.9 to 15.3)

* First and second reader not identified.

† The estimates in parentheses are the figures for Scotland excluding Glasgow.

‡ The results quoted in this article are R_1 versus post-discussion and cannot strictly be compared with the other articles.

estimate obtained with the MSSC with equal numbers in R_1+R_2- and R_1-R_2+), with a 95% confidence interval of (6.9% to 17.7%).

Results

Seven relevant papers were identified by our search strategy. Table 4 illustrates the main procedural differences between the various studies examined. None of the studies illustrated was a randomised trial, and they generally relied on retrospective examination of the data. Thus biases could be introduced, with no way of determining where the bias lay and in which direction. Recall criteria and blinding are both clinical procedures, where blinding refers to the second reader not knowing the decision of the first reader. As will be seen in table 5 the stated improvements in cancer detection rates due to the second screener vary considerably, but how much of this variation is due to the clinical procedures and how much to the method of calculating the improvement?

Three of the studies had used the MSSC method for estimating the improvement in cancer detection rate due to the second reader, and all of these studies reported the reading to be blinded. The other four studies used a variety of methods of calculation (table 5). The reported range of improvement was from 1.5% to 15% over these studies. Use of the recommended method of calculation gives a range of 6.25% to 11.7% for the improvements in cancer detection rates (with the exclusion of the Warren paper[7] from these calculations, as their comparison was between the first radiologist and the post-discussion result, rather than between first reader and second reader). We note that the confidence intervals for these improvements are rather wide, reflecting the relatively small numbers of patients with cancer. A weighted average of the improvements due to the second reader is 11.4% with a 95% confidence interval of 10.0% to 12.9%. This figure is highly influenced by the results of the Deans paper, which is based on 2473 cancers detected—nearly 10 times greater than any other study.

The overall benefits from an improvement in sensitivity cannot be gauged unless the corresponding impact on specificity is also reported. However, a number of papers failed to take this into consideration. Measuring changes in recall rates due to double reading depends on the recall rate for the first reader being an accurate approximation of single reading. In table 5 the reported effects of double reading on recall rates indicate that where there is a system of discussion or third reader arbitration to decide recalls an improvement in specificity is observed. However, as noted by Wells[11] and Warren,[7] a bias may be introduced because individual readers are aware that their initial recall decisions may be reversed. The changes to the recall rates quoted in the articles are usually calculated differently from the improvement in cancer detection. In the Antinnen[5] and Warren[7] papers this calculation is the difference between the recall rates before and after discussion (see table 6 for exact details). As such, it is obvious that the recall rate will decrease, whereas the methods of calculating the change in recalls that we have quoted have been the amount of additional recalls made by the second reader with respect to the first. The weighted average of the estimated increase is recall rate is 38%, with a 95% confidence interval of 36.9% to 39.2%.

Discussion

Although we have focused on blinding as a potential confounder in our search for an estimate of the improvement in cancer detection due to double reading, other factors may play a part. For example, it is entirely possible that reading order has an effect on the improvement, where the psychology of the reader

Table 6 Methods of calculating the change in recall rate due to double reading

	Method of calculating change	*Stated change (%)*	*MSSC (%)*	*c/a+b (%)*	*95% CI of (c/a+b) (%)*
Denton[3]	Recalls not mentioned	Not given	Unknown	Unknown	Unknown
Thurfjell[4 10]	Recalls not mentioned	Not given	Unknown	Unknown	Unknown
Antinnen[5]	Mean reduction of post-discussion cases per reader	−45	66.9 (41)*	46 (45.3)	(39.8 to 53.2) ((36.5 to 56.3))
Deans[6]	Reader 1 compared with double reading (figures excluding Glasgow)	+37 (4.2 raised to 6.6)	23.6	37.3†	(36.1 to 38.6)
Warren[7]	First reader recall rate compared with post-discussion recall rate	−39.1 (6.9 lowered to 4.2)	48 (44.5)‡	43.9 (16.4)	(40.8 to 47.3) ((14.7 to 18.3))
Anderson[8]	Specificity	−1.8	34.2	Not calculable	Not calculable
Ciatto[9]	MSSC	+15	15	18.7	(15.4 to 22.8)

* The figures in parentheses are the post-discussion results.

† Scotland excluding Glasgow.

‡ The figures in parentheses are the R_1 versus actual decision results.

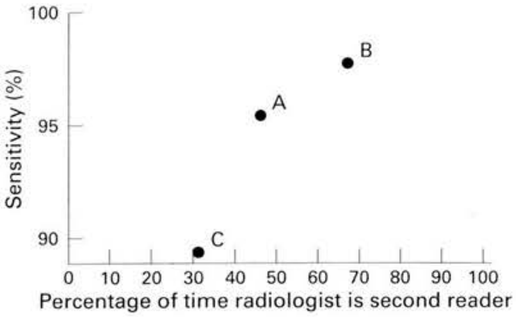


Figure 1 Relation between reading second and sensitivity.

knowing that they are first or second may influence the outcome. Unfortunately, given the nature of these studies, it becomes difficult to separate the order effect from the other effects.

A good example of this is the 1994 Anderson paper.[8] The study was retrospective and non-blinded, where the radiologists were first or second based only on who was available to read at the time. The relative number of first and second readings for the three available radiologists was significantly different (χ^2 test: $p<0.0001$). Radiologist A read second on 46% of the mammograms s/he saw, radiologist B read second on 67%, and radiologist C read second on 31%. Given that the sensitivities for these three radiologists were 95.4%, 97.8%, and 89.4% respectively (fig 1), there seems to be a strong linear relation between these two variables. However, given the nature of the data, it is also possible that the effect is due to radiologist differences.

The Ciatto study[9] (blinded, "worst case" recalls) indicates a gain for double reading which is similar to those for unblinded studies also operating a worst case recall policy. This is not what might be expected if the "blinding effect" hypothesis is accurate. In a properly randomised blinded study b and c would be approximately equal. Thus MSSC and $c/a+b$ would be roughly equal. However, there is a large difference between the cancers detected by R_1 only and those detected by R_2 only in the Ciatto study (two and nine, respectively), leading to the difference between the two metrics, as seen in table 4. In this case it seems unlikely that each radiologist was equally assigned to be first or second reader. Ciatto offers a high degree of experience as an explanation for the low improvement in cancer detection, stating that the performance of single reading is maximised, but this fails to explain the worse performance of the first reader compared with the second. It is unclear from this article whether there was true randomisation during the study. One marked difference between the Ciatto studies and the others is that only a fraction of the total clinical throughput is double read (20%), thus the main experience of the participating radiologists will be with single reading. It is an interesting possibility that where readings are predominantly double readings the performance of individual radiologists may decline while the sensitivity of the clinic as a whole is maintained. Double reading studies will still show gains in detection because of the assumption that one or other radiologist is still equivalent to a single reader. However, the gain over true single reading may only be minimal in reality.

A further issue is one of accuracy. It is necessary to quantify accurately the performance gains due to double reading to inform policy decisions effectively. Trials should be planned in advance, with firm protocols in place to limit the sources of variability and set a definite end point. Sample sizes need to be adequate; for example, to detect a 6% relative increase in cancer detection over a 5 in 1000 base rate, with 80% power requires a trial size of 114 103 women (at an agreement of 90% between radiologists). Higher agreement between radiologists will reduce the required sample size (for example, 43 881 at 98% agreement), but it would be unwise to overestimate the agreement when estimating the trial size and discover that there is insufficient power at the end of the experiment.

Conclusion

For an unambiguous answer as to how much better double reading is over single reading it is necessary to compare both practices directly. Such a study would be difficult to contrive, and probably would prove infeasible to put into practice. Thus the studies cited necessarily take a pragmatic approach—making the assumption that the first or second reader's performance is equivalent to that of a single reader when calculating performance gains. Only if this assumption is accurate will the studies reflect the true performance gain achievable due to double reading, though even this has been challenged. Furthermore, the studies cited have additional methodological problems. They were either retrospective, or were conducted—again, for pragmatic reasons—without optimal control over particular sources of error (for example, without proper randomisation of radiologists). It is also necessary to attend carefully to the impact of double reading on specificity, paying particular attention to how the decision pathway might bias any comparisons.

Methodological problems not withstanding, these studies still offer the "best evidence" currently available for performance gains due to double reading, though care is required in their interpretation. Because of variations in experimental design, close attention should be given to choosing the most appropriate statistics for reporting and comparing results. We have shown particularly that the MSSC is an inappropriate metric if comparing blinded and unblinded studies, and have proposed $c/(a+b)$ as a more accurate comparator.

In conclusion, from the papers we have examined, we estimate the mean increase in cancer detection due to double reading to be 11.4% (excluding the Warren result), and the mean increase in the recall rate to be 38%.

1 Beam CA, Sullivan DC. What are the issues in the double reading of mammograms? [letter]. *Radiology* 1994;**193**:582.

2 Working group chaired by Sir Patrick Forrest. *Breast cancer screening.* (Report to the Health Ministers of England, Wales, Scotland, and Northern Ireland.) London:HMSO, 1987.

3 Denton ERE, Field S. Just how valuable is double reporting in screening mammography. *Clin Radiol* 1997;**52**:466–8.

4 Thurfjell E, Lernevall K, Taube A. Benefit of independent double reading in a population-based mammography screening programme. *Radiology* 1994;**191**:241–4.

5 Antinnen I, Pamilo M, Soiva M, *et al.* Double reading of mammography screening films - one radiologist or two? *Clin Radiol* 1993;**48**:414–21.

6 Deans HE, Everington D, Cordiner C, *et al.* Scottish experience of double reading in the national breast screening programme. *The Breast* 1998;**7**:75–9.

7 Warren RML, Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *Br J Radiol* 1995;**68**:958–62.

8 Anderson EDC, Muir BB, Walsh JS, *et al.* The efficacy of double reading mammograms in breast screening. *Clin Radiol* 1994;**49**:248–51.

9 Ciatto S, Del Turco MR, Morrone D, *et al.* Independent double reading of screening mammograms. *J Med Screen* 1995;**2**:99–101.

10 Thurfjell E. Mammographic screening: one versus two view and independent double reading. *Acta Radiol* 1994;**35**: 345–50.

11 Wells JC, Cooke J. Film reading practice of UK breast screening units. *The Breast* 1996;**5**:404–9.

4th Meeting of the International Society for Neonatal Screening

13–16 June 1999, Stockholm, Sweden

The latest scientific achievements in the field of neonatal screening will be presented. Invitations are extended to scientists, laboratory directors, technicians, paediatricians, nurses, nutritionists, and all other professionals engaged in neonatal screening, and to relevant parent/patient organisations and commercial companies.

The topics will include blood sample collection, quality assurance, new techniques (for example, tandem mass spectrometry), evolving programmes (for example, cystic fibrosis, hearing defects), neonatal screening in developing countries, and neonatal screening in the next millennium.

Further details: Congress Secretariat: Stockholm Convention Bureau, Box 6911, 102 39 Stockholm, Sweden. Tel/fax: +46 8 736 15 00. Email: isns@stocon.se

Home page: http://www.stocon.se/isns

Drawing the line between perception and interpretation in computer-aided mammography

M. Hartswood[a], R. Procter[a], L. Williams[b], R. Prescott[b], P. Dixon[a]

[a]Department of Computer Science, Edinburgh University, Edinburgh, EH9 3JZ, Scotland.
[b]Department of Public Health Sciences, Edinburgh University, Edinburgh, EH8 9AG, Scotland.

ABSTRACT

Screening mammography calls for a combination of perceptual skills to find what may be faint and small features in a complex visual environment, and interpretive skills to rate their diagnostic significance. Evidence suggests radiologists' performance of this task can be improved by computer-aided prompting of target features.

The introduction of computer-aided mammography provides an interesting case study of 'allocation of function' issues. One is where to 'draw the line' between perception and interpretation when determining the system's functional role. Our investigations indicate that radiologists find a system which is 'perceptually acute', but 'interpretatively naive', more acceptable than predicted by earlier work. We present evidence that this is because drawing the line in this way helps radiologists to understand, and to monitor, the system's behaviour.

A second issue concerns the impact of computer-aided mammography on existing practices. Our studies reveal informal, but important, collaborative practices which help to make radiologists' work routinely available to each other. We argue that such practices must be properly understood and accommodated within computer-aided mammography if its benefits are to be fully realised.

1. Introduction

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

In the UK, women between the ages of 50 and 64 are invited to attend a clinic for screening every three years. In the UK, the rate of detection of abnormalities through screening is about 6% for women undergoing their first screening, falling to 3% for second and subsequent screenings. Currently about 0.6% of those screened are found to have malignancies. The radiologists' task is a difficult one, not least because the small number of cancers is hidden amongst a large number of normal cases. It is a task which demands a high level of perceptual and interpretative skill: under certain circumstances normal tissue can have an abnormal appearance — and vice versa.

The goal of screening is to achieve a reliable and controlled cancer detection rate. Two performance parameters are particularly important: specificity and sensitivity. A high specificity (high true positive rate) means that few women will be recalled for further tests unnecessarily; a high sensitivity (low false negative rate) means that few cancers will not be found. Achieving high specificity *and* high sensitivity is difficult.

The UK screening programme is continually investigating ways of improving detection rates: current practice involves each mammogram being 'double read' (examined independently by two radiologists) which has been shown to improve true positive rates compared with single reading (examination by one radiologist only). In the past five years, interest has grown in the possibility of developing computer-based image analysis tools which will enable a single radiologist to achieve performance equal to that achieved by double reading.

Computer-aided mammography (CAM) raises some important questions regarding the allocation of function between human and computer-based agents. We begin by reviewing allocation of function issues within the general medical application context, and then briefly outline the UK breast screening programme and the nature of reading work. We then present evidence from our investigations, and finally we discuss its implications for computer-aided mammography.

2. Allocation of Function Issues in Medical Work

The early promise that expert systems would master the intellectual aspects of medical practice (Schwartz, 1970) remain largely unfulfilled. Of the many medical decision support systems (MDSSs) implemented, few have found routine use (Forsythe, 1992a; Heathfield and Wyatt, 1993). Explanations for the failure of MDSSs fall broadly into three 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 issues: scope, role and work practice.

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 (Blois, 1980). 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 (Miller, 1994).

Role

Some MDSSs support decision-making by simply providing information that can assist physicians to reach their own conclusion, e.g., 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.

Work practices

Work practice issues in MDSS applications are inevitably multi-faceted, and problematic for designers. 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 (Kaplan, 1988), whereas MDSSs that give useful reminders or alerts have been well received (Clayton and Hripcsak, 1995).

3. Allocation of function issues in computer-aided mammography

Radiologists' expertise in reading mammograms is a combination of the perceptual skills needed to find what may be very faint and small features in a complex visual environment, with the interpretative skills required to rate their diagnostic significance (Tabar and Dean, 1985). False negatives can be attributed to a number of factors:

1. incomplete visual search, e.g. because of fatigue, attention diversion,
2. missing of features e.g. because they are very faint, and
3. 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 seen. The third represents an error of interpretation.

We are involved in a project to develop a CAM system to analyse mammograms for signs of features known to be associated with the early stages of breast cancer. For each feature found, the system generates a prompt on a paper copy of the mammogram (see Figure 2). The approach is based upon experimental evidence that shows prompting can improve radiologists' performance by reducing errors of perception (Hutt, 1996).

A CAM system poses a challenge with respect to each of the MDSS allocation of function issues outlined earlier. In the case of scope, problems may occur for two reasons. First, current image analysis techniques are not able to find all the various types of mammographic feature in which radiologists are interested. Second, some kinds of feature may be hard to distinguish from one another, and features may also overlap, with the result that radiologists may misattribute a prompt to a feature which the system is not actually capable of detecting. Together, these two factors raise the possibility that radiologists may 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 discretionary and flexible use to be made of the prompting information: the radiologist will be free to determine when to consult the prompts and may choose to ignore them. It is evident, however, that changing from double reading to computer-aided single reading could present significant problems, and should not be attempted without a much better understanding of current clinic practices.

The issue of role concerns the question of where to draw the line between perception and interpretation when determining the functional role of the CAM system. The project's goal is to increase radiologists' sensitivity by reducing the number of false negatives attributable to errors of perception. The system is not intended to address the issue of false negatives attributable to errors of interpretation. In principle, the system's functional role may therefore be defined as perceptual, and not interpretative. However, in practice, the question of where to draw the line in CAM between perception and interpretation is problematic.

Drawing the line so as to limit the system's interpretative function has the virtue of achieving a complementary synthesis of system and radiologists' strengths: the former is more consistent in its visual search performance and the latter has interpretative skills which the system cannot match (Claridge, 1997). However, given the nature of the mammogram image, drawing the line in this way may lead to many 'low value' false positive prompts, i.e., prompts for features that radiologists can see are obviously benign. The danger is that radiologists may find such prompts distracting and ignore them, including some true positive prompts. In contrast, drawing the line so as to increase the system's interpretative function, and so reduce false positive prompts, is likely to cause its false negative prompt rate to increase.

In practice, some interpretative function (even if relatively simplistic) is essential in a CAM system. As with the human observer, perception and interpretation are operationally closely linked. For example, a CAM system which was unable to distinguish between a random distribution of microcalcifications and microcalcification clusters would be useless. The problem is to find the correct balance between perception and interpretation: too little interpretation and the system will fail in its objective of reducing false negative rates; 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 investigation of screening practices at a number of clinics in the UK. This included experiments, semi-formal interviews with radiologists and radiographers, and ethnographic-styled observation of work practices.

4. Breast screening in the UK

The UK Breast Screening Program (UKBSP) is a national service with a regional organisation. Each region is served by a number of screening clinics, each with two or more radiologists. The initial screening test is by mammography, where one or more X-ray films (mammograms) are taken of each breast by a radiographer. Each mammogram is examined for evidence of abnormality by two experienced radiologists. Types of feature that are indicators of malignancy include:

Microcalcifications are small deposits of calcium visible on a mammogram as tiny bright specks. They can be due to benign processes: 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.

Ill-defined lesions are areas of radiographically-dense tissue appearing as a 'bright patch' on the mammogram that might indicate a developing tumour. Typically, lesions that are well-defined are the result of benign processes: for example, they may be cystic. Lesions that do not have a well-defined edge are considered susp cious.

Stellate lesions are visible as a radiating structure with ill-defined borders. The rad ating components (or spicules) are the result of malignant processes infiltrating th breast tissue.

Architectural distortion may be visible when breast tissue around the site of a devel oping tumour contracts. In the absence of other signs this might give a subtle clu to the presence of a tumour.

Asymmetry between left and right mammograms may be the only visible sign of som hard to detect features. Asymmetry can be difficult to interpret as there is often natural asymmetry in the distribution of breast tissue.

When reading, radiologists may consult information provided by the radiographer tha could have a bearing on mammogram interpretation: for example, the location of scars whether the woman is taking HRT, etc. In this way, information from several sources i combined in the reading process. However, screening largely relies on radiologists' per ceptual and interpretative skills. Radiologists are highly trained and their work practice have evolved to reduce the likelihood of mistakes, especially false negatives.

Reading practices

Double reading involves each mammogram being examined independently by two ra diologists. Various studies have indicated that double reading may give a 5% to 15% improvement in cancer detection (Anderson et al., 1994; Warren and Duffy, 1995). Ther are variations in the way that double reading is implemented. The most simple metho is to recall on a 'worst opinion recalls' basis, i.e., if either, or both, radiologists decide t recall. Alternative methods include calling in a third radiologist to make the final decisio when radiologists disagree.

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

However, the reading process is more complex than it appears at first sight. Ou investigations indicate that categorisation of feature types is less clearly delineated tha the taxonomy described above suggests, particularly for ill-defined lesions. For example the appearance of some features may be ambiguous. Any linear structures associated wit a lesion might be interpreted as evidence for spiculation. Such structures are examine closely. If they are perceived to pass through, rather than originate within the feature then grounds for suspicion are diminished.

Radiologists may alter their recall threshold according to the type of tissue presen in a given mammogram. A feature in a mammogram that has a lot of asymmetricall distributed ('patchy') tissue might be treated with less suspicion than a similar featur appearing in a mammogram that has more evenly distributed tissue.

Monitoring and articulation of work

All aspects of screening work are closely monitored to reduce mistakes, particularly false negatives. Clinic staff monitor their own, and each others' performance through formal procedures for quality assurance and work documentation. Clinic staff hold regular meetings 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 public — aspects of their work which they perform as individuals, such as their reasons for giving a 'recall' or 'no recall' opinion. This emphasises the fact that despite its apparent individualised character, reading work is performed within a specific "community of practice" (Jordan, 1996). Review meetings, for example, serve to establish, reinforce and review where radiologists should be setting the recall threshold. It is important, for example, that differences between radiologists' recalls are maintained within a manageable range: the 'virtuous' difference which accounts for the improved detection rates observed in double reading. If the difference is too large, however, clinic specificity targets may be jeopardised, and changes in procedure may follow, like changing from a 'worst case' to a 'third reader arbitration' recall decision-making policy.

In many workplaces, a more informal kind of work articulation is achieved through the public character of documents (Hughes et al., 1996). In the screening clinic, the reporting form provides a particularly noteworthy example of this. Its design, together with the work practices in which it is embedded, mean that second readers see the first reader's opinion when they record their own. This provides second readers with the opportunity to compare their performance with that of their colleagues, *within* the context of their own reading work. We found no evidence that the availability of the first reader's opinion directly influences the second reader's opinion: on the contrary, we believe that radiologists do reach their decisions independently. Instead, we suggest that it serves to maintain a more general awareness of each others work.

We observed that in some clinics this informal articulation of work has evolved further: first readers sometimes annotate the reporting form. In a significant number of instances we found that these annotations related to features that the first reader had interpreted to be in category C2 (benign), i.e., cases which the first reader had decided not to recall. Figure 1 (1) shows one example of such an annotation. The first reader has marked on the breast schematic printed on the reporting form the site of a feature with an "X" and written "NRC" (no real change) beside it. In Figure 1 (2), the first radiologist has marked the site of a feature with "?" and written "BT" (breast tissue). Discussions with radiologists revealed that these annotations serve several purposes. First, in the event of the second reader deciding to recall, the first reader's annotations will provide useful information should the case go to third reader arbitration. Of particular interest, however, was that the radiologists emphasised how this practice of annotating benign features plays a less overt, but important role of keeping each other informed about their work. One radiologist remarked:

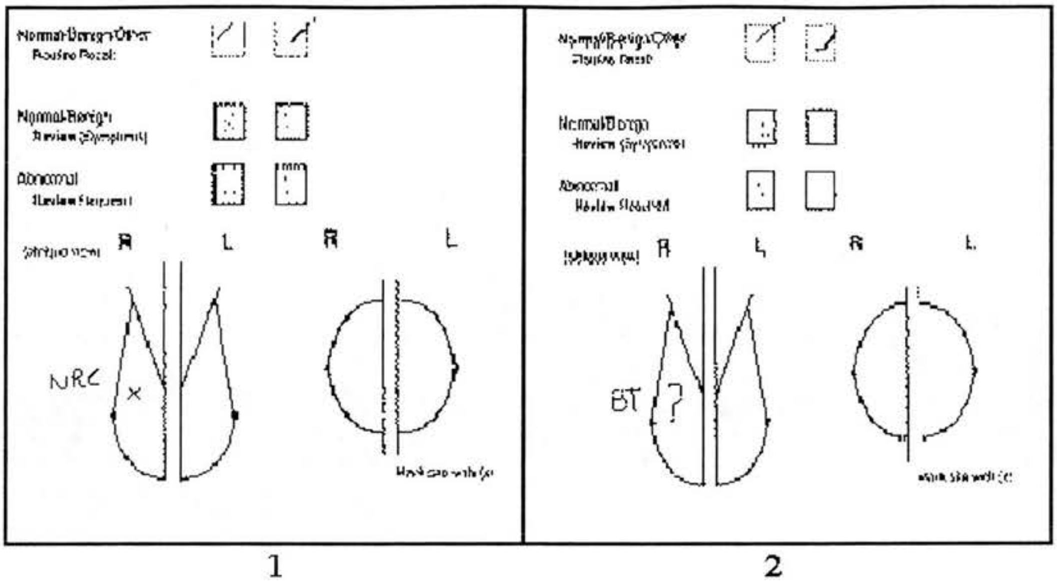


Figure 1: Examples of first readers' benign feature annotations.

> "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 reasoning behind her 'no recall' opinion. The first example (1) suggests little doubt in the first reader's mind that her opinion is correct: the annotation seems intended merely to reinforce 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 draw attention to — the first reader's uncertainty.

The fact that these informal work articulation practices should focus on features that fall on the benign side of the recall threshold may seem surprising. However, the region around the recall threshold is where most false positive and false negative decisions are likely to occur, and where the impact of differences in radiologists' opinions will be most significant. In choosing to document this aspect of their work, radiologists display an orientation to the collective monitoring and management of their recall decision-making and community of practice.

5. Previous investigations of prompting

Experimental evidence suggests that prompting can improve radiologists' performance by directing their attention towards suspicious features, but it was also found that if the false positive (FP) prompt rate is more than 1.5 times the True Positive (TP) prompt rate, then prompting ceased to be effective (Hutt, 1996). Since, in screening mammography, the underlying cancer rate is approximately 0.5%, then for a 90% sensitivity target, we may conclude that a prompting system may only be allowed 0.68 FP prompts per 100 cases. This represents a combination of specificity and sensitivity which is far superior to that achieved by any existing image analysis techniques, and, indeed, to that of any radiologist.

Though this conclusion is pessimistic of the value of current CAM techniques, it is

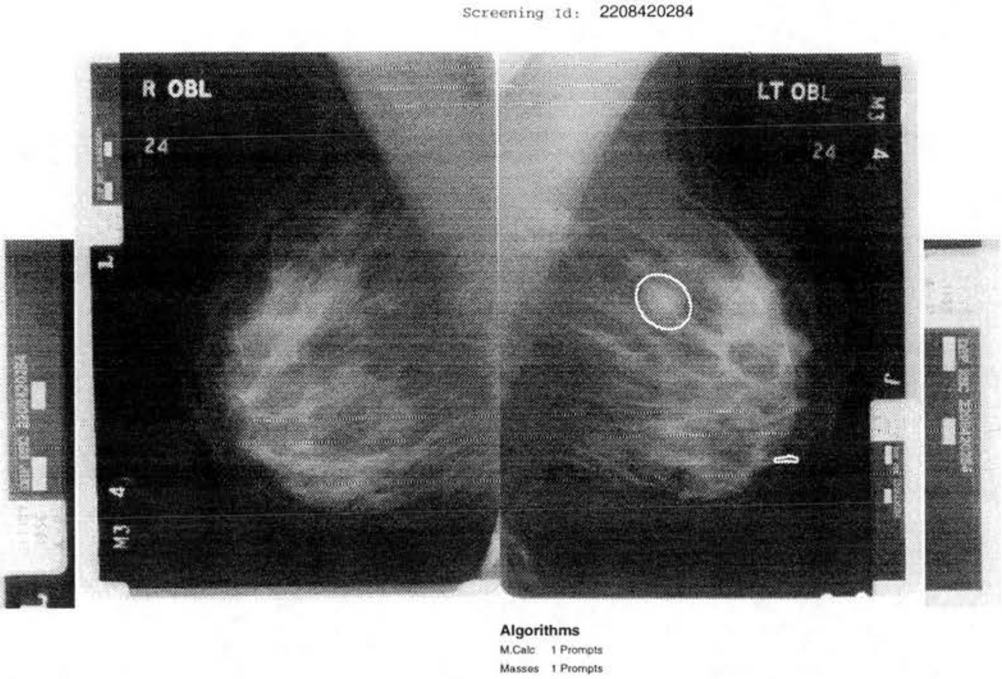


Figure 2: Example prompt sheet.

open to question. The studies employed heavily biased test sets in order to obtain a statistically significant measure of sensitivity improvement, and so the results may not be directly applicable to the circumstances in which reading is performed in the clinic. To investigate this further, we decided to explore how radiologists assessed the value of prompts under conditions more typical of reading in the clinic.

6. Investigating radiologists' assessment of prompting

In a series of experimental sessions, realistic reading conditions were simulated, including use of standard reporting forms and attaching reporting forms and prompt sheets to a film bag (Hartswood et al., 1997). Outputs from two of the CAM system's feature detection algorithms were used to generate prompts for microcalcification clusters (Hume et al., 1996) and ill-defined lesions (Miller and Ramsay, 1996). 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.

Prompt sheets consisted of a hard-copy, low resolution image of the mammogram pair with prompt information superimposed (Procter et al., 1994). Prompts for ill-defined lesions consisted of an ellipse surrounding the suspect region, and for microcalcifications, an irregular outline of the potential cluster (see Figure 2). 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. Before the experiment, subjects were given an overview of how the CAM system worked, including the types of feature it was capable of detecting.

The experiment consisted of four conditions. In three, subjects were prompted at different 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	Ill-defined lesions		Microcalcifications	
Condition	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%

control. The data recorded included recall rate and time taken to read each condition. Subjects were recorded on video, and their actions subsequently transcribed. Questionnaires were administered before and after the experiment, and after each condition. Subjects' attitudes to the system were assessed after each condition using a 20 point Likert test, with the higher the total score, the more favourable the assessment.

Results and discussion

Subjects were asked to rate each prompt 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 less distracting after the experiment than they were before it (p<0.05) (Hartswood, et al., 1997).

For the cases they recalled, subjects were asked to indicate whether the relevant feature had been correctly prompted. For the majority of subjects, a monotonically increasing Likert score was apparent as the number of prompts they judged to be correct increased. This suggests that subjects were more favourably disposed towards the system when the prompts corresponded with their expectations: i.e., when the 'opinion' of the system and that of the subject broadly coincided.

The protocol for the experiment instructed subjects to examine the films, examine the prompt sheets, and then record their opinion. The video transcripts revealed that subjects sometimes failed to follow instructions correctly. Table 2 shows the number of occasions when the subjects either failed completely to examine the prompt sheet (Type 1 error), and when they recorded their opinion before examining the prompt sheet (Type 2 error). In the latter case, subjects may have turned the reporting form over after recording their opinion, and then gone back to it realising that they had forgotten to examine the prompt sheet. Taking radiologists' 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. These results suggest that at lower prompting rates there was insufficient information to hold the subjects' attention, either because of the frequency, or quality, (or both) of the prompts.

In eliciting post-condition comments, we sought to explore how use of the CAM system contributed to subjects' understanding of its behaviour. The results were mixed: for example, since there were so few pathology-proven cancers in the test set, we had expected that subjects would not be able to assess the system's sensitivity accurately. In fact, their unanimous opinion that the sensitivity of the system for ill-defined lesions (62% maximum) was too low showed their grasp of this aspect of system behaviour was good. In contrast, several subjects expressed the belief that the system was detecting asymmetries, even though it could not.

Overall, the results of this experiment indicated that under more realistic conditions,

Table 2: Number of occasions subjects did not examine prompt *at all* (Type 1 error), or only 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

radiologists' tolerance level for FP prompts was appreciably higher than the upper limit previously established for improved radiologist performance. Of course, positive assessment by radiologists may not necessarily coincide with improvement in performance, but these results raised the possibility that, perhaps because of its artificiality, earlier work had underestimated the FP prompt upper limit. One explanation is that subjects' tolerated FP prompts in the new experiment because they provided useful information about the CAM system's behaviour. To test this, a follow-up study was devised to examine in more detail how radiologists might use prompts to construct and confirm a model of the system's behaviour.

7. Classification of prompts

Three experienced radiologists were asked to examine the prompts produced at the highest sensitivity in the earlier experiment, and to decide whether they would recommend a recall on the basis of the prompted feature. They were also asked to asked to classify each of the features prompted according to their own confidence scale: C1 (normal), C2 (benign), C3 (equivocal), C4 (suspicious) and C5 (malignant), and whether they thought prompting for these features would be acceptable in routine screening. In addition, radiologists were encouraged to vocalise their thoughts, and these were recorded and transcribed.

Evaluating the system

Table 3 compares how different pairs of radiologists classified the same set of prompted features as a 'recall' or 'no recall'. The interesting cases are those where the radiologists

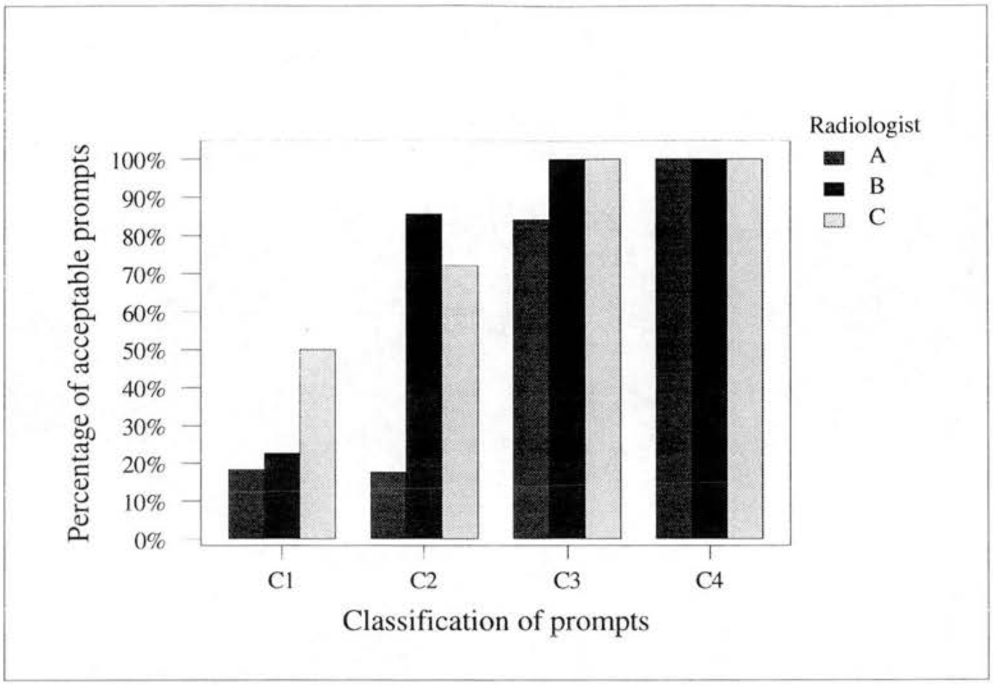


Figure 3: Percentage of acceptable prompts by prompt classification.

disagreed (the highlighted cells). As noted earlier, radiologists do not always agree on which cases to recall. It is not surprising therefore that radiologists' classification of prompted features also shows some marked differences of opinion.

Figure 3 shows radiologists' ratings of prompt acceptability broken down by prompt classification (C1:C4). It is clear from these results that the boundary between 'not acceptable' and 'acceptable' lies within the C2 category, i.e., prompts for benign features. This leads us to suggest that the effect of FP prompts will depend on their classification. When reading unaided, radiologists perceive and interpret *candidate* features which include members of the C2 (probably benign) category: i.e., features that have some properties in common with those they interpret as suspicious (i.e., C3:C5). We argue that prompts for *candidate* features may be acceptable to radiologists in the clinical setting, whereas prompts for *other* features (i.e., C1) would be distracting. This may explain the results seen in earlier work. We conclude therefore that the appropriate place to draw the line between perception and interpretation is so that the system can distinguish between C1 and C2 features.

What is also interesting about these results is the parallel between radiologists' apparent tolerance of C2 prompts and their ad-hoc practice of annotating C2 features. There will always be cases where the absence of a prompt may give ambiguous evidence of the system's performance. It could mean that the system found no feature (a possible 'error' of perception), or that it found a feature and then determined it to be benign (a possible 'error' of interpretation). We suggest that radiologists may find this ambiguity a source of confusion when attempting to understand where the system draws the line between perception and interpretation. In this critical region of performance, they prefer to have the less unambiguous evidence of a prompt because of its capacity to document the CAM system's behaviour.

We argue that radiologists find some of the CAM system's FP prompts useful in much the same way as they find each others annotations of benign features useful. To reiterate, radiologists annotate these features as a way of documenting a particularly critical region of their reading performance. It is not so surprising therefore that radiologists should also show an interest in this same region of the system's performance. We conclude that prompts for *candidate* features afford learning about — and confirmation of — the system's behaviour.

Making sense of the system

The following extracts of session verbal protocols illustrate how radiologists tried to make sense of the CAM system's behaviour from the evidence of the features it had prompted. In a number of instances, the transcripts show examples of misunderstanding of the extent of the system's capabilities, and confusion because of apparent inconsistencies in its behaviour.

In this first set of extracts, the comments suggest radiologists were unable to accurately place the CAM system's operational scope: i.e., the types of feature it is capable of detecting:

> "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."

In the first extract, the radiologist interpreted as an asymmetry a feature which the system prompted as an ill-defined lesion. In fact, the system does not prompt asymmetries, but it was evident from the transcript as a whole that radiologists explained the behaviour of the system by assuming that it was capable of detecting asymmetry. This led to expectations that were difficult to fulfil. In the second extract, the radiologist expressed disappointment with the system precisely because it had failed to prompt an area of asymmetry.

The next set of extracts focuses on radiologists' problems with understanding how the system 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."

In the first extract, the radiologist did not interpret the particles of microcalcification as forming a series of discrete clusters: her interpretation was that there was simply a widespread random distribution. The system prompts a region of microcalcifications if it identifies five or more particles in the neighbouring area. In this instance, by chance, some of the randomly distributed particles met the system's criteria, and so a prompt was produced. The system only examines the image locally to determine if the cluster criterion is met. In contrast, the radiologist is able to make a global appraisal, and can discover larger scale trends that are not apparent to the CAM system. In this case, the radiologist concluded there were no microcalcification clusters, and was perplexed as to why one part of the "random distribution" should be prompted over any other.

In the second extract, confusion arose because the radiologist automatically classified the calcification present as being vascular, then posed the question: "why some vascular calcification and not others?" Again, the system has a much simplified interpretation: part of the vascular calcification had fragmented into number of particles which were sufficiently close together for the system to interpret them as a cluster. The remainder of these vascular calcifications maintained their characteristic tram line appearance, and were not prompted.

The third extract is particularly interesting. Once more, the radiologist was perplexed because the system's interpretation of a cluster was less sophisticated than her own. Initially, the radiologist decided that there was a single cluster of lobular calcification, and several clusters of vascular calcification present. The radiologist was more interested in the former than the latter, and so was disappointed when only the vascular clusters were prompted. The lobular calcifications were very subtle, and so would have needed to form a tight cluster in order to be prompted. On the other hand, some vascular calcifications qualified as clusters according to the system's interpretation. The radiologist made a qualitative distinction between the vascular and lobular clusters, but the system has no such interpretative capacity, and so fell short of the radiologist's expectations.

In the final set of extracts, the radiologists indicated that they had not see anything of significance 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 subtle calcifications present. Radiologists do make a point of looking for subtle clusters, however, very subtle clusters occur relatively frequently, are mostly benign, and present insufficient information in terms of size, shape and distribution for a radiologist to identify malignant ones. Furthermore, if there is disease present, at this stage it is likely to develop relatively slowly, and so there is a reduced risk in waiting until the subsequent screening round when there might be more evidence. In these examples the system is too perceptually acute, producing prompts for features that are difficult for radiologists to locate, and that also have little diagnostic relevance.

8. Conclusions and future work

Computer-aided mammography raises allocation of function issues with regard to scope, role and work practice. Taking the issue of role first, our investigations indicate that a CAM system should have sufficient interpretative capability to distinguish between *candidate* and *other* features. Our evidence suggests that from the radiologists' point of view, this would mean drawing the line between perception and interpretation at the C1 (normal) : C2 (benign) boundary. We acknowledge that these are subjective effects, and that so far we have no evidence that radiologists' actual reading performance will be improved. Large scale trials are being planned to obtain statistically reliable measures of the latter.

Our investigations also show that radiologists may have problems in understanding the operational scope of CAM systems, particularly at their boundaries. This points to the importance not only of determining where to draw the line between perception and interpretation, but also of radiologists *knowing* where it is. Our evidence suggests that prompts not only serve as a cue to examine particular features, but also as an aid to the development of radiologists' understanding of how the system works, and what its capabilities are.

Training sessions and manuals are useful resources for explanation, and we have used the results of our investigation to inform the content of such materials. However, for sustainable understanding, systems need to provide accounts of their behaviour which are both relevant to, available, and understandable within the actual *doing* of the work they support. The problem is that CAM systems are complex, and that prompts only provide a very limited account of their behaviour. We are currently exploring ways in which these accounts can be enriched. Our approach is informed not only by the way individual radiologists' make decisions, but also by the ways in which they sustain their broader community of practice.

This brings us to the final issue of work practice. Through the public character of the reporting form, double reading provides a means by which radiologists can make their work available to each other *as they do it*, i.e., where this information is most likely to be relevant, and understandable. Double reading therefore contributes to the collective maintenance of clinics' screening performance in ways, which though they are informal and sometimes ad-hoc, may be just as important as its nominal effect. Double reading has evolved practices through which radiologists' work can simultaneously both be explicitly distributed *and* implicitly collective. The adoption of computer-aided single reading would inevitably mean the disruption of these practices. We conclude, therefore, that the collective dimension of reading work must be better understood if the potential benefits of computer-aided mammography are to be fully realised. Further investigation of this issue is also planned.

References

Anderson, E. D. C., Muir, B. B., Walsh, J. S. and Kirkpatrick, A. E., 1994, The Efficacy of Double Reading Mammograms in Breast Screening, *Clinical Radiology*, **49**, pp. 248-251.

Blois, M. S., 1980, Clinical judgment and computers, *The New England Journal of Medicine*, **303**, pp. 192-197.

Claridge, E. 1997, Experts' assessment as a "gold standard" for characterisation of lesions? In Proceedings of Medical Image Analysis and Understanding '97, Oxford.

Clayton, P. D., and Hripcsak, G., 1995, Decision support in healthcare. *International*

Journal of Biomedical Computing, **39**, pp. 59-66.

Forsythe, D. E., 1992a, Blaming the user in medical informatics: The cultural nature of scientific practice, *Knowledge and Society*, **9**(3), pp. 95-111.

Forsythe, D. E., 1992b, Using ethnography to build a working system: Rethinking basic design assumptions. In Proceedings of the 16th Symposium on Computer Applications in Medical Care, pp. 505-509.

Heathfield, H. A. and Wyatt, J., 1993, Philosophies for the Design and Development of Clinical Decision Support Systems, *Methods of Information in Medicine*, **32**(1), pp. 1-8.

Hartswood, M., Procter, R., Williams, L., and Prescott, R., 1997, Subjective Reaction to Prompting in Screening Mammography. In Proceedings of Medical Image Analysis and Understanding '97, Oxford.

Hughes, J., King, V., Mariani, J., Rodden, T., and Twidale, M., 1996, Paperwork and its Lessons for Database Systems. In The Design of Computer Supported Cooperative Work and Groupware Systems by D. Shapiro, M. Trauber and R. Traunmuller (eds.) (North-Holland), pp. 43-62.

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

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

Jordan, B. Ethnographic Workplace Studies and CSCW. In The Design of Computer Supported Cooperative Work and Groupware Systems by D. Shapiro, M. Trauber and R. Traunmuller (eds.) (North-Holland), pp. 17-42.

Kaplan, B., 1982, The influence of medical values and practices on medical computer applications. In Proceedings of the First International Conference on Medical Computer Science/Computational Medicine, pp. 83-88.

Kaplan, B., 1988, Development and acceptance of medical information systems: an historical overview, *Journal of Health and Human Resources Administration*, **1**, pp. 9-29.

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

Miller, R. A., 1994, Medical Diagnosis and Decision Support Systems — Past, Present and Future, *Journal of the American Medical Informatics Association*, **1**(1), pp. 8-27.

Procter, R., Thanisch, P., Astley, S., and Hutt, I., 1994, User interface design and data management for digital mass mammography. In Proceedings of the Second International Workshop on Digital Mammography, York.

Schwartz, W. B., Medicine and the Computer, 1970, *New England Journal of Medicine*, **283**, pp. 1257-1264.

Sutton, G. C., 1989, Computer-aided diagnosis: a review, *British Journal of Surgery*, **76**, pp. 82-85.

Tabar, L. and Dean, P., 1985. Teaching Atlas of Mammography, Theime.

Warren, R. M. L., and Duffy S. W., 1995, Comparison of single reading with double reading of mammograms, and changes in effectiveness with experience, *The British Journal of Radiology*, **68**, pp. 958-962.

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

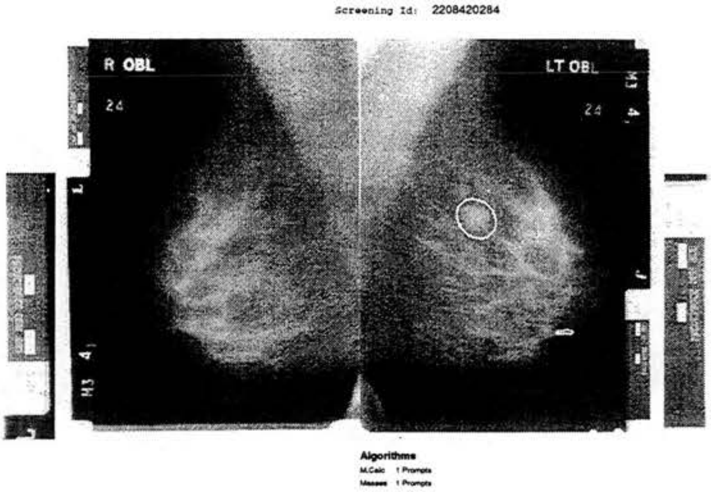


Fig. 1. Example prompt sheet

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.

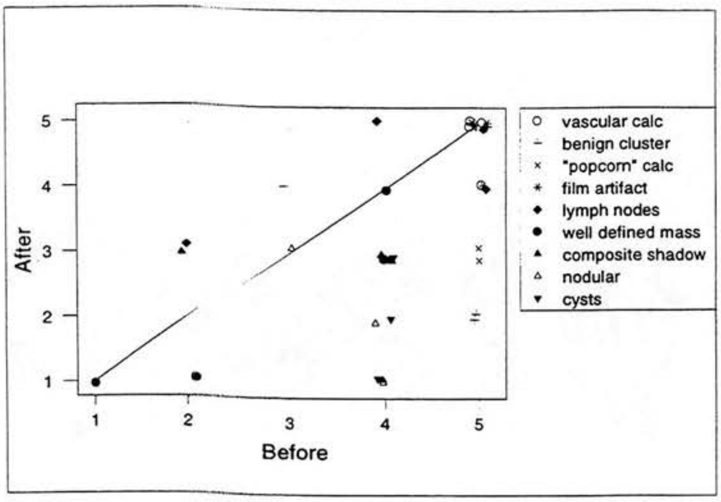


Fig. 2. Subjects' assessment of value of particular types of FP prompt (1 = useful and 5 = distracting) before and after the experiment.

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.

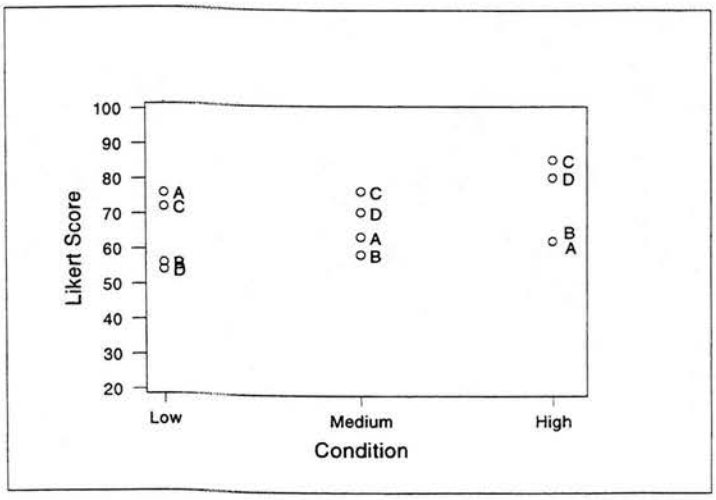


Fig. 3. Likert score against condition for subjects A to D.

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

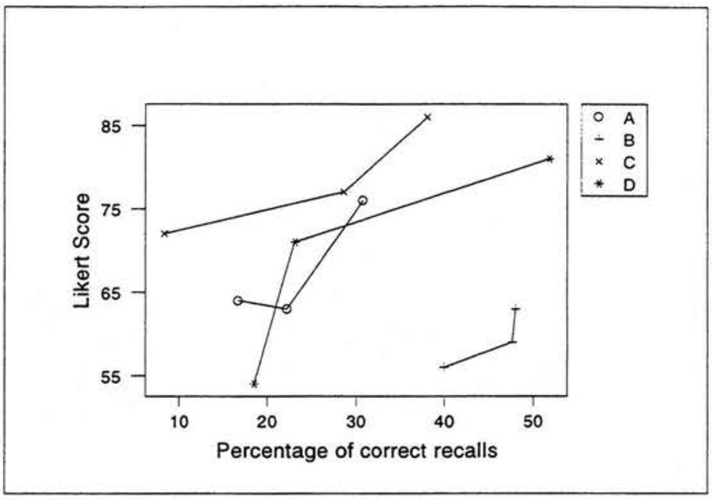


Fig. 4. Percentage of correctly prompted recalls against Likert score for each subject.

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.

Prompting in mammography: Computer-aided Detection or Diagnosis?

M. Hartswood[a], R. Procter[a], L. J. Williams[b]

[a]Department of Computer Science, Edinburgh University, Scotland.

[b]Department of Public Health Sciences, Edinburgh University, Scotland.

Corresponding Author:
Mark Hartswood
Department of Computer Science
James Clark Maxwell Buildings
King's Buildings
Edinburgh EH9 EJZ
Scotland
United Kingdom

email: mjh@dcs.ed.ac.uk
tel: 0131 650 5899
fax: 0131 667 7209

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

This work is concerned with the use of Computer-Aided Detection systems by radiologists to assist with screening mammography. Our focus is on Human Factors issues, particularly 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 the information supplied by a detection system as evidence to support diagnostic decisions in cases where radiologists are uncertain about the interpretation of a lesion. It is possible that this mode of use may have a detrimental effect on performance.

We have previously suggested that radiologists are able to use false positive prompts as evidence to assess the capabilities of a prompting system. We concluded that some categories of false positive prompts can be useful in providing an account of system behaviour over and above that available from true positive prompts [2]. Our investigations revealed that users of a prompting system were able to form an accurate assessment of system performance but were less well able to determine accurately the system's scope and function from the evidence of the prompts alone. The work informed the development of a prototype training package. Our aim was to present a model of 'best practice' for using the prompt information and also to provide an account of system behaviour through a set of examples of true positive, false positive and false negative prompts [3].

In subsequent investigations we have focussed on how radiologists use the prompts in their decision making. In this paper we outline how diagnostic decisions may be affected, we also give a more detailed treatment of these results in [4].

Others have assessed the impact of prompting on performance and decision making, for example, Mugglestone [5] and Chan [1]. Typically, in this type of work, somewhat artificial conditions are required to obtain quantitative measures of performance. In contrast, we have examined how prompting systems are used in conditions closely resembling normal clinical practice, employing techniques such as participant observation, interviews and questionnaires to inform our interpretation of outcome data.

The setting for our work is the continuing evaluation of the PROMAM prompting system. In this paper we report data obtained from small scale clinical trial of PROMAM involving 5 radiologists from a Scottish breast screening centre reading two thousand archive cases [6].

References

[1] Chan, H. et al. (1990) Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms. Investigative Radiology, vol 25, p. 1102–1110.

[2] Hartswood, M., Procter, R., Williams, L., and Prescott, R. (1997) Subjective Reaction to Prompting in Screening Mammography. In proceedings of Medical Image Analysis and Understanding, Oxford.

[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 proceedings of the First International Conference on Allocation of Functions. Galway. IEA Press, pp. 275-291.

[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] Mugglestone, M., Lomax, R., Gale, A. G. and Wilson, A. R. M. (1996) The effect of prompting mammographic abnormalities on the human observer. In Doi, K. et al. (Eds.) Proceedings of the Third International Workshop on Digital Mammography. Chicago, June.

[6] Williams, L. J., Prescott, R. and Hartswood, M.(1997) Computer-aided cancer detection and the UK National Breast Screening Programme. Submitted for inclusion in these proceedings.

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

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 [3, 4].

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 [1]. Another study reported that prompting can entail an increase in False Positive (FP) decisions without necessarily having an overall effect on confidence levels [6]. 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 prompts per case; mass sensitivity 72.9%, FP rate of 0.66 prompts per case [7]. 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 [5]. In particular, they were instructed that they should not use prompts as contributory evidence in their recall/normal decisions.

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

3 Results and Discussion

In each of the post-prompted session interviews, subjects were asked if the prompts had 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.

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 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 Figure 1 (a) and (b) respectively.

The first question shows little difference between subjects' pre- and post-trial opinions, with only one subject changing their opinion from 'Uncertain' to 'Agree'. This is perhaps not remarkable — if there is uncertainty in diagnosis, it might be expected that the default position would be to recall.

The second question shows that there is a consolidation of opinion post-trial, with subjects being more likely to believe that the absence of a prompt might influence their recall decisions. In a sense this is counter intuitive if, given uncertainty, the default position might be to recall. However, subjects' responses can be explained in a similar way to those in the first question if prompting information is being used to aid diagnosis.

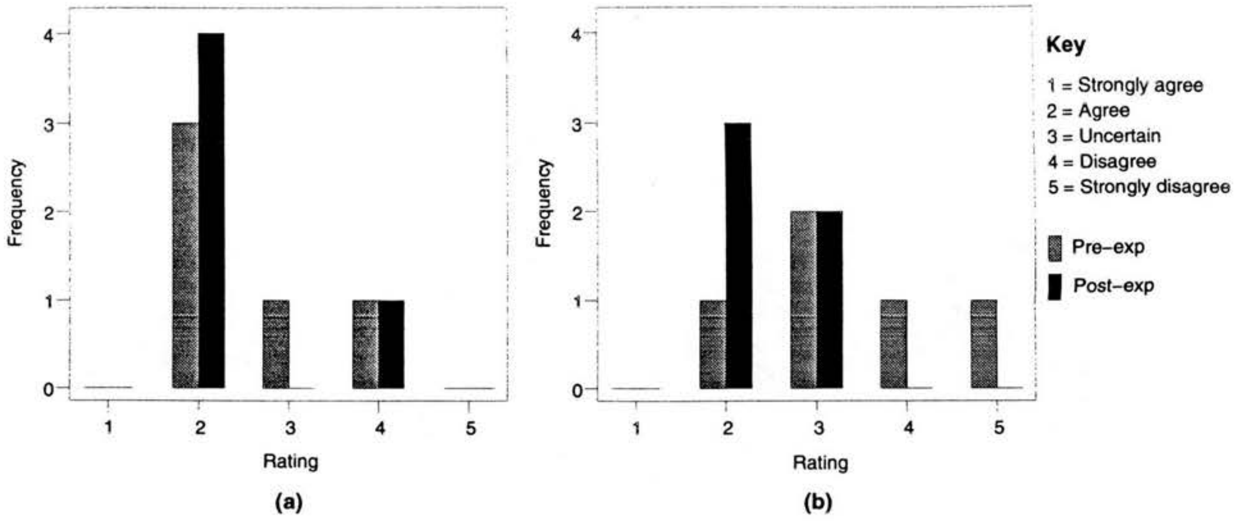


Fig. 1. (a) the presence of a prompt will make me more inclined to recommend recall (b) the absence of a prompt will make me less likely to recommend recall.

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
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 the number of prompted single reader recalls and the number of unprompted single reader recalls. Of the prompted single reader recalls, 50.7% were correctly prompted for by the system, where as 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 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. It is also consistent, however, with the conclusions drawn from both subjects' comments in the interview data, and with their questionnaire responses, that the presence and absence of prompts influences their diagnosis.

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 [2]. 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 modifying 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 radiologists 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 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. Chan, H. et al. (1990) Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms. Investigative Radiology, vol 25, p. 1102–1110.
2. Giger, M. (1993) Computer-aided Diagnosis. In Haus, A. and Yaffe, M. (Eds.) A categorical course in physics: Technical aspects of breast imaging, p. 283–298.
3. 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.
4. 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.
5. 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.
6. Mugglestone, M., Lomax, R., Gale, A. G. and Wilson, A. R. M. (1996) The effect of prompting mammographic abnormalities on the human observer. In Doi, K. et al. (Eds.) Proceedings of the Third International Workshop on Digital Mammography. Chicago, June.
7. 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 CANCER DETECTION IN THE UK BREAST SCREENING PROGRAMME

LINDA J. WILLIAMS, ROBIN J. PRESCOTT AND MARK HARTSWOOD[1]
Department of Public Health Sciences, University of Edinburgh
[1] Department of Computer Science, University of Edinburgh

1. Introduction

PROMAM (PROmpting for MAMmography) is a computer-aided prompting system, designed to detect the early signs of breast cancer on mammographic films. These films are scanned at 42 micron resolution, then analysed by two algorithms, specifically searching for micro-calcification clusters and ill-defined masses. Suspicious areas are then highlighted on a low-resolution image of the films, and printed onto a piece of paper. This 'prompt sheet' is then given to the radiologists as part of the usual information to which they would have access.

In this paper, we present the results of an experiment using PROMAM in a clinical setting, using archive film. The aims of this investigation were twofold: to ensure that there would be no additional recall burden; and to obtain an estimate of the agreement between prompted and unprompted radiologists, to better assess the sample size required for the full trial.

2002 cases (the sets of films belonging to one woman) were scanned and analysed. Of these, 102 cases were pathology proven malignancies, chosen so that the proportion of cancers to normals was approximately the same ratio as would normally be found in the relationship between recalls and non-recalls.

2. Method

Five radiologists from Ardmillan House (the South East Scotland Breast Screening Centre) were asked to participate in the experiment, outwith their normal working hours. The 2002 cases were divided into batches of roughly one hundred cases, with each batch being read by both a prompted and an unprompted radiologist.

2.1 ALLOCATION OF CANCERS TO BATCHES

In order that the radiologists would not learn how many cancers were in a batch of 100, the number was randomised around a mean of 5. They were told that the batches were weighted with more malignancies than they would usually find in screening, but that the actual number of cancers per batch would vary, and could conceivably be as low as zero.

Due to the high volume of cases that needed to be read, the limiting factor in this experiment was the availability of radiologists' time. Hence, normal randomisation was impractical and a system of minimisation was devised, based on the number of cancers the available pair of radiologists had read first/second and prompted/not prompted, and, above all, the number of cancers that had been prompted first and second.

2.2 BLIND DOUBLE READING

To ensure that each reading practice (with or without prompt) was treated equally, radiologists were asked to read each case blind. In other words, they had no access to the decisions made by the previous reader. Since the reporting forms were colour-coded to aid data input, there was little reason to conceal whether a radiologist was the first or second reader. Hence, they were not blinded to reading order.

Due to technical difficulties, it was not possible to allow the readers unsupervised access to the case histories, although information was always available when requested.

2.3 RECORDING OF DATA

With respect to the aims of this paper, the principal data were the decisions to recall or not recall each case. This was recorded in a custom-designed database using the Access utility. Since the cancers (and the previously recalled non-cancers) were known to us, it was later possible to subdivide the data according to malignant or non-malignant. In addition, radiologists were asked to note if, for the cases that they recalled, the system had correctly prompted the suspicious feature.

3. Results

For ease of illustration, cancers and non-cancers will be treated as two separate groups. All recall rates are based on non-cancers only (since the number of cancers present in the full set is unrealistically high). The participating radiologists will be referred to as A, B, C, D, and E, for the sake of anonymity.

Radiologist	Cancer detection		Recall rate	
A	27/32	84.4%	34/563	6.0%
B	31/36	86.1%	54/770	7.0%
C	45/46	97.8%	82/758	10.8%
D	25/28	89.3%	49/572	8.6%
E	55/62	88.7%	78/1139	6.8%

Table 1: summary of response data

Table 1 is a summary of the overall performance of the radiologists. Cancer detection is given as 'the number of correctly identified cancers'/'the number of cancers shown to the radiologist'. Similarly for the recall rate; 'the number of (non-cancer) recalls made'/'the number of non-cancer films seen by the radiologist'. From these results, it would appear that a high sensitivity is gained at the expense of a high recall rate

3.1 RECALL RATE

As mentioned previously, one of the more important functions of this experiment was to ensure that the recall rate would not increase when the prompting system was added to the usual clinic procedures.

		Unprompted			
		Recalled	Not recalled	Tech recall	Total
Prompted	Recalled	65	69	0	134 (7.1%)
	Not recalled	97	1650	10	1757
	Tech recall	1	7	1	9
	Total	163 (8.6%)	1726	11	1900

Table 2: recalls made under prompted and unprompted conditions

Contrary to expectation, the recall rate for the prompted readings was lower than that for the unprompted readings (p=0.0175). However, further examination of the data by conditional logistic regression showed that the radiologist was more influential in the number of recalls than the presence or absence of a prompt, with prompting having only a marginal effect after the radiologist and batch effects had been taken into account.

3.2 CANCER DETECTION

		Unprompted		
		Recalled	Not recalled	Total
Prompted	Recalled	86	5	91 (89.2%)
	Not recalled	6	5	11
	Total	92 (90.2%)	10	102

Table 3: cancers detected under prompted and unprompted conditions

Disappointingly, there was no discernable difference between the prompted and unprompted conditions for the cancers. However, this may be due to one 'outlier' batch, where the unprompted radiologist correctly identified all six cancers present in the batch, and the prompted radiologist only identified three. Why this is the case, we do not know, as the three cancers that were missed were all correctly prompted by the system.

3.3 ALGORITHMS

Algorithm performance is of major importance to the project; how the radiologists perceive the algorithms' accuracy will determine how much reliability they imbue the prompts. There would be little advantage in presenting prompts that are then disregarded out of hand by the radiologist.

	correctly prompted	not correctly prompted	total
Recalled by both	72	14	86
Recalled by prompted only	5	0	5
Recalled by unprompted only	4	2	6
Recalled by neither	2	3	5
Total	83	19	102

Table 4: algorithm performance on malignancies

Overall, the system was 81.4% sensitive on pathology proven malignancies, with 83.6% of women prompted. However, this can be further subdivided by the algorithm involved; micro-calcification was 93.8% sensitive, ill-defined lesions was 72.9% and for cases where both a lesion and a micro-calcification cluster were present, the sensitivity was 81.8%. The related prompt rates were 17.6% (micro-calcification prompt only), 29.7% (mass prompt only) and 36.4% (both types of prompts) of women prompted.

4. Conclusions

Despite the lack of improvement in cancer detection, the experiment was deemed a success due to the non-increase in recall rate when the radiologists were prompted. As the experiment was designed as a pre-clinical trial, it was never expected that we would find any significant improvement in cancer detection. Results revealed a level of agreement between the prompted and unprompted readers in cancer detection of 93%. Since the national cancer detection rate is 5.45 per 1000 (1995/6 figures), to detect a relative improvement of 6% would require a trial size of approximately 90,000 women.

Acknowledgements
We wish to extend our grateful thanks to the staff of Ardmillan House for their time and patience, especially Mrs. Pat Dixon; to the algorithm developers (also presenting a paper at this conference); and to Mr. Bill Cormack for support, encouragement and leadership of the PROMAM team.

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.

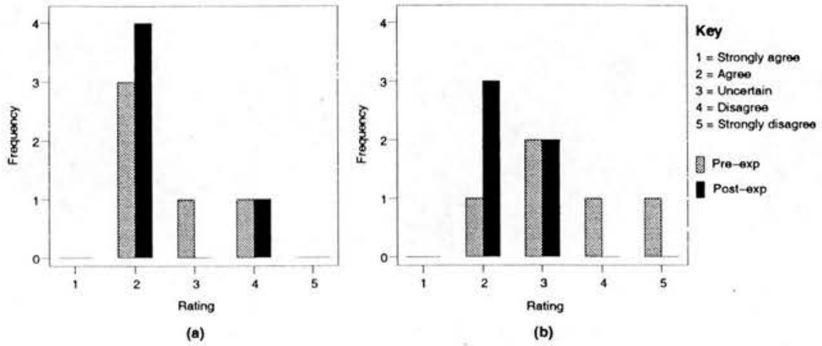


Figure 1. (a) the presence of a prompt will make me more inclined to recommend recall; (b) the absence of a prompt will make me less likely to recommend recall.

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.