

Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography

P Taylor, J Champness, R Given-Wilson,
K Johnston and H Potts



February 2005

**Health Technology Assessment
NHS R&D HTA Programme**





INAHTA

How to obtain copies of this and other HTA Programme reports.

An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (<http://www.hta.ac.uk>). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

- fax (with **credit card** or **official purchase order**)
- post (with **credit card** or **official purchase order** or **cheque**)
- phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque

If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card

The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order

You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of HTA on CD?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography

P Taylor,^{1*} J Champness,² R Given-Wilson,³
K Johnston⁴ and H Potts⁵

¹ Centre for Health Informatics and Multiprofessional Education (CHIME),
Royal Free and University College Medical School, London, UK

² Cancer Services Collaborative, St George's Hospital, London, UK

³ Radiology Department, St George's Hospital, London, UK

⁴ Scottish Executive, Edinburgh, UK

⁵ Adamson Centre, St Thomas' Hospital, London, UK

* Corresponding author

Declared competing interests of authors: none

Published February 2005

This report should be referenced as follows:

Taylor P, Champness J, Given-Wilson R, Johnston K, Potts H. Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography. *Health Technol Assess* 2005;**9**(6).

Health Technology Assessment is indexed and abstracted in *Index Medicus/MEDLINE*, *Excerpta Medica/EMBASE* and *Science Citation Index Expanded (SciSearch®)* and *Current Contents®/Clinical Medicine*.

NHS R&D HTA Programme

The research findings from the NHS R&D Health Technology Assessment (HTA) Programme directly influence key decision-making bodies such as the National Institute for Clinical Excellence (NICE) and the National Screening Committee (NSC) who rely on HTA outputs to help raise standards of care. HTA findings also help to improve the quality of the service in the NHS indirectly in that they form a key component of the 'National Knowledge Service' that is being developed to improve the evidence of clinical practice throughout the NHS.

The HTA Programme was set up in 1993. Its role is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care, rather than settings of care.

The HTA programme commissions research only on topics where it has identified key gaps in the evidence needed by the NHS. Suggestions for topics are actively sought from people working in the NHS, the public, consumer groups and professional bodies such as Royal Colleges and NHS Trusts.

Research suggestions are carefully considered by panels of independent experts (including consumers) whose advice results in a ranked list of recommended research priorities. The HTA Programme then commissions the research team best suited to undertake the work, in the manner most appropriate to find the relevant answers. Some projects may take only months, others need several years to answer the research questions adequately. They may involve synthesising existing evidence or designing a trial to produce new evidence where none currently exists.

Additionally, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme is able to commission bespoke reports, principally for NICE, but also for other policy customers, such as a National Clinical Director. TARs bring together evidence on key aspects of the use of specific technologies and usually have to be completed within a limited time period.

Criteria for inclusion in the HTA monograph series

Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this monograph was commissioned by the HTA Programme as project number 98/16/04. As funder, by devising a commissioning brief, the HTA Programme specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme or the Department of Health.

Editor-in-Chief: Professor Tom Walley
Series Editors: Dr Peter Davidson, Professor John Gabbay, Dr Chris Hyde,
Dr Ruairidh Milne, Dr Rob Riemsma and Dr Ken Stein
Managing Editors: Sally Bailey and Caroline Ciupek

ISSN 1366-5278

© Queen's Printer and Controller of HMSO 2005

This monograph may be freely reproduced for the purposes of private research and study and may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising.

Applications for commercial reproduction should be addressed to NCCHTA, Mailpoint 728, Boldrewood, University of Southampton, Southampton, SO16 7PX, UK.

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.



Abstract

Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography

P Taylor,^{1*} J Champness,² R Given-Wilson,³ K Johnston⁴ and H Potts⁵

¹ Centre for Health Informatics and Multiprofessional Education (CHIME), Royal Free and University College Medical School, London, UK

² Cancer Services Collaborative, St George's Hospital, London, UK

³ Radiology Department, St George's Hospital, London, UK

⁴ Scottish Executive, Edinburgh, UK

⁵ Adamson Centre, St Thomas' Hospital, London, UK

* Corresponding author

Objectives: To determine the value of computer-aided detection (CAD) for breast cancer screening.

Design: Two sets of mammograms with known outcomes were used in two studies. Participants in both studies read the films with and without the benefit of a computer aid. In both studies, the order of reading sessions was randomised separately for each reader. The first set of 180 films, used in study 1, included 20 false-negative interval cancers and 40 screen-detected cancers. The second set of 120 films, used in study 2, was designed to be favourable to CAD: all 44 cancer cases had previously been missed by a film reader and cancers prompted by CAD were preferentially included.

Setting: The studies were conducted at five UK screening centres between January 2001 and April 2003.

Participants: Thirty radiologists, five breast clinicians and 15 radiographers participated.

Interventions: All cases in the trial were digitised and analysed using the R2 ImageChecker[®] version 2.2. Participants all received training on the use of CAD. In the intervention condition, participants interpreted cases with a prompt sheet on which regions of potential abnormality were indicated.

Main outcome measures: The sensitivity and specificity of participants were measured in both intervention and control conditions.

Results: No significant difference was found for readers' sensitivity or specificity between the prompted and unprompted conditions in study 1 [95% confidence index (CI) for sensitivity with and without CAD is 0.76 to 0.80, for specificity it is

0.81 to 0.86 without CAD and 0.81 to 0.87 with CAD]. No statistically significant difference was found between the sensitivity and specificity of different groups of film reader (95% CI for unprompted sensitivity of radiologists was 0.75 to 0.81, for radiographers it was 0.71 to 0.81, prompted sensitivity was 0.76 to 0.81 for radiologists and 0.69 to 0.79 for radiographers). Thirty-five readers participated in study 2. Sensitivity was improved in the prompted condition (0.81 from 0.78) but the difference was slightly below the threshold for statistical significance (95% CI for the difference -0.003 to 0.064). Specificity also improved (0.87 from 0.86); again, the difference was not significant at 0.05 (95% CI -0.003 to 0.034). A cost-effectiveness analysis showed that computer prompting increases cost.

Conclusions: No significant improvement in film readers' sensitivity or specificity or gain in cost-effectiveness was established in either study. This may be due to the system's low specificity, its relatively poor sensitivity for subtle cancers or the fact the prompts cannot serve as aids to decision-making. Readers may have been better able to make use of the prompts after becoming more accustomed to working with them. Prompts may have an impact in routine use that is not detectable in an experimental setting. Although the case for CAD as an element of the NHS Breast Screening Programme is not made here, further research is required. Evaluations of new CAD tools in routine use are underway and their results should be given careful attention.



Contents

Glossary and list of abbreviations	vii	Why are the prompts ignored?	36
Executive summary	ix	Impact on professional groups	37
1 Background and aims	1	Modelling the impact of CAD on the screening programme	37
Mammography and breast cancer screening	1	Economic assessment	37
Non-radiologist film readers	2	Usability of CAD	39
Computer aids for mammography	2	Evaluation of rapidly changing technology	40
Assessing CAD for the NHSBSP	7	The prospects for CAD	40
2 Methods and materials	9	Conclusions and recommendations for future research	41
Study 1	9	Acknowledgements	43
Study 2	11	References	45
3 Results	13	Appendix 1 Data collection form	49
Study 1	13	Appendix 2 Model for a test of reading accuracy	51
Study 2	17	Appendix 3 Modelling the cost-effectiveness of computer prompting	53
4 Economic assessment	21	Health Technology Assessment reports published to date	59
Introduction	21	Health Technology Assessment Programme	69
Previous research on cost-effectiveness of breast screening	21		
Methods	22		
Sensitivity analysis and presentation of results	28		
Results	28		
Cost per cancer detected	29		
Cost per life-year gained and cost per QALY	29		
5 Discussion	33		
Impact on readers' sensitivity and specificity	33		



Glossary and list of abbreviations

Technical terms and abbreviations are used throughout this report. The meaning is usually clear from the context, but a glossary is provided for the non-specialist reader. In some cases, usage differs in the literature, but the term has a constant meaning throughout this report.

Glossary

Arbitration The use of a third film reader to assess disagreements between film readers working together in double-reading.

Asymmetry A sign of cancer in which a small area of dense tissue is visible in one breast and not the other.

Biopsy An invasive procedure to collect cells to make a definitive diagnosis of cancer.

Bootstrapping A statistical technique whereby random subsamples of a population are used to create a sampling distribution that can be used to determine a confidence interval.

Breast clinician A physician specialising in breast disease, usually employed as a non-radiologist film reader.

Craniocaudal mammogram A top-down projection image of the breast.

Digitisation Scanning an analogue film to make a digital image.

Double-reading A screening protocol in which all films are viewed separately by two film readers.

Ductal carcinoma *in situ* A non-invasive lesion having some characteristics of cancer and which requires treatment.

False negative A cancer missed at screening.

False negative minimal signs A cancer missed at screening which was perhaps only visible with the benefit of hindsight.

False positive A normal case incorrectly identified as cancer.

False-positive-rate The number of false-positive prompts per case.

Film reader A health professional who reads screening mammograms, normally a consultant radiologist but sometimes a trained radiographer or a breast clinician.

Full-field digital mammography The direct acquisition of a digital X-ray image without the need for film.

ImageChecker[®] A computer-aided detection/diagnostic system developed by R2 Technology.

Interval cancer A cancer that presents clinically between screening rounds.

Logit function A function that transforms a variable (such as a probability) that is constrained to lie within a certain range to take values that will be normally distributed.

Markov model A technique used to estimate the number of events affecting a cohort of patients over a period, using data about frequency of events in specified states and the probabilities of transitions between states.

Mass One of the common abnormalities observed on mammograms.

Mediolateral oblique mammogram A sideways oblique projection image of the breast.

Microcalcification One of the common abnormalities observed on mammograms.

Nottingham Prognostic Index A formula used to calculate the risk of death associated with a cancer based on a number of facts about the cancer and the patient.

continued

Glossary continued

Positive predictive value The number of correctly identified cancer cases as a proportion of the total number of cases identified as cancer.

Prognostic group A set of cancers associated with a particular prognosis.

Quality-adjusted life-year A unit used in economic assessment to quantify the health gain from an intervention.

Radiographer A healthcare professional with a specialist training in the acquisition of radiological images. Radiographers have not traditionally been involved in the interpretation of images, but are being used as film readers in breast screening.

Radiologist A medical doctor with training and a qualification in radiology.

Receiver operating characteristic analysis A method of comparing diagnostic tests in which sensitivity is plotted against specificity for a range of decision points.

Rollers The equipment used to allow film readers to view large numbers of films.

Sensitivity The number of correctly identified cancer cases as a proportion of the total number of cancer cases.

Specificity The number of correctly identified normal cases as a proportion of the total number of normal cases.

Stellate lesions One of the common abnormalities observed on mammograms.

True positive A cancer correctly identified as such.

True-positive fraction The number of true positives expressed as a proportion of the total number of cancers.

List of abbreviations

ANOVA	analysis of variance	NA	not applicable
BCD	breast cancer diagnosed	NHSBSP	National Health Service Breast Screening Programme
CAD	computer-aided detection or diagnosis	NPI	Nottingham Prognostic Index
CI	confidence interval	PG	prognostic group
DCIS	ductal carcinoma <i>in situ</i>	QALY	quality-adjusted life-year
DR	distant recurrence	RCT	randomised controlled trial
EAC	equivalent annual cost	ROC	receiver operating characteristic
FDA	Food and Drug Administration	RR	regional recurrence
FFDM	full-field digital mammography	SD	standard deviation
FNA	fine-needle aspiration	SMF	Standard Mammogram Format
GLM	generalised linear model	UCL	University College London
LR	local recurrence		

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.



Executive summary

Background

This report describes two studies carried out in order to assess the potential role of computer aids in the UK NHS Breast Screening Programme (NHSBSP).

Objectives

The objective was to determine the value of computer-aided detection (CAD) for breast cancer screening. The impact of the R2 ImageChecker[®] on the sensitivity and specificity of radiologists and film-reading radiographers was assessed in two experiments, referred to here as study 1 and study 2, and the resulting data were used in an economic evaluation.

Methods

Design

Two sets of mammograms with known outcomes were used. Participants in both studies read the films with and without the benefit of the computer aid. In both studies, the order of reading sessions was randomised separately for each reader. The first set of 180 films, used in study 1, included 20 false-negative interval cancers and 40 screen-detected cancers. The second set of 120 films, used in study 2, was designed to be favourable to CAD: all 44 cancer cases had previously been missed by a film reader and cancers prompted by CAD were preferentially included.

Setting

The studies were conducted at five screening centres: South-West London, Norfolk and Norwich, Luton and Dunstable, Worthing, and Bristol. Study 1 was conducted between January 2001 and July 2002, and study 2 between September 2002 and April 2003.

Participants

Thirty radiologists, five breast clinicians and 15 radiographers participated.

Interventions

All cases in the trial were digitised and analysed using the R2 ImageChecker version 2.2. Participants all received training on the use of CAD. In the intervention condition participants interpreted cases with a prompt sheet on which regions of potential abnormality were indicated.

Main outcome measures

The sensitivity and specificity of participants were measured in both intervention and control conditions.

Results

No significant difference was found for readers' sensitivity or specificity between the prompted and unprompted conditions in study 1 [95% confidence index (CI) for sensitivity with and without CAD is 0.76 to 0.80, for specificity it is 0.81 to 0.86 without CAD and 0.81 to 0.87 with CAD]. No statistically significant difference was found between the sensitivity and specificity of the different groups of film reader (95% CI for unprompted sensitivity of radiologists was 0.75 to 0.81, for radiographers it was 0.71 to 0.81, prompted sensitivity was 0.76 to 0.81 for radiologists and 0.69 to 0.79 for radiographers). Thirty-five readers participated in study 2. Sensitivity was improved in the prompted condition (0.81 from 0.78) but the difference was slightly below the threshold for statistical significance (95% CI for the difference -0.003 to 0.064). Specificity also improved (0.87 from 0.86); again, the difference was not significant at 0.05 (95% CI -0.003 to 0.034). A cost-effectiveness analysis was performed based on data from studies 1 and 2. The analysis showed that computer prompting is cost-increasing.

Conclusions and recommendations for research

No significant improvement in film readers' sensitivity or specificity or gain in cost-effectiveness was established in either study. This may be due to

the system's low specificity, its relatively poor sensitivity for subtle cancers and the fact the prompts cannot serve as aids to decision-making. It may be that readers would be better able to make use of the prompts if they had longer to become accustomed to working with them. Prompts may have an impact in routine use that is not detectable in an experimental setting.

Although the case for CAD as an element of the NHSBSP is not made here, further research is required. Evaluations of new CAD tools in routine use are underway and their results should be given careful attention.

There should be a clearer and speedier route to commissioning evaluations of rapidly changing technologies.

Chapter I

Background and aims

This report describes a series of studies carried out as part of the investigation on behalf of the NHS HTA board to assess the potential role of computer aids in the UK NHS Breast Screening Programme (NHSBSP). This chapter describes the context. It starts by giving a brief overview of the role of mammography in the breast cancer screening programme and outlines some of the challenges facing the programme. The potential role of computer aids is then introduced, with a brief survey of the state of the art in computer-aided detection and diagnosis in mammography.

Mammography and breast cancer screening

The UK NHSBSP was set up in 1987 following the recommendations of the Forrest Report.¹ From the start it was clear that the programme was a major exercise in public health, one that would demand considerable human and financial resources as well as the cooperation of women in the screening age group. Fifteen years on, the programme is generally regarded as a success. The programme in the UK has screened more than 14 million women and has detected over 80,000 cancers.² It was estimated in 2000 that the programme was saving at least 300 lives per year. That figure is set to rise to 1250 by 2010.³

Through the screening programme women between the ages of 50 and 64 years are invited every 3 years to attend for a screening appointment. At the appointment, which takes place either at a screening centre or in a mobile van, X-ray images known as mammograms are taken. Mammograms are projection images and it is now the policy, at screening, to take two views of each breast: a mediolateral oblique and a craniocaudal view. The films are processed at the screening centre and then loaded onto roller viewers for interpretation. Unless the woman is attending for her first screening visit, the films from the previous visit are also loaded onto the viewer, to allow any changes in breast appearance to be assessed. In almost all centres, the policy is that the films are double-read; that is to say, they are read separately by two different professionals, normally consultant radiologists. There are

different ways of doing double reading, but the gold standard would be for each reader to identify films that warrant discussion at a consensus or arbitration meeting, where the assessment of a third radiologist can be called upon. The decision made here is whether or not to recall the women for further investigation. A small percentage of women are recalled for technical reasons, if it is felt that the mammogram is not of adequate quality to allow a safe decision to be made.

Women who are recalled are seen at assessment clinics. Whereas the interpretation of screening mammograms is a quick and simple task, normally completed in a matter of seconds rather than minutes, the assessment process is a much more costly one, involving further investigations that include clinical examination, additional mammography, ultrasound, and needle and core biopsies. The pressure on assessment clinics is a critical problem for screening centres. If the waiting lists for assessment grow too long, the national programme will suspend screening at the centre, which inevitably attracts adverse attention, often reported in the national media.

In 2001, 1,361,881 women were screened and 9866 cancers detected through the programme.⁴ Official figures for missed cancers are not published, but can be estimated by assuming that missed cancers will either present clinically (e.g. because the patient detects a lump) before the next screening visit, or be detected at the next screening visit. Blinded review of the previous screening films for both interval cancers and for cancers detected at the screening visits other than the first suggests that as many as 25% of radiographically visible cancers are missed at screening.⁵

The core of the screening programme, then, is the assessment of mammograms by trained film readers, normally consultant radiologists. The task is a difficult one. The architecture of breast tissue is variable and mammograms have a noisy, unstructured appearance. Cancer is usually not revealed directly; rather, its presence must be inferred from the assessment of a number of signs, which are all too often ambiguous. The overwhelming majority of films will be normal

(cancer detection rates are around six per 1000 cases). The detection of cancers therefore requires a large number of readers with good perceptual skills, judgement and vigilance.

The NHSBSP is, however, facing an acute staffing crisis. Regular news stories surface in the media about staff shortages contributing to failures in screening.⁶ A recent survey by Field found that 61% of units are staffed by no more than one or two consultants.⁷ These units could never provide a comprehensive double reading service using radiologists alone. Field found a 50% failure rate in the recruitment of consultants to screening posts, with no applications at all received for four posts advertised in the *British Medical Journal*. If the situation is bad now, it is set to get much worse. The cohort currently entering the age range covered by the screening programme is that of the post-war 'baby boom', with the result that the number of women eligible for screening is set to increase by 16% over the next 10 years. In addition, the NHS Plan announced that the government intends to extend routine invitations to women up to the age of 70 years and states that all women will have two views of the breast taken at every screening.⁸ This latter improvement began to be introduced across the UK during 2003. These changes will increase the demand for film readers. The Cancer Plan reports on initiatives to increase the number of radiologists and radiographers working in the NHSBSP, but makes it clear that it is also essential that alternative working practices be introduced; the most significant change currently being piloted is the use of trained radiographers as film readers.⁹

Non-radiologist film readers

Two groups of non-radiologists are currently employed as film readers in the breast screening programme: radiographers and breast clinicians. Radiographers are being trained as film readers in a number of centres in the UK. Trained radiographers are currently working as film readers in at least six screening centres. The Association of Breast Clinicians was founded in 1996 and currently has 53 members, all of whom are registered medical practitioners working in breast diagnostics.

Haiart and Henderson compared the sensitivities and specificities of a radiologist, a radiographer and a clinician, and found similar sensitivities but greatly inferior specificity for the non-radiologists.¹⁰ Pauli and colleagues studied a

group of seven trained radiographers working as second readers. They found that the increase in sensitivity due to the second reader was 6.4%.¹¹ This was reported as being equivalent to the increase found for second reading by radiologists. Cowley and Gale provide data on the comparative performance of the different professional groups on PERFORMS, a voluntary self-assessment exercise.¹² The results show that untrained radiographers have some interpretative skills and that trained radiographers perform as well as radiologists. The performance of breast clinicians was slightly better than that of radiologists in early rounds of assessment and similar in the most recent round.

The available data, therefore, seem to suggest that both breast clinicians and radiographers can be trained to perform as well as radiologists. The samples studied are, however, relatively small and likely to be highly selected (nine trained radiographers and ten breast clinicians participated). It may be that only a limited number of radiographers feel comfortable with the additional responsibility associated with film reading. Attempts to involve large numbers of radiographers in the interpretation of films may fail to maintain the performance achieved in the studies carried out to date.

It seems likely that an increased reliance on radiographers will be an essential component of any solution to the staffing crisis. There are, however, outstanding questions concerning the use of non-radiologist film readers and it would be wrong to assume that using non-radiologist film readers can solve the crisis by itself. Attempts to encourage large numbers of radiographers to move into film reading could generate a shortage of radiographers to take the films. It therefore seems likely that technical solutions, such as computer aids, will have to be considered. Nevertheless, any assessment of the impact of technology on film readers must take into account the fact that in the future many film readers will not be radiologists.

Computer aids for mammography

The computer analysis of mammograms has been a field of research since the 1960s and the recent advent of direct digital mammography (X-ray machines that generate a digital mammogram without the need for an image to be produced on film) has generated increased activity.¹³ The most important work in the computer analysis of

mammograms has been in the detection of two important classes of mammographic abnormality: microcalcifications and masses. A great deal of effort has been put into the development of these techniques and they can be made extremely sensitive. However, this sensitivity is achieved only at the cost of specificity, with high numbers of false prompts. The interpretation of mammograms cannot, at the moment, be completely automated; this remains a distant prospect. Some authors, however, have suggested that the combination of computer plus human reader can produce better results than a human reader alone, and have concluded that computer-aided detection (CAD) systems can be used to prompt human readers.¹⁴ Recently, authors have suggested that CAD systems could be used as an alternative to second reading in programmes that are currently using two human film readers to look at each case.¹⁵

A CAD system is used to alert a human film reader to regions of a mammogram where computerised analysis suggests that abnormalities may be found. Most systems are currently delivered to work with analogue films. Such systems include a scanner that creates a digital image of each film. The digital images are then analysed by the software that places the prompts. The prompts can be displayed either on paper or on computer monitors placed next to a roller viewer. The radiologist (or other film reader) views the image, checks the CAD display or prompt sheet and, if appropriate, reassesses the image. CAD systems have been around in research settings for over a decade, and in clinical use for a little less. The US Federal Government has, since 2001, reimbursed radiologists who use CAD, an initiative that provided a massive stimulus to the market for CAD in the USA: R2 (the current market leader) now claims an installed base in the USA of over 1000 systems.¹⁶ Uptake outside the USA has been slower.

Many authors have assumed that CAD will only become commonplace once full-field digital mammography (FFDM) is the norm. That is, however, not an immediate prospect. FFDM machines are roughly five or six times as expensive as conventional analogue machines and many centres, including those that make up the NHSBSP, are still choosing to invest in analogue machines that will not be replaced for 10 years or more.

The following three subsections give a brief overview of how CAD works, and summaries of the

four leading CAD providers and of the available evidence on the impact of CAD.

CAD algorithms

The algorithms used by the suppliers of CAD systems are all a matter of commercial confidence, so there is little point in reviewing the published literature in this report. It is, however, appropriate to make a number of general observations.

There are two broad classes of approach to the analysis of digital mammograms. One, the physics-based approach, attempts to understand the physical processes that underlie the image formation processes and to build a mathematical model of the relationship between pixel value and breast anatomy. Such a model provides a sound mathematical basis for future processing. The best known example of a physics-based approach is the Standard Mammogram Format (SMF) developed by Mirada Solutions. SMF allows mammograms to be normalised, removing variation due to differences in the image acquisition process or between different mammography devices. SMF has been used in a number of CAD algorithms developed in academia, but the authors are not aware of any commercial CAD tools that incorporate it.¹⁷

In contrast, most of the successful algorithms for detecting abnormalities use purely statistical approaches. Probabilistic calculations are performed to assess the likelihood that a particular image region contains an abnormality. Different detectors are developed for the different forms of abnormality, most notably microcalcifications, stellate lesions and well-defined masses. For the most part these algorithms analyse each image separately. Algorithms that detect left-right asymmetry or temporal change have been developed, but these are generally still regarded as unsolved problems.

R2 Technology provides a guide to the algorithms used in the R2 ImageChecker®.¹⁸ This guide, however, does not give any technical details, but rather provides users with explanations of what the algorithms are looking for and therefore of where they may go wrong. The microcalcification detector is said to place prompts only if at least three calcifications have been detected and if they are no more than 2.5 mm apart. The prompt is placed on the centroid of the cluster. The mass detector uses a ring with an outer diameter of 32 mm and an inner diameter of 6 mm. The ring is moved over the image, and lines detected within

the ring, if appearing to radiate from the centre, lead to the placing of a mass prompt. Pronounced radiating lines will be prompted whether or not there is a central mass. Less pronounced radiating lines will be prompted only if associated with a central mass. The software is optimised to search for lesions of 10–20 mm in diameter. Small and larger lesions may be missed. The manufacturers suggest a continuum for masses that is related to the likelihood that the ImageChecker will spot them.

Current CAD providers

R2 Technology

R2 Technology was founded in 1993 and is based in California, USA.¹⁹ Their tool, the ImageChecker System, was the first US Food and Drug Administration (FDA)-approved CAD system for breast imaging. Over 550 film-based ImageChecker systems and more than 150 R2 CAD systems for digital mammography have been shipped. Over 5 million mammograms have been interpreted with ImageChecker assistance. The company is also bringing a lung nodule detection tool to market. The R2 ImageChecker uses a Canon digitiser with a resolution of under 50 μm and a depth of 12 bits per pixel. Up to 50 films can be loaded into the digitiser at once, taking around 1 hour to process. The prompts are then displayed on a liquid crystal display (LCD) screen placed next to the roller viewer.

CADx

CADx merged with Qualia Computing in September 2002 to create the second largest CAD provider.²⁰ One hundred CADx SecondLook systems have been shipped since gaining FDA approval in January 2002. The company, which is based in Ohio, now aims to bring out new tools, including SecondLook CT Lung and SecondLook CT Colon. CADx SecondLook uses a modified Howtek digitiser with a 43- μm pixel size. The loader and processing speed are comparable to R2. Prompts are displayed on a paper sheet.

CADVision

CADVision was founded in 1993 and is based in Jerusalem, Israel.²¹ CADVision has two products, a prompting system similar to that of the other providers, and a diagnostic system that classifies breast abnormalities and ranks them according to their likelihood of malignancy. The company has received the Conformité Européenne (CE) Mark (European regulatory approval) for its software and expects to receive FDA certification (US regulatory approval). The company and its

scientific and clinical partners have regularly published articles about their algorithms and the performance of their systems.²¹

iCAD

This company received approval from the US FDA to market its MammoReader™ systems in the USA in January 2002.²² The authors are not aware of any published research about this system. iCAD is the only vertically integrated company in its market; it also manufactures medical film digitisers for a variety of medical imaging and other applications. Established in 1984, iCAD has sold over 30,000 quality-imaging systems worldwide. The company is based in New Hampshire, but its principal research and development facilities are in Florida.

Evidence for CAD in mammography

There are three forms of evidence about CAD. The simplest kind of experiment one can do is to use CAD to analyse films with known outcomes and then present data on the number of cancers detected and the false prompt rate. These experiments reveal something about the sophistication of the algorithms used. More useful information is provided by experiments that assess the impact of CAD on radiologists' decision-making. Such experiments normally involve taking small test sets of films with known outcomes, presenting them to radiologists with and without CAD, and measuring the radiologists' sensitivity and specificity with and without CAD. These evaluations are necessarily performed under test conditions, usually with sets of images that have artificially high frequencies of cancer. The most valuable investigation is therefore an assessment of CAD in routine use. Very few such evaluations have, however, been performed. The next three subsections review the existing evidence from the three kinds of study.

Tests of the systems' sensitivity and false prompt rate

The largest evaluation of the sensitivity of CAD was performed as part of the evaluation carried out by R2 Technology to gain FDA approval. They tested the system on 1083 cases of biopsy-proven cancer collected from 13 centres.²³ Updated algorithms were subsequently tested on the same images.²⁴ Data from both evaluations are presented in *Table 1*.

Version 1.2 of the software generated an average of 4.1 false prompts per four-view case. In version 2.2 this was reduced to 1.0 markers per four-view case, on normal cases.²⁴

TABLE 1 Sensitivity of the R2 ImageChecker versions 1.2 and 2.2

	No. of cases	No. correctly marked, 1997 (v1.2)	% correctly marked, 1997 (v1.2)	No. correctly marked, 2000 (v2.2)	% correctly marked, 2000 (v2.2)
Microcalcifications	404	396	98%	399	99%
Masses	679	507	75%	580	85%
Total	1083	903	83%	979	90%

A comparable study was performed to obtain FDA approval for SecondLook.²⁵ In total, 906 mammograms of biopsy-proven cancer were obtained from 17 institutions. The system correctly marked the cancer in 809 cases (89%). The sensitivity for clustered microcalcifications was 95% (280/296) and 87% (529/610) for masses.

The performance of the two main suppliers is very similar and although tests have been done comparing the systems on the same set of images, the likelihood of detecting a statistically significant difference must be low. One study compared ImageChecker, CADx SecondLook and a third system, MammoReader.²⁶ Altogether, 120 biopsy-proven cancer cases were processed through three CAD systems, ImageChecker, MammoReader and SecondLook. The case-level sensitivity was very similar for the three systems (Table 2).

Several further studies have been performed to assess the sensitivity of CAD systems to particular categories of cancers. Warren Burhenne and colleagues obtained the prior films for 427 of the 1083 cases used in the study described above.²³ They present data for the performance of R2 on

TABLE 2 Sensitivity of the three CAD systems that currently have FDA approval

System	Sensitivity
ImageChecker	91%
SecondLook	89%
MammoReader	91%

TABLE 3 Sensitivity of the R2 ImageChecker to missed cancers according to the consensus on the actionability of abnormality on the prior

Consensus on actionability	No. of cases	No. correctly marked, 1997 (v1.2)	% correctly marked, 1997 (v1.2)	No. correctly marked, 2000 (v2.2)	% correctly marked, 2000 (v2.2)
3/5, 4/5, 5/5	112	91	81%	101	90%
4/5, 5/5	74	63	85%	70	95%
5/5	36	33	92%	33	92%

those priors where a majority of five radiologists agreed, at a blinded review, that an actionable finding was visible. The ImageChecker correctly marked 89 out of 115 (77%) of these missed cancers, 30 out of 35 (86%) missed calcifications and 58 out of 80 (73%) missed masses. Masses, where CAD performs less well, are responsible for many more missed cancers than calcifications. Table 3 shows these data, along with data on the performance of the later release.²⁴

CADx quotes similar results for a study of 121 cancers classed as actionable missed cancers, of these actionable cases, at least 86 (71%) were prompted by SecondLook.²⁵

Destounis and colleagues identified 52 actionable prior mammograms that had been missed by consensus double-reading.²⁷ Only eight of these were calcifications, seven of which were marked by the ImageChecker, whereas 19 (65%) of the masses were marked.

There is, therefore, clear evidence that CAD algorithms are extremely sensitive even for the detection of relatively subtle abnormalities. The evidence that they are effective in improving radiologists' decision-making is, however, less strong.

Tests of the impact of systems on radiologists' decision-making

A large number of studies has now been performed to assess the impact of prompts on radiologists' decision-making. Most of these

studies were not prospective evaluations of the prompting systems, but tests using archive cases with known outcomes selected to contain large numbers of cancers.

Thurfjell and colleagues carried out a study with three film readers: an expert screener, a screening radiologist and a clinical radiologist.²⁸ The expert's sensitivity of 86% was unchanged, but use of the ImageChecker improved that of the screening radiologist from 80% to 84% and that of the clinical radiologist from 67% to 75%. The specificities of the expert and of the clinical radiologist were unchanged, while that of the screening radiologist fell from 83% to 80%. Funovics and colleagues tested the system using a test set including 40 proven spiculated lesions and three radiologists. They found an average improvement in sensitivity of 9%, with some cost in specificity.²⁹ Moberg and colleagues evaluated the R2 ImageChecker in a study in which three radiologists looked at a test set including 59 interval cancers. They found no significant change in sensitivity or specificity.³⁰

Marx and colleagues conducted a study in which five radiologists viewed a test set containing 36 cancers with and without CADx SecondLook. They found sensitivities of 80.6% and 80.0% without and with CAD, and specificities of 83.2% and 86.4%, respectively.³¹

The largest reported study of sensitivity using the R2 ImageChecker is that of Brem and Schoonjans, who used a sample of 106 cases including 42 malignant microcalcifications, 40 with benign microcalcifications and 24 normals.³² Five radiologists participated. Forty-one out of 42 (98%) malignant microcalcifications and 32 out of 40 (80%) benign microcalcifications were prompted at a prompt rate of 1.2 markers per image. The radiologists' sensitivity without and with the system ranged from 81 to 98% and from 88 to 98%, respectively. No statistically significant changes in sensitivity were found and there was no significant compromise in specificity.

Ciatto and colleagues present results of a study in which ten radiologists used CADx SecondLook to review a test set containing 17 cancers. They found no evidence of an increase in sensitivity, but found evidence of a loss of specificity.¹⁵

The results of these studies are disappointing. It has not seemed possible to show conclusively that CAD does allow for an improvement in sensitivity.

It is worth noting that the studies listed above all attempt to show an improvement due to the prompts by studying sensitivity at a particular threshold. Several of them are small and likely to be inadequately powered. Studies carried out in academia using research systems have shown an improvement in sensitivity if measured using receiver operating characteristic (ROC) analysis across a range of thresholds.³³

Various techniques are used to analyse the impact of diagnostic information on film readers' decision-making. The most sophisticated approach is to analyse the ROC. The data required in ROC analysis is a measure of the confidence that the film reader has in each decision to recall. Each level of confidence is used as a threshold to separate recalls from non-recalls. The set of thresholds is then used to create a set of points on a plot of false-positive rate against false-negative rate. The best fit curve between the points is then taken as a measure of the quality of decision-making that the diagnostic information enables. Two conditions (such as prompted and unprompted reading) can be compared by comparing the area under the curves for the two conditions.

One could take the view that ROC analysis is a more sensitive technique, and therefore better able to show an improvement; or it may be that the improvement at the clinically important operating point is relatively minor, and that this is obscured in ROC analysis.

Prospective studies of radiologists' sensitivity and specificity

No one has yet attempted a randomised controlled trial (RCT) of CAD. Freer and Ullissey conducted a 12-month prospective evaluation of the impact of CAD using the R2 ImageChecker.³⁴ The trial involved 12,860 screening mammograms. Each was initially interpreted without CAD. Areas marked by the CAD system were then re-evaluated. Data were recorded both before and after the CAD prompts were consulted. The authors report an increase in recall rate from 6.5 to 7.7% with the use of CAD and an increase from 3.2 to 3.8/1000 in the cancer detection rate. A study by Morton and colleagues looked at 12,646 patients, again with the R2 ImageChecker. A comparison of outcomes for interpretations made without and with CAD assistance revealed a 6.64% increase in breast cancer detection rate and an increase in screening recall rate from 9.82% to 10.89%.³⁵

Young and colleagues carried out a similar study on 12,082 mammograms that were each read by two radiologists. The recall rate for two radiologists was 21%. The recall rate increased to 24% when using CAD. The overall accuracy of CAD in marking cancers was 73%. The overall accuracy of the two radiologists was 94%. They concluded that two radiologists detect more cancers than one radiologist and CAD.³⁶

Warren Burhenne and colleagues prospectively studied the recall rates of 14 radiologists using CAD. These radiologists had previously had a recall rate of 8.3% (assessed from historical data on 23,682 cases) and using CAD for 14,817 had a recall rate of 7.6%, suggesting that there is no significant impact on specificity due to CAD. The design of this study meant that it was unable to report on sensitivity.²³

Assessing the evidence

The key issue in evaluating the above evidence on decision-making is the contrast between the somewhat equivocal results of studies using test sets drawn from archive data, with a small number of prospective studies. The best known of these, that of Freer and Ulissey, found an increase in the number of cancers detected, although at a non-negligible cost in terms of increased recall rate.³⁴ It should also be noted that 75% of the additional cancers were ductal carcinoma *in situ* (DCIS) rather than invasive cancers. The study could also be questioned because it compares a judgement made on the radiologist's first look with a judgement made on the basis of the first and second looks, where CAD is available on second look. One might expect that the performance of the radiologists on the first look would be slightly compromised because they would have been anticipating the second look; further, the increment due to the second look cannot be entirely attributed to the availability of CAD.

The results of the three prospective trials are somewhat different. By their nature they are difficult and time-consuming to set up and involve only small numbers of readers. There are substantial performance differences between radiologists, and there are likely to be substantial differences in the use that they make of prompts. It will therefore be necessary to review the results of large numbers of such trials before a clear conclusion can be drawn about the impact of prompts. The trial by Freer and Ulissey found a large increase;³⁴ the increases in the other two trials suggest that the impact will not allow prompts to replace human second readers.^{35,36}

Assessing CAD for the NHSBSP

The research described in this report was carried out by a team of academics and clinicians based at University College London (UCL), St George's NHS Trust and the University of Oxford, UK. The work was led by researchers at UCL, who were responsible for the design of the evaluation, the data collection and overall management. The senior clinical partner, Dr Rosalind Given-Wilson, was based at St George's NHS Trust and provided clinical input and advice as well as the images used in the evaluation. The CAD system was installed at St George's for the duration of the project. Statistical advice was provided by Dr Henry Potts, who was based at the School of Public Policy at UCL. The economic evaluation was performed by Dr Katharine Johnston, who at the time was working at the University of Oxford.

Looking at the available evidence on the use of the computer aids and on film-reading radiographers, it was decided from the start that there was a need to examine the potential for the use of computer aids as part of a screening programme employing radiographers as film readers. Therefore, a study was conducted to ascertain whether (1) computer aids could be used in the role of second readers, allowing a single radiologist to achieve the accuracy currently achieved by double-reading, and (2) whether a radiographer using a computer aid performed as well as a radiologist.

The authors took the view, at the start of the project, that at that time the available evidence about CAD could not justify the setting up of an RCT. Such a trial would have had to elicit consent from 60,000 women in order to have two arms with 180 cancers in each. It would have been unable to publish definitive results until the women had returned for a subsequent screening 3 years later. It would have required resources that, in the authors' view, could not be justified. The research team considered carrying out a prospective study using the design later followed by Freer and Ulissey, but was advised that this would not be ethical without seeking patient consent, which would have posed serious logistical and administrative difficulties.

The researchers also wanted to include a large sample of different film readers in the study and therefore elected to carry out a test of the impact of prompts on decision-making. The study used archive films with a high proportion of cancers

that would be read by volunteer film readers in experimental conditions. The study was designed to test the hypothesis that R2 would improve sensitivity at no cost in specificity. The intention was to perform a single study and use the data collected in an economic evaluation. This led to one of the key decisions made in the design of the study, to analyse the data to detect changes in sensitivity and specificity at a clinically important decision threshold.

ROC analysis was not used because of concern that the area between the two curves might reflect the impact of the prompts on the film readers' reported confidence in their decisions at the margins and not a clinically significant impact on the proportion of women that they actually would

recall. Therefore, a relatively large study was designed using many more readers than previous studies, who were required to look at a fairly large number of films.

The researchers decided to use the R2 ImageChecker, which was then, as now, the clear market leader, as the CAD tool to assess. The data from the study were, however, inconclusive, so a second study was performed, using a more selected test set of cancers. These two studies are described in Chapter 2 under the headings Study 1 and Study 2. The results appear in Chapter 3. The economic evaluation, which draws on both sets of results, is presented in Chapter 4. Chapter 5 presents a discussion of the results and the conclusions that can be drawn from it.

Chapter 2

Methods and materials

This chapter describes the two studies carried out as part of the investigation commissioned by the HTA programme. Study 1 was carried out between March 2001 and July 2002. The results are presented in Chapter 3. Study 2 was designed to answer questions raised by the analysis of the results of study 1. It was carried out between September 2002 and March 2003. The results are also presented in Chapter 3.

Study 1

The first of the studies performed was designed as a test of the impact of the computer-placed prompts on the sensitivity and specificity of radiologists and radiographers reading test rollers containing an artificially high proportion of cancer cases. The aim was to test the hypothesis that sensitivity would be improved by the prompts and to provide data that could be used to assess the potential value of prompts.

Films

The sample contained 180 cases. The films were taken from the South-West London Breast Screening Service. The original films were used in the study. All were films of women aged 50–64 years undergoing routine breast screening. The South-West London Breast Screening Service uses two views and double-reading with arbitration. All cases had a proven outcome, being either a malignant result at biopsy or, for controls, a normal result at this round and a subsequent round of screening. Cases where a suspicious appearance proved negative at biopsy were excluded. The sample included 60 cancers: 40 consecutive cancers detected through routine screening and 20 interval cancers. A breakdown of cancer cases by primary radiological abnormality is given in *Table 4*. The interval cancers had been reviewed by a panel of radiologists and classified as false negatives. Controls were selected from the same period as the screen-detected cancers.

The mammograms were divided into three sets of 60 cases, each to be interpreted at a single sitting. The ratio of cancer to non-cancer cases was varied slightly across the three sets. An additional set of 60 cases was used as a training set to accustom

TABLE 4 Breakdown of cancer cases by primary radiological abnormality

Primary abnormality	No. of cases
Round mass	5
Ill-defined mass	7
Stellate/spiculated lesion	22
Asymmetry	11
Microcalcification	15

film readers to using the CAD system. All films were temporarily anonymised and assigned a study number.

Prompts

The films were processed using the R2 ImageChecker M1000. This uses a Canon digitiser with 50- μ m resolution and 12 bits per pixel. The films are digitised using a bulk loader that allows approximately 12 four-view cases to be digitised in a single batch. Each film takes around 4 minutes to process. The digitised films are analysed by CAD algorithms for the detection of masses and calcifications. Prompts are placed on areas where the algorithm suggests that a mass or calcification may be present. Emphasised prompts are placed on regions that elicit a particularly strong response from the algorithms. The prompts were printed on sheets of A4 paper, with four low-resolution images shown in a row across the top half of the sheet. An example is shown in Appendix 1.

Readers

Fifty film readers participated: 30 consultant radiologists, five breast clinicians and 15 trained radiographers. All had undergone a rigorous training programme to meet the requirements of NHSBSP. All were currently working in the screening programme and reading at least 5000 screening cases per annum.

Procedure

The study was conducted at five screening centres: South-West London, Norfolk and Norwich, Luton and Dunstable, Worthing, and Bristol.

All film readers were given training including an explanation of prompting and of the behaviour of

the system. They were told that they should look at the films in the normal way, before looking at the prompts. They were advised on the typical frequency of prompts and given examples of cases where CAD often generates inappropriate prompts. All film readers then read a training roller, using the prompts. This roller contained a similar mix of cancers and non-cancers to the subsequent test rollers.

The order in which each reader viewed the test sets was separately randomised, as was the order in which they viewed the conditions (prompted and unprompted). A minimum period of 1 week was left between reading a test set in the two conditions.

Readers viewed the films on a standard roller viewer in the normal viewing conditions pertaining at the centre. Viewing conditions were the same for the prompted and unprompted conditions. Readers were asked to complete a report form for each case. An example form is included in Appendix 1. Low-resolution images of the films were included on the form. In the prompted condition these were shown with prompts; in the unprompted condition they were shown without prompts. The reader was asked to circle any area of abnormality and state their degree of suspicion for each abnormality in a table below the images. Readers were then asked to give an overall decision on recall for each case. The time taken to read each test set was recorded.

All of the cancer cases were reviewed by a consultant radiologist (RGW), who indicated the extent of visible signs of cancer on each image. Two members of the team (PT and JC) compared these annotations with the prompt sheets generated by the ImageChecker to ascertain whether or not the cancer was correctly prompted. A case was considered to have been correctly prompted if a prompt appeared in either view at a location indicated by the radiologist. Three borderline cases were reviewed by RGW.

Sample size calculation

Performing power calculations for film-reading studies is difficult, and for complex designs involving multiple readers and multiple observations, analytical expressions for sample size become intractable. The researchers wrote a computer program based on a mathematical model of the proposed design. This program allowed them to run simulations based on different sample sizes under different assumptions in order to determine study power. Details of the

model are provided in Appendix 2. A fuller account of the approach is provided by Pepe and colleagues.³⁷ The model requires the following parameter values: $S_{\text{pre}}^{\text{D}}$, sensitivity of the average reader on the average film; $S_{\text{post}}^{\text{D}}$, sensitivity of the average reader on the average film, with CAD; a^{D} , variation in sensitivity due to the varying difficulty of films; b^{D} , variation in sensitivity due to the varying ability of readers; R , number of readers; and I^{D} , number of disease cases.

A value for b^{D} was taken from a survey by Beam and colleagues, which suggested that the sensitivity of film readers can differ from the mean by up to 20%.³⁸ There are few data on the varying difficulty of images, so the figure for a^{D} was also set, somewhat arbitrarily, at 20%. The mean sensitivity of film readers in experimental conditions depends on the difficulty of the test set, and on the reading protocol. The model with $S_{\text{pre}}^{\text{D}}$ was set to 70% and 80%. The difference between intervention and control conditions was set at 10%, which roughly corresponds to the effect of CAD found by Funovics and colleagues.²⁹ A realistic maximum number of images for a film reader to interpret in a study is 300 and it was felt that at least two normals were required for every cancer in the set; therefore, the model was run with values of 60 and 90 for I^{D} .

A particular difficulty for this study was recruiting a sufficient number of non-radiologist film readers, and the model was run using values of 7 and 10 for R . Where one of the conditions being considered is double-reading, the number of readers required is double R . Results are presented in Table 5. Four-hundred simulations were obtained with each setting of parameter values. The value given for power is the proportion of simulations where the result was significant at the 0.05 level.

Given the above calculations, it was assumed that ten readers will be required for each group to be compared, but that 60 cancers will be adequate. The analysis assumes paired observations. So, for example, if a figure for 'Normal sensitivity' is obtained by having images double-read by radiologist X and radiologist Y, and a figure for 'Sensitivity with CAD' is obtained by having radiologist X read films with CAD, the unit of analysis is the difference between the two sensitivities for each radiologist X. An analysis for non-paired observations was also performed, in which the mean of the non-CAD sensitivities is compared with the mean of the CAD sensitivities. This allowed completely different readers to be

TABLE 5 Estimated power of various simulated film-reading studies

No. of readers	Cancers	Normal sensitivity	Sensitivity with CAD	Variability of readers	Variability of films	Study power
10	90	0.7	0.8	0.2	0.2	0.96
10	90	0.8	0.9	0.2	0.2	0.97
10	60	0.7	0.8	0.2	0.2	0.85
10	60	0.8	0.9	0.2	0.2	0.91
7	90	0.7	0.8	0.2	0.2	0.77
7	90	0.8	0.9	0.2	0.2	0.75
7	60	0.7	0.8	0.2	0.2	0.59
7	60	0.8	0.9	0.2	0.2	0.64

used in both conditions. With ten readers per group (20 for double-reading) and 60 cancers, this gives a power of 0.76 or 0.83 depending on whether the initial sensitivity is 70% or 80%.

In the above analysis, the difference to be detected between the two conditions is 10%. The other parameters, however, were set rather cautiously and one might reasonably have hoped to detect a smaller difference. Using these estimates this study had a 71% chance of detecting a 5% change in sensitivity.

Study 2

The second study was performed to test the hypothesis that an improvement in radiologists' and radiographers' sensitivity due to the use of prompts could be detected using a specially selected set of images. An improvement due to the prompts will not be detected if the computer algorithms are unable to detect the cancer. An improvement will not be detected if the overwhelming majority of film readers can detect the cancer in the control condition (without the prompts). Therefore, it was decided to assemble a test of cases that were prompted by R2 but had been missed in the past by radiologists. Any effect detected would have to be evaluated in the light of an assessment of the extent to which the cancers in this test set are representative of the cancers detected in the screening programme.

Films

The sample for study 2 contained 120 cases. The films were again taken from the South-West London Breast Screening Service. Cancer cases were confirmed by positive biopsy and normal cases had a subsequent normal screen. The cancer cases were selected for subtlety before being digitised and checked to see whether the ImageChecker would prompt the cancer. The

criterion used for subtlety was that the case must have been missed by at least one film reader in the past. Three different categories of case meeting this criterion were used in putting together the set of cancer cases prompted by the ImageChecker: false-negative interval cancers ($n = 7$), cases used in a previous experiment for which data sheets were still available ($n = 2$) and cases missed by the first reader in normal double-reading ($n = 31$).

The procedure for identifying cases in the latter category was as follows. The records of the screening centre were used to retrieve lists of screen-detected cancer cases. The film packets were then checked. If the cancer had been marked for recall by the second but not the first reader the case was selected as a candidate for inclusion. Checking the records for 1 April 1999 to 31 March 2000, 262 screen-detected cancer cases were found, 31 of which met this criterion but only 19 of which were prompted by R2. Twelve cases meeting the same criteria were obtained by repeating the process with a subsample of cases from the previous year. Forty cases that were successfully detected by the ImageChecker were included in the study. Four further cancer cases not correctly prompted were also included; these were all cases missed by first reader. Control films were unselected normal cases from the same period as the cancer cases.

The frequency of the different radiological abnormalities on the prompted cancer cases is shown in *Table 6*. Only the primary abnormality is listed.

Prompts

The same machine and the same algorithms were used to print paper prompts sheets as described above for study 1.

Readers

Readers who had completed reading all test films in the original study were invited to take part in

TABLE 6 Types of abnormality in the 40 prompted cancers used in study 2

Abnormality	n
Microcalcification	13
Asymmetry	9
Spiculated mass	10
Architectural distortion/ill-defined mass	4
Round mass	4

the study extension. Thirty-five film readers participated: 18 consultant radiologists, 15 trained radiographers and two breast clinicians. All had undergone a rigorous training programme to meet the requirements of the NHSBSP. All were currently working in the screening programme and reading at least 5000 screening cases per annum.

Procedure

The study was conducted at the same five screening centres as study 1: South-West London, Norfolk and Norwich, Luton and Dunstable, Worthing, and Bristol.

The procedure used for study 1 was followed. The order in which each reader viewed the test sets was separately randomised, as was the order in which they viewed the conditions (prompted and unprompted). A minimum period of 1 week was left between reading a test set in the first and in the second conditions. Readers viewed the films

on a standard roller viewer and were asked to complete a report form for each case. Low-resolution images of the films were included on the form. In the prompted condition these were shown with prompts; in the unprompted condition they were shown without prompts. The reader was asked to circle any area of abnormality and state their degree of suspicion for each abnormality in a table below the images. Readers were then asked to give an overall decision on recall for each case. The time taken to read each test set was recorded.

Power calculation

Data for the first 40 film readers in study 1 were analysed. The standard error on a transformed scale (see Chapter 3, section 'Impact of CAD prompts on film readers' sensitivity and specificity', p. 13) for the primary treatment effect was 0.06, which is approximately equivalent to a standard error of about 0.015 on the raw scale for the typical sensitivities seen. The value of this precision may be undermined by the difficulties in the data set, as discussed above. Using the simulation tool set with the parameter values derived from the data for the first 40 readers also gave a figure of 0.06 for standard error. Running the simulation with the parameter values for the proposed extension gave 0.09. It was calculated that revisiting 40 readers (which will require the participation of just three large centres) and using an additional 40 cases (on top of the 15 cases from study 1 meeting the criteria of this study) gave the study 85% power.

Chapter 3

Results

This chapter presents the results and data analysis from the two studies described in Chapter 2.

Study 1

Several analyses were carried out on the data collected in study 1. First, the sensitivity and specificity of the CAD system were assessed. Then the impact of CAD on film readers' sensitivity and specificity was tested, looking first at all film readers, then at radiologists and radiographers separately. The study then attempted to calculate what values for sensitivity and specificity would have been obtained had the radiologists been double-reading with radiographers and to compare these with the values obtained for radiologists using CAD. Several specific hypotheses about the system were tested, a set of which concerned factors that may lead film readers to ignore prompts. Finally, the hypothesis was tested that poor readers may be differentially affected by CAD. The results of all these analyses are presented in the following sections.

Sensitivity and specificity of the R2 ImageChecker

The ImageChecker was judged to have correctly prompted 45 out of the 60 cancer cases. Thirty-six out of 40 screen-detected cases were prompted, a sensitivity of 90%. Only 11 out of 20 interval cases were prompted, a sensitivity of 56%. In the 15 unprompted cancers, there were seven with no prompts in the breast where the cancer was found, while the other eight contained a mean of 1.5 false prompts. The unprompted cancers are classified by abnormality in *Table 7*. The number of false prompts per case is shown, classified according to the type of case, in *Table 8*. The false prompt rate overall was 1.9 per case, with normals producing more false prompts per case than cancer cases.

Impact of CAD prompts on film readers' sensitivity and specificity

Readers' responses were collected on a scale of 1–4 (definitely recall, discuss probably recall, discuss probably not recall, definitely not recall), but this was collapsed to a binary response for the analysis

TABLE 7 Numbers of cancers missed by CAD for the different classes of radiological abnormality

Abnormality	No. of missed cancers
Irregular mass	2
Stellate lesion/spiculated mass	5
Subtle distortion	2
Asymmetry	5
Calcification	1

TABLE 8 Number of false-positive prompts per case

	Normals	Screen-detected	Interval
No. of cases	120	40	20
No. of false prompts	246	58	35
Mean no. of false prompts/case	2.0	1.4	1.7

(recall versus not recall). The sensitivity and specificity of each reader on each roller under each condition were then calculated. (On a small number of occasions, there were missing data for particular films. The sensitivity and specificity were then calculated ignoring that case.) Sensitivities and specificities are on a 0–1 scale and tend not to be normally distributed. A modified logit transform was therefore applied to the data.³⁹

The transformed data were then entered into a generalised linear model (GLM) [repeated measures analysis of variance (ANOVA)] with two within-subject factors: for the prompt (two levels: with prompt, without prompt) and for the roller (three levels for three rollers). One analysis was done for sensitivity and another for specificity. Testing for a difference in sensitivity, a significant effect was found for the roller ($F_{2,76} = 26$, $p < 0.001$), but not for the prompt ($F_{1,38} = 0.003$, $p = 1.0$) or for the interaction between roller and prompt ($F_{2,76} = 2.2$, $p = 0.1$). The data presented in *Table 9* show the effect of taking the model estimated means and 95% confidence intervals (CIs) on the transformed scale and back-transforming them on to the original sensitivity scale of 0–1.

The results of a comparable analysis for specificity are presented in *Table 10*. Again, there is a significant effect for the roller ($F_{2,76} = 5.0$, $p = 0.009$), but not for the prompt ($F_{1,38} = 0.13$, $p = 0.7$) or for the interaction between roller and prompt ($F_{2,76} = 0.9$, $p = 0.4$).

There is, therefore, no evidence that use of the prompts provided by the R2 ImageChecker affected readers' sensitivities or specificities. Power calculations (using a model based on that published by Pepe and colleagues, as explained in Appendix 2³⁷) suggested that the study had a greater than 80% chance of detecting a 10% improvement from an initial sensitivity of 70%. In the event, the initial sensitivities were higher than expected, but as can be seen from the confidence intervals in *Tables 9* and *10*, the study was sufficiently powered to detect differences of less than 0.1 in the sensitivity or specificity.

TABLE 9 Means and 95% CIs for film readers' sensitivities with and without prompts

Sensitivity	Unprompted	Prompted
Roller 1	0.80 (0.76 to 0.83)	0.79 (0.76 to 0.81)
Roller 2	0.83 (0.81 to 0.84)	0.82 (0.80 to 0.84)
Roller 3	0.71 (0.67 to 0.74)	0.74 (0.70 to 0.77)
Overall	0.78 (0.76 to 0.80)	0.78 (0.76 to 0.80)

TABLE 10 Means and 95% CIs for film readers' specificities with and without prompts

Specificity	Unprompted	Prompted
Roller 1	0.83 (0.80 to 0.86)	0.85 (0.82 to 0.88)
Roller 2	0.82 (0.79 to 0.85)	0.81 (0.77 to 0.85)
Roller 3	0.85 (0.81 to 0.88)	0.85 (0.81 to 0.88)
Overall	0.84 (0.81 to 0.86)	0.84 (0.81 to 0.87)

Impact of CAD prompts on radiologists' and radiographers' sensitivities and specificities

Taking the above data for specificity and analysing the results for radiologists and radiographers separately (there were too few breast clinicians in the final sample to make comparisons with that group) gave the results presented in *Table 11*. There was a significant effect for the roller ($F_{2,70} = 21$, $p < 0.001$), but not for prompt, reader type or any of the interactions (p -values > 0.2). Repeating the exercise for specificity, again there was a significant effect due to the roller ($F_{2,70} = 5.1$, $p = 0.009$), but not for prompt, reader type or any of the interactions (p -values > 0.1).

Reading times

Table 12 reports the median and mean reading times for all readers. The means are used in the economic assessment reported in Chapter 4.

Since the time data were generally positively skewed and not normally distributed, analysis of differences was performed using a log transformation. The transformed data were entered into a generalised linear model with two within-subject factors. The two within-subject factors are for the prompt (two levels: with prompt, without prompt) and for the three rollers. Mauchly's test of sphericity was used and there was significant deviation from sphericity for the main

TABLE 12 Reading times for all reader types

	Unprompted		Prompted	
	Median	Mean (SD)	Median	Mean (SD)
Roller 1	45	55.7 (23.5)	46	50.1 (22.5)
Roller 2	50	50.4 (21.3)	50	56.5 (25.8)
Roller 3	50	59.7 (32.6)	55	54.9 (17.1)
Overall	48	55.3 (26.3)	50	53.8 (22.1)

TABLE 11 Impact of CAD prompts on radiologists' and radiographers' sensitivities (means and 95% CIs)

Sensitivity	Roller	Unprompted	Prompted
Radiologist	Roller 1	0.80 (0.75 to 0.83)	0.79 (0.75 to 0.82)
	Roller 2	0.83 (0.81 to 0.85)	0.82 (0.80 to 0.84)
	Roller 3	0.71 (0.66 to 0.76)	0.75 (0.71 to 0.79)
	Overall	0.78 (0.75 to 0.81)	0.79 (0.76 to 0.81)
Radiographer	Roller 1	0.79 (0.71 to 0.85)	0.77 (0.71 to 0.82)
	Roller 2	0.81 (0.76 to 0.85)	0.80 (0.75 to 0.84)
	Roller 3	0.69 (0.60 to 0.77)	0.66 (0.57 to 0.74)
	Overall	0.76 (0.71 to 0.81)	0.74 (0.69 to 0.79)

effect of the roller ($W = 0.78, p = 0.03$) and the interaction between roller and prompt ($W = 0.75, p = 0.02$). There was no significant effect on the times of the roller ($F_{2,58} = 1.9, p = 0.2$), the prompts ($F_{1,29} = 0.2, p = 0.6$) or the interaction between roller and prompt ($F_{2,58} = 2.6, p < 0.1$).

Table 13 shows the median and mean times taken by radiologists to read the rollers.

Table 14 shows the median and mean times taken by radiographers to read the rollers. The tables show that for both prompted and unprompted, radiologists took less time to read a roller than radiographers (the result is true for both median and mean times). There is, however, little difference between the reading times for prompted and unprompted for radiologists and radiographers.

TABLE 13 Reading times of radiologists

	Unprompted		Prompted	
	Median	Mean (SD)	Median	Mean (SD)
Roller 1	40	45.8 (14.7)	45	45.9 (18.3)
Roller 2	42	43.8 (20.1)	45	48.2 (22.7)
Roller 3	50	57.2 (35.7)	50	55.0 (18.7)
Overall	44	48.9 (25.1)	48	49.7 (20.0)

TABLE 14 Reading times of radiographers

	Unprompted		Prompted	
	Median	Mean (SD)	Median	Mean (SD)
Roller 1	75	78.6 (28.5)	45	50.7 (30.5)
Roller 2	70	65.7 (20.5)	75	70.7 (26.2)
Roller 3	65	65.0 (31.5)	55	53.6 (17.0)
Overall	70	69.8 (27.2)	58	58.3 (25.2)

TABLE 15 Figures for sensitivity and specificity for radiologists double-reading with radiographers, with a third reader arbitrating on disagreement; figures for sensitivity and specificity of prompted and unprompted radiologists single-reading are presented for comparison

	Roller	Unprompted	Prompted	Double-reading
Sensitivity	Roller 1	0.80	0.79	0.81
	Roller 2	0.83	0.82	0.82
	Roller 3	0.71	0.75	0.68
	Overall	0.78	0.79	0.77
Specificity	Roller 1	0.81	0.80	0.87
	Roller 2	0.82	0.79	0.86
	Roller 3	0.82	0.82	0.91
	Overall	0.81	0.80	0.88

Comparison with double-reading

The data from this study were used to simulate a comparison between a radiologist using CAD and a radiologist double-reading with a radiographer. All the possible combinations of radiologist and radiographer were generated, and a calculation done for each case of what the decision would have been if they had been double-reading. The aim was to simulate double-reading with arbitration so that, in cases where they disagreed over the recall decision, the assessment of a reader chosen at random from others in the study was used as the final arbiter. The sensitivity and specificities that would be achieved with this double-reading protocol were calculated, given the data. These are shown in Table 15, with the corresponding values for single-reading by prompted and unprompted radiologists provided for comparison.

The differences between the conditions were not great. To generate confidence intervals for these data, focusing on the comparison between single-reading by a prompted radiologist and double-reading by radiologists working with radiographers, values for the change in specificity and sensitivity between these two conditions were generated. The total set of values was then sampled by random selection with replacement to simulate a sampling distribution for the data (a procedure known as bootstrapping). This was used to determine 95% confidence intervals for the difference between the intervention (single-reading with prompts) and control conditions (double-reading). These are presented in Table 16.

It appears that sensitivity was higher in the intervention condition on roller 3, but elsewhere the effect was in the other direction. The most noticeable effect was that specificity was higher with double-reading.

TABLE 16 Means and 95% CIs for the difference in sensitivity and specificity between intervention (single-reading by prompted radiologists) and control conditions (double-reading by radiologists working with radiographers)

	Sensitivity	Specificity
Roller 1	-0.023 (-0.054 to 0.0003)	-0.067 (-0.10 to -0.030)
Roller 2	-0.019 (-0.044 to 0.0048)	-0.048 (-0.044 to 0.0048)
Roller 3	0.049 (0.016 to 0.085)	-0.09 (-0.13 to -0.057)

TABLE 17 Results of an analysis attempting to find a difference between two groups of readers: those who saw CAD first and those who saw CAD second.

	Roller 1	Roller 2	Roller 3
Improvement in sensitivity	0.17 (-0.044 to 0.078) $t_{48} = 0.56, p = 0.58$	0.0094 (-0.025 to 0.044) $t_{48} = 0.55, p = 0.58$	0.042 (-0.027 to 0.11) $t_{48} = 1.2, p = 0.23$
Improvement in specificity	0.0091 (-0.038 to 0.056) $t_{48} = 0.39, p = 0.70$	-0.036 (-0.099 to 0.027) $t_{48} = -1.2, p = 0.25$	-0.028 (-0.086 to 0.029) $t_{48} = -0.99, p = 0.33$

The measure analysed is the improvement in the CAD condition; this is presented separately for each roller and for both sensitivity and specificity. Improvement is calculated for each reader and means calculated for the two groups. The table presents the difference between the two means and the 95% CIs for that difference, and p -values for the t -tests comparing the change in sensitivity and specificity of the two groups of readers.

Impact on poorly performing readers

The possibility was considered that R2 might be of more value for less able readers. To test this hypothesis the readers were divided into two groups: 19 less able readers and 20 able readers. This was then used as a between-subject factor in the analysis. As a measure of ability, average sensitivity scores (mean on transformed scale) across all conditions were used to avoid any regression towards the mean effects.

For sensitivity there was a significant effect of the roller ($F_{2,74} = 27, p < 0.001$) and of reader skill ($F_{1,37} = 62, p < 0.001$), but not for prompt or any of the interactions (p -values > 0.07). For the key interaction between prompt and reader skill, $F_{1,37} = 0.6, p = 0.4$. For specificity there was a significant effect of the roller ($F_{2,74} = 5.3, p = 0.007$) and of reader skill ($F_{1,37} = 14, p = 0.001$), but not for prompt or any of the interactions (p -values > 0.1). For the key interaction between prompt and reader skill, $F_{1,37} = 0.8, p = 0.4$. There was no evidence that the use of R2 affects able and less able readers' sensitivities or specificities differently.

Impact of reading order

For practical reasons it was not possible to specify a minimum interval of longer than 1 week to elapse between a participant reading a roller in one condition and then reading it in the other condition. This meant that it was possible that

readers could be remembering cases. The authors' clinical colleagues assured them that the impact of this on busy film readers would be minimal, but it remained a concern. If readers who were randomised to see cases first in the prompted condition were remembering the location of the prompts when they read in the unprompted condition, the overall assessment of the difference between the two conditions would be diminished.

The following analysis was performed to test for this effect. The readers were separated into two groups: those who read with the prompts first and those who read with the prompts second (this had to be done separately for each of the three rollers). For each combination of reader and roller, the improvement in the prompted condition over the unprompted condition was then calculated. A t -test revealed that there was no difference between the mean improvement in the group who saw the prompts first and the group who saw them second (Table 17).

Why are the prompts ignored?

To determine why there was no impact due to the prompts, several hypotheses were considered. The focus was on those cancers that were the source of the most errors in the unprompted condition and where the ImageChecker had placed a correct prompt. Table 18 shows the percentage of readers correct with and without prompts for the 14

TABLE 18 The 14 cases called correctly by fewer than 90% of readers, ranked according to the impact of the prompt

% of readers correct (prompted)	% of readers correct (unprompted)	Impact of prompt	Prompts on both views	Confident marker	Prompt for calcifications
0.14	0.26	-0.12			
0.46	0.54	-0.08			
0.31	0.36	-0.05	Y	Y	
0.46	0.50	-0.04			Y
0.60	0.58	0.02	Y	Y	
0.90	0.88	0.02		Y	Y
0.90	0.88	0.02	Y	Y	
0.90	0.86	0.04	Y		
0.87	0.82	0.05		Y	Y
0.48	0.42	0.06	Y		
0.96	0.88	0.08			Y
0.58	0.50	0.08			
0.78	0.68	0.10	Y		Y
0.72	0.56	0.16		Y	
0.64	0.60	0.04			
Y, yes.					

cancer cases where fewer than 90% of the film readers made a correct decision without prompts, but which the ImageChecker prompted.

The 14 films that fewer than 90% of readers called correctly in the unprompted condition were ranked according to the difference in the percentage who called it correctly in the prompted versus unprompted conditions; this difference was taken as a measure of the impact of the prompt. *Table 18* shows for which of the 14 films a microcalcification prompt, an emphasised prompt or a prompt in both views was displayed for the primary abnormality. They are shown ranked according to the difference in the percentage of readers making a correct decision between the conditions. Looking at the distribution of cases that were prompted correctly, on two views, prompted with a confident marker or prompted for calcifications, there is no obvious correlation between any of these three factors and the impact of the prompt. There is, however, a clear, but weak, effect suggesting that in these cases prompts can improve performance.

Therefore, a decision was made to perform a second study with a much larger sample of cases that met these two criteria, to establish the potential impact of prompts in these kinds of cases. The results of this study are presented in the next section.

Study 2

The analyses carried out on the data from study 2 were similar to those carried out on the data from study 1. The impact of CAD on film readers' sensitivity and specificity was tested, looking first at all film readers, then at radiologists and radiographers separately. The test of impact was repeated looking at all readers, putting the data from study 2 together with data from a subset of images from study 1, those meeting the criteria used to select cases for study 2. The hypothesis was tested that readers may be differentially affected by CAD. The study also considered whether readers paid more attention to emphasised prompts or to calcification prompts. The results of all these analyses are presented in the following sections.

Impact of CAD prompts on film readers' sensitivity and specificity

Readers' responses were collected on a scale of 1–4, as for study 1, but this was again collapsed to a binary response for the analysis (recall versus not recall). The sensitivity and specificity of each reader on each roller under each condition were then calculated. (On a small number of occasions, there were missing data for particular cases. The sensitivity and specificity were then calculated ignoring that case.) The data were, as for study 1, transformed using a modified logit function and

TABLE 19 Comparison of sensitivities in the prompted and unprompted conditions for all readers on all films meeting the study 2 criteria (means and 95% CIs)

Sensitivity	Unprompted	Prompted
Roller 4	0.83 (0.78 to 0.87)	0.84 (0.79 to 0.88)
Roller 5	0.72 (0.66 to 0.77)	0.78 (0.73 to 0.82)
'Roller 6'	0.64 (0.59 to 0.68)	0.65 (0.60 to 0.71)
Overall	0.74 (0.70 to 0.77)	0.76 (0.72 to 0.80)

then entered into a GLM with two within-subject factors (a repeated measures ANOVA). The two within-subject factors were for the prompt (two levels: with prompt, without prompt) and for the roller (two levels for two rollers). One analysis was done for sensitivity and another for specificity.

Sensitivity

There was a significant effect of the roller ($F_{1,34} = 18.4, p < 0.001$), but not of the prompt ($F_{1,34} = 2.8, p = 0.10$) or the interaction between roller and prompt ($F_{1,34} = 1.2, p = 0.3$).

Specificity

There was a significant effect of the roller ($F_{1,34} = 60.1, p < 0.001$), but not of the prompt ($F_{1,34} = 3.1, p = 0.088$) or the interaction between roller and prompt ($F_{1,34} = 1.6, p = 0.2$).

Overall, performance improved with R2, but the effects were small and not statistically significant.

A 'virtual' third roller was also created by taking data from the first study, using cases that fitted the criteria for the second study. The inclusion of these films had been assumed when performing the power calculation for the study. This mock roller had 15 cases and 28 non-cases. The basic GLM was performed as before (assuming sphericity), giving the following results.

Sensitivity

There was a significant effect of the roller ($F_{2,68} = 39.4, p < 0.001$), but not of the prompt ($F_{1,34} = 2.3, p = 0.14$) or the interaction between roller and prompt ($F_{2,68} = 1.0, p = 0.4$). Taking the model estimated means and 95% confidence intervals on the transformed scale and back-transforming them on to the original scale gave the data shown in *Table 19*.

Specificity

There was a significant effect of the roller ($F_{2,68} = 23.4, p < 0.001$) and of the prompt ($F_{1,34} = 6.0, p = 0.019$), but not of the interaction

TABLE 20 Comparison of specificities in the prompted and unprompted conditions for all readers on all films meeting the study 2 criteria (means and 95% CIs)

Specificity	Unprompted	Prompted
Roller 4	0.92 (0.86 to 0.92)	0.92 (0.89 to 0.94)
Roller 5	0.81 (0.76 to 0.84)	0.81 (0.79 to 0.84)
'Roller 6'	0.83 (0.80 to 0.86)	0.85 (0.81 to 0.88)
Overall	0.85 (0.82 to 0.87)	0.87 (0.84 to 0.89)

between roller and prompt ($F_{2,68} = 0.7, p = 0.5$). Taking the model estimated means and 95% confidence intervals on the transformed scale and back-transforming them on to the original scale gave the data shown in *Table 20*.

The values in *Table 20* may look as if there is less of an effect than in *Table 19*, but the logit transformation makes more of a difference for scores nearer 1, as here. Hence, this result was significant, but looked no more marked than that for sensitivity. There is evidence here that R2 improves specificity, but not sensitivity.

Impact of CAD prompts on radiologists' and radiographers' sensitivities and specificities

This analysis assessed whether performance varied by reader type. The breast clinicians were excluded as there were too few of them to perform any meaningful analysis, so this analysis was conducted on 18 radiologists and 15 radiographers.

Sensitivity

There was a significant effect of the roller ($F_{1,31} = 15.3, p < 0.001$), but not for any of the other effects (p -values > 0.1).

Specificity

There was a significant effect of the prompt by roller ($F_{1,31} = 4.3, p = 0.047$), but not for any of the other effects (p -values > 0.07).

There were no statistically significant effects of reader type, or interactions with reader type, for sensitivity or specificity. In this study, radiographers were as accurate as radiologists and equally unaffected by using R2.

Impact on poorly performing readers

This analysis assessed whether R2 was of more help for poor readers. The readers were divided by their average sensitivity scores (mean on transformed scale) across all conditions (to avoid any regression towards the mean effects) into two halves: 18 less

TABLE 21 Mean (95% CI) sensitivity and specificity for the two studied conditions and the simulation of double-reading

	Single-reading	Single-reading with CAD	Double-reading
Sensitivity	0.77 (0.73 to 0.81)	0.80 (0.76 to 0.84)	0.81 (0.79 to 0.84)
Specificity	0.85 (0.81 to 0.87)	0.86 (0.84 to 0.88)	0.88 (0.86 to 0.90)

TABLE 22 Means and 95% CIs for the comparisons between single-reading and reading with CAD and between single- and double-reading

	Single-reading with CAD compared with single-reading	Double-reading compared with single-reading
Increase in sensitivity	0.03 (−0.0027 to 0.064)	0.04 (0.014 to 0.077)
Increase in specificity	0.01 (−0.0033 to 0.034)	0.03 (0.0099 to 0.062)

able readers and 17 better readers. This was then used as a between-subject factor in the analysis.

Sensitivity

There was a significant effect of the roller ($F_{1,33} = 18.4, p < 0.001$) and of reader skill ($F_{1,33} = 67.2, p < 0.001$), but not for prompt ($F_{1,33} = 2.7, p = 0.11$) or any of the interactions (p -values > 0.2). For the key interaction between prompt and reader skill, $F_{1,33} = 1.2, p = 0.3$.

Specificity

There was a significant effect of the roller ($F_{1,33} = 60.9, p < 0.001$) and of reader skill ($F_{1,33} = 13.9, p = 0.001$), but not for prompt ($F_{1,33} = 3.1, p = 0.086$) or for any of the interactions (p -values > 0.18). For the key interaction between prompt and reader skill, $F_{1,33} = 0.7, p = 0.4$. There is no evidence that use of R2 affects good and worse readers' sensitivities or specificities differently.

Simulation of double-reading

Data from the performance of individual readers working without CAD were used to simulate the effect of double-reading with arbitration. For each pair of readers the result was taken to be recall if both agreed on recall, no recall if both agreed on no recall, and if they disagreed, the result was determined using the judgement of a third reader, selected at random.

Confidence intervals on the data were generated using a bootstrapping technique in which means were calculated in each of the three conditions (single-reading, single-reading with CAD and double-reading) for 999 random simulated samples generated from the sets of scores. The 95% confidence interval was taken to be the range between the 25th and 975th means. Again, means

were calculated following a logit transformation but the critical values are presented here following an inverse transformation. The results are shown in *Table 21*.

The same bootstrapping technique was used to calculate 95% confidence intervals for two comparisons, one between single-reading and single-reading with CAD and one between single-reading and double-reading. The results are shown in *Table 22*.

Double-reading increased sensitivity compared with single-reading. The sensitivity for single-reading with CAD was greater than that for single-reading; however, since the lower end of the 95% confidence interval for the difference was below zero, it was not statistically significant. The mean specificity was also improved both by CAD and by double-reading; again the difference due to double-reading was statistically significant, whereas that due to CAD was not.

Why are the prompts ignored?

Study 2 involved 35 readers and 40 cases of correctly prompted cancer. There were six missing observations, hence 1394 reports of a film reader using correctly placed prompts to interpret cases. Of these, 305 were not recalled. Several factors were considered that may affect whether or not a prompt was recalled. Two factors relate to the case: difficulty and size of lesion. *Figure 1* shows a plot of failure to recall against the percentage of confident 'no recall' decisions (readers had a choice between 'definitely no recall' and 'discuss, probably no recall', 'discuss, probably recall' and 'definitely recall'). When readers were less confident about their decision in the unprompted condition, they were more likely to act on a correct prompt and to recall the case in the prompted

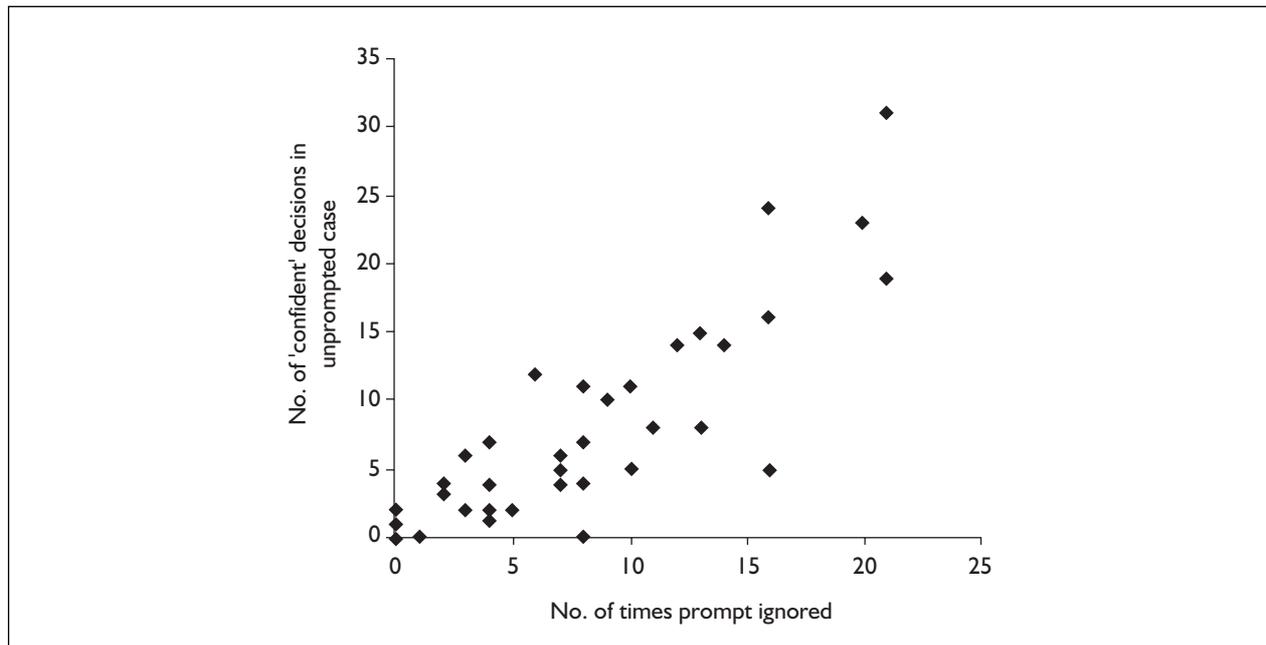


FIGURE 1 Number of times a cancer case was given a confident 'no recall' decision in the unprompted condition, plotted against the number of times the prompt was ignored

TABLE 23 Effect on impact of three characteristics of prompts

	1st quartile (n = 11)	2nd quartile (n = 10)	3rd quartile (n = 9)	4th quartile (n = 10)
% of emphasised prompts	91	60	30	50
% of calcification prompts	36	30	33	40
No. of prompts per case	4.4	4.6	3.8	2.6

condition (Pearson's coefficient of correlation = 0.85, $p < 0.001$).

There was no correlation between the size of the lesion, as assessed at biopsy, and the rate of failure to recall (Pearson's coefficient of correlation = -0.9 , $p < 0.9$).

Three factors relating to the prompts were also considered: whether or not the prompt was emphasised, whether or not the prompt was for calcification and whether or not there were any distracting prompts. The 40 cancers were divided into four quartiles according to the rate of failure to recall in the prompted condition (Table 23). Prompts in the first quartile were ignored on zero to three occasions, in the second on 4 to 7 occasions, in the third on eight to 11 occasions and in the fourth on 12 to 21 occasions.

Readers were much less likely to ignore emphasised prompts (Mann-Whitney $Z = 2.2$, $p = 0.029$). There was no evidence that prompts

for calcification were more likely to be responded to (Mann-Whitney $Z = 0.33$, $p = 0.74$). There was a slight trend towards cases with more prompts being more likely to be recalled correctly (Spearman's $r = -0.29$, $p = 0.07$).

Three factors relating to the reader were considered: years of experience, prior expectation and ability. Data were collected from 31 of the readers about the number of years that they had been working in the screening programme. Thirty-three of the readers completed a questionnaire after using the system in study 1, in which they rated the value of the ImageChecker on a scale of 1–100. This score was used as an index of the reader's expectations. Mean sensitivity achieved in study 1 was used as a measure of ability. For each of these three factors a test was done for a correlation with the impact of the prompting system on each reader. Impact was measured as the improvement in sensitivity in the prompted condition compared with the unprompted condition. None of the three factors considered correlated with impact.

Chapter 4

Economic assessment

Introduction

Background

Accurate estimates of cost-effectiveness are increasingly required before the introduction of health technologies. Such estimates quantify whether the benefits, such as improved quality of life and/or survival, are sufficient to justify the scarce resources required. Screening technologies are no different to other technologies in this respect and therefore estimates of the cost-effectiveness of a new screening programme, or changes to an existing screening programme, are required. Assessment of screening technologies does, however, raise some specific issues when estimating their cost-effectiveness. In particular, there is a need to address the harms of any screening programme (such as false-positive and false-negative outcomes) in addition to any potential benefits arising from early detection. Furthermore, in the UK, decisions about screening programmes tend to be made at national level rather than by individual GPs or clinicians. Consequently, formal screening programmes require substantial investment and this generates a further need to produce estimates of cost-effectiveness.

In the UK, the report that led to the introduction of the NHSBSP¹ was one of the earliest cost-effectiveness studies to estimate a cost per quality-adjusted life-year (QALY) through a combination of estimates of costs, quality of life effects and survival. When the NHSBSP was introduced it was recommended that economic components should be built into future studies of potential changes to policy options since the costs and benefits of screening would be affected by any change. Today, the focus of breast screening policy, and thus of cost-effectiveness, is on how the NHSBSP can best be delivered.

Costs and benefits associated with computer prompts in breast screening

Computer-aided prompts represent one policy option to explore in attempting to improve the delivery of the NHSBSP. The introduction of computer prompts would be associated with a number of different costs and benefits. On the cost side, introduction of computer-aided prompts would require investment in image processors and

this may represent a significant capital investment across the NHSBSP. Additional costs may be associated with training of staff to use the prompting system as well as time taken to digitise films. There are, however, two areas with the potential to realise cost savings using computer prompts. The first is that if the time taken to read films using computer prompts is lower than conventional reading methods this may result in cost savings. The second is that if the recall rate for assessment is lower with computer prompts than conventional reading methods this may result in cost savings.

If computer prompts reduce the specificity of the test then more women will experience false-positive outcomes and the associated quality of life effects of a false alarm. Furthermore, if computer prompts reduce the sensitivity of the test then more women will experience false-negative outcomes and the associated quality of life effects of false reassurance, as well as the quality of life effects associated with the fact that the cancer may only be detected at a more advanced stage and may therefore require more intensive treatment. Further changes in costs and benefits may arise if computer prompts affect not only the rate of cancer detection but also the type of cancers identified. If computer prompting detects cancers with better prognoses, cost savings may arise as a result of lower treatment costs and benefits may arise in terms of improved quality of life and survival. In a prospective study using computer prompting, Freer and Ulissey found that a high proportion of the additional cancers detected with computer prompts were DCIS and therefore had a better prognosis, but they did not consider the impact of this on the cost-effectiveness of prompting.³⁴

To date, no study has estimated and compared the costs and benefits associated with the use of computer prompts in mammography.

Previous research on cost-effectiveness of breast screening

Despite the fact that there has been no cost-effectiveness analysis of computer-aided prompts, there has been considerable interest in the cost

and cost-effectiveness of breast screening policies in many countries. This interest stems from the fact that the implementation (and running) of a breast screening programme is likely to involve substantial investment for governments or health insurance organisations. Two of the earliest studies estimating the cost-effectiveness of introducing a breast screening programme were in the UK¹ and in The Netherlands.⁴⁰ Since these studies were published, the cost-effectiveness of introducing breast screening programmes has been estimated in many other countries.^{41–44}

In recent years, the focus of interest in the cost-effectiveness of breast screening has been on how breast screening services are delivered. These include studies exploring the cost-effectiveness of the age of screening,^{45,46} the screening test⁴⁷ and digital mammography.⁴⁸ In the UK, recent cost-effectiveness studies have explored a range of reading policies, including the number of views and the number and combination of readers.^{49–53} All of these studies concluded that increasing the number of views or increasing the number of readers was cost-effective. The interest in the UK in exploring the cost-effectiveness of reading policies may have arisen as a result of using such policies to reduce false-negative rates and improve the accuracy of cancer screening in general.

Studies on the impact of cost-effectiveness of the number of views and number and combination of readers are relevant to the design of a cost-effectiveness analysis of computer-aided prompts. The aim of this section is not to provide a systematic review of these studies; rather, it highlights the points that have influenced the methodology adopted in this report.

Almost all previous studies adopted an intermediate outcome measure (cancers detected) as the main outcome measure and thus presented a cost-effectiveness ratio measuring the cost per additional cancer detected. The only exception is the study by Wald and colleagues which, in addition to a cost per cancer detected, estimated a cost-effectiveness ratio of cost per year of life saved.⁴⁹ This latter estimate was, however, based on a crude assumption that the 24% increase in cancer detection would result in a 24% increase in lives saved. Although the number of cancers detected is a commonly reported outcome measure, it can only be used to judge the relative cost-effectiveness of a technology compared with technologies that have also estimated a cost per cancer detected. It cannot allow a judgement as to the relative value for money of a technology

compared with other health technologies. Only estimation of a cost per life-year gained or cost per QALY can achieve that. Hence, this study estimates both a cost per cancer detected and a cost per QALY, so that a range of comparisons can be made. Estimation of outcomes beyond cancers detected also allows consideration of quality of life effects of cancer, recurrences for cancer and life-years gained through cancer detection.

The second key feature of the methods adopted in the cost-effectiveness studies of reading policies is that, with the exception of the study by Wald and colleagues,⁴⁹ all studies synthesise costs and effectiveness data from a number of sources and are not based on trial data. Hence, although this study is based primarily on effectiveness data from archived films, it is not unusual for cost-effectiveness studies to be based on non-randomised studies.

Methods

Aims and scope of cost-effectiveness analysis

The aim of the economic analysis was to estimate the cost-effectiveness of mammographic film reading using computer prompts. The cost-effectiveness analysis used estimates of sensitivity and specificity from the reading studies reported in Chapter 3 and combined these estimates with estimates of health service costs. The study adopted a health service perspective for the estimation of costs since these types of cost represent the main difference between prompted and unprompted reading. Cost estimates were derived from five breast screening centres.

Two forms of cost-effectiveness were estimated. The first was a cost per cancer detected at 12 months. This figure can be compared with previous estimates of cost per cancer detected for mammographic reading policies to indicate the relative value for money of computer prompting. The second was a cost per QALY at 10 years. This latter estimate used a modelling approach to extrapolate from cancers detected. The aim of the extrapolation method was to estimate the gain from detecting any additional cancers (life-years or QALYs) and to quantify any changes in longer term costs as a result of any additional cancer detection. The cost per QALY figure can be compared with estimates of cost per QALY of other health technologies or threshold cost-effectiveness ratios to determine the relative cost-effectiveness of computer prompting.

TABLE 24 Description of studies and comparisons made in the economic assessment

Characteristic	Study 1	Study 1 (double-reading simulation)	Study 2	Study 2 (virtual roller)
No. of cases	180	180	120	163
Selection of cases	'Representative'	'Representative'	'Selected': previously missed by film reader and corrected prompted by R2	As study 2 with the addition of cases from study 1 meeting the criteria for study 2
No. of cancers	60	60	40	?
Comparison made in economic assessment	Prompted versus unprompted	Double-reading (unprompted) versus single-reading (prompted)	Prompted versus unprompted	Prompted versus unprompted

The main study was powered to detect differences in sensitivity and specificity, rather than to detect differences in economic end-points, such as cost per cancer detected and cost per QALY. As this is often the case for cost-effectiveness studies, the recommended approach in such situations is to estimate uncertainty around the cost-effectiveness estimates, since it is possible that the lack of significance in the effect difference may have arisen as a result of insufficient power to detect differences.⁵⁴

The comparisons made in the economic assessment are summarised in *Table 24*, along with a summary of key features of the studies discussed in detail in Chapter 3.

Health service costs

Overview

The key potential changes in health service costs arising with the introduction of a computer prompted system compared with an unprompted system are:

- additional equipment costs of R2 ImageChecker (including maintenance costs)
- additional costs of training staff to use prompting system
- additional staff time taken to digitise films for prompting
- change in film reading time
- change in total assessment costs resulting from any change in recall rate
- change in total treatment and future costs resulting from any change in cancer detection rates and/or prognosis of cancers.

It is clear that although prompting will be cost-increasing in terms of additional costs associated with equipment, training and digitising time, it

may save costs if savings arise as a result of reading costs, total assessment, total treatment and future costs. The methods used to estimate these potential changes are now described (treatment costs and future costs are explained in the section on cost per QALY). Cost differences between prompted and unprompted reading are presented as a cost per 1000 women screened. All costs are presented in 2001/02 prices. All future costs are discounted at 3.5%, the rate recommended by HM Treasury since April 2003.⁵⁵

Equipment, training and digitising costs

In estimating the average equipment cost per screen it is important to consider the number of computer prompting installations required per centre. A conservative estimate would be to assume that each screening centre would require one computer prompting system (the R2 ImageChecker), but such an assumption would limit the generalisability of results since the size of the screening centre is known to vary widely.⁵⁵ If screening centres are large then they will require more than one ImageChecker to be able to keep up. To estimate the number of installations and hence equipment cost per screen, an estimate of the throughput of the ImageChecker and an estimate of screening unit size are required.

In the authors' experience, it is difficult to process more than 180 cases per week with a single digitiser. The average number of weeks for which a screening centre is open is 49 weeks per year.⁵⁵ Hence, the average throughput of one ImageChecker is 8820 cases per year. The size of screening centre can be estimated from data from the same survey.⁵⁶ Centre-specific data on the number of screens was divided by the average throughput of R2 (8820) to estimate the number of installations required per centre. Clearly, capital

investments are 'lumpy' in that the number of screens divided by the throughput of an ImageChecker is unlikely to be a whole number and will therefore have to be rounded up to reflect the real number of installations required. In this study, the following rounding procedure is used. If, after dividing the number of screens by the throughput of the ImageChecker, the number of installations required is estimated to be greater than zero and less than or equal to 1.3 then 1 is assumed to be required; if greater than 1.3 or less than or equal to 2.3 then 2; if greater than 2.3 and less than or equal to 3.3 then 3. These numbers are then used to estimate the average equipment costs. Any lower number of installations would affect the efficiency of the current screening centres and screening programme as a whole. Centre-specific data on the number of women screened and number of installations required are combined with capital costs (now described) to estimate the average equipment cost per screen.

The capital costs of the ImageChecker system are estimated based on list prices, converted to annual equivalents using the equivalent annual cost (EAC) method. Annuity factors are based on a 3.5% discount rate.⁵⁵ The capital cost of the ImageChecker is £108,000 (NHSBSP 2001). The lifespan of equipment (7 years) is based on manufacturers' estimates. The maintenance costs per machine are £10,000 per annum.⁵⁷

The introduction of a prompting system requires staff to be trained in its use and generates additional staff time costs. These costs are estimated on the basis of a radiographer and radiologist within a screening centre having a 3-hour training session with retraining every 3 years. The time is then weighted by cost per hour of a radiologist (£45.11) and a radiographer (£13.71).⁵⁸ The average training cost per screen is calculated based on the average number of radiographers and radiologists per screening centre (average of 4.89 radiographers per centre and 0.89 radiologists per centre) as well as the number of centres (97).⁵⁵ Since training costs span 3 years they are converted to an annual cost using the equivalent annual cost method, annuitised at 3.5%. The total annual training cost for the NHSBSP is divided by the number of women screened per year to estimate the average training cost per screen.

Computer prompting generates additional costs in terms of the time spent scanning and digitising films before the system is able to place prompts. The time involved in digitising one case (four

films) is 4.5 minutes, based on experience of digitising the films for the study. The time taken to digitise is then weighted by the cost per hour of a radiographic assistant (£5.90) to estimate staff costs of digitising.⁵⁸

For the simulation of double-reading, based on study 1 data, comparing double-reading (one radiologist and one radiographer) with single-reading (one radiologist using computer prompting), only the latter option includes the additional costs of equipment, training and digitising associated with computer prompting.

Reading time

A potential area of cost saving with computer prompting is if the time taken to read films using the prompted system is shorter than with the unprompted system. Several approaches were considered for estimating reading time. Initially, observational work was considered as the preferred approach to estimating reading time. Given, however, that it would not have been possible to observe all 40 readers reading three rollers each using both prompted and unprompted conditions (240 reading sessions), the option was either to observe a subsample or to ask readers to self-report their reading times. The latter approach was adopted to allow the maximum number of observations on reading time. Each reader was asked to record the time the reading of the roller started and the time the reading finished. Other details of the reading session were also recorded, such as whether the reader was disturbed and the time of day. The average time taken to read a roller (containing 60 cases, four films per case) was estimated for each reader type. The average time was then weighted by the cost per hour of a radiographer (£13.71) and radiologist (£45.11) to estimate the reading time costs. The estimate of reading costs used in the base-case analysis is an average of the reader types.

For the simulation of double-reading, based on study 1 data, comparing double-reading (one radiologist and one radiographer) with single-reading (one radiologist using CAD), the following approach was adopted to estimate reading costs. For double-reading, the average time cost of radiologists (unprompted) plus the average time of radiographers (unprompted) was used as the reading time cost. Although radiographers take longer to read films, the unit cost of their time is lower. For single-reading, the average time cost of radiologists (prompted) was used as the reading time cost.

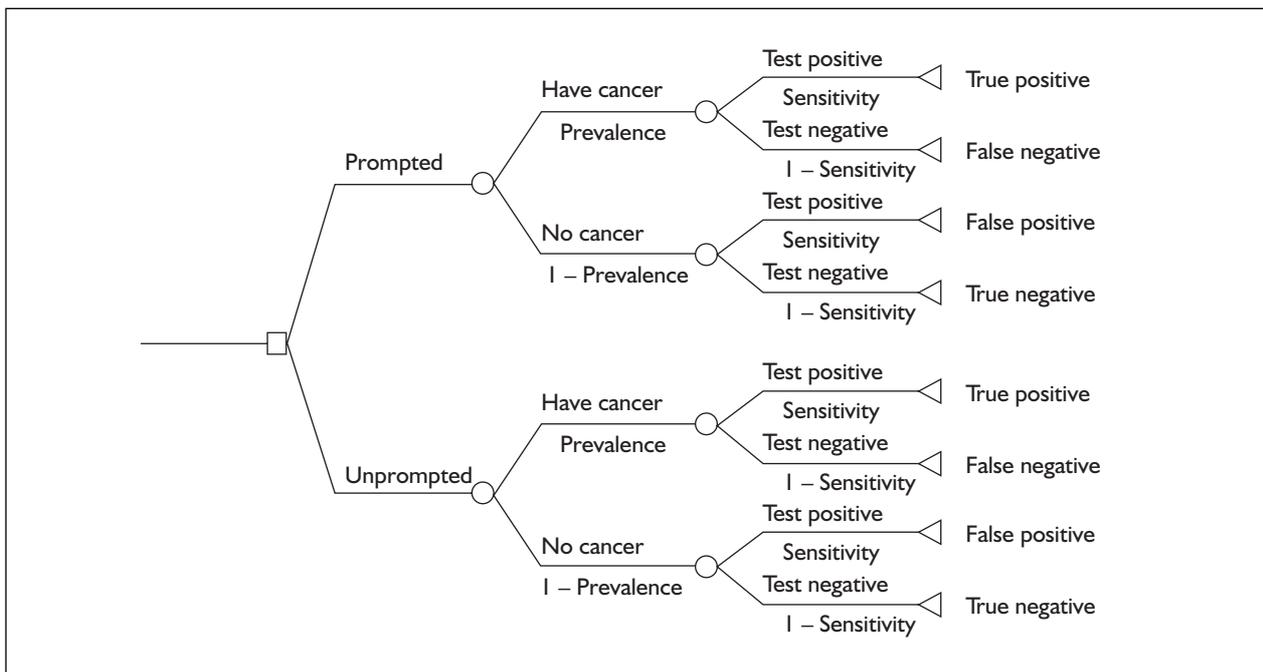


FIGURE 2 Screening pathways

Other screening costs and assessment costs

Other screening costs, such as the costs of invitation and the costs of films, consumables and screening equipment, do not differ between prompted and unprompted. The unit cost of these screening costs is £25.40.⁵⁸ This unit cost excludes reading costs.

The total assessment costs between prompted and unprompted costs may differ if the recall rate differs. The unit cost of assessment was taken from a study of the costs of assessment from five screening centres.⁵⁸ The unit cost estimate was estimated by weighting the costs of individual assessment procedures by the proportion of women having those procedures, and represents an estimate of cost of assessment. The unit costs included staff time in running the assessment session and in the administration of results, consumables per procedure [films for mammography, needles, syringes for fine-needle aspiration (FNA), laboratory processing for FNA], equipment per procedure (screen taking, processing, reading), ultrasound machine (for ultrasound), stereotactic machine (for FNA), buildings and overheads. The unit cost of assessment is £69.04.

Cost per additional cancer detected

Decision analysis

Cost data were combined with data on the sensitivity and specificity of prompted and

unprompted systems to estimate a cost per cancer detected up to the point of detection, and to provide an estimate of cost-effectiveness in the short term.

Figure 2 shows the pathways leading to four screening outcomes: true positives, false negatives, false positives and true negatives. The decision analysis was split into 'cancer' and 'no cancer' so that the sensitivity and specificity estimates from the reading of the test sets could be entered directly.

For each of the four screening outcomes, costs were assigned. Costs for the true-positive outcome represent the cost of the screen using either the prompted or unprompted system plus the cost of assessment. (Treatment costs for true-positive cases are addressed when estimating future costs and benefits, discussed below.) Costs for the false-negative outcome represent the cost of the screen using either the prompted or unprompted system. This is because false-negative cases are not detected until later and at this stage it is not yet known that the case is actually cancer. (False-negative costs and outcomes are addressed when estimating future costs and benefits, discussed below.) Costs for the false-positive outcome represent the cost of the screen using either the prompted or unprompted system plus the cost of assessment. Costs for the true-negative outcome represent the cost of the screen using either the prompted or unprompted system.

Sensitivity and specificity estimates from the respective studies reported in Chapter 3 were entered into the model to determine costs and recall rates. A further variable, 'prevalence', was entered into the model to determine the number of cancers detected. The test sets used in the studies contained a higher proportion of cancers than would normally be the case in a routine screening situation. If the ratio of cancers to control cases in the test sets was used to estimate the number of cancers detected from computer prompting in routine screening, it would greatly exaggerate the number of extra cancers detected. Therefore, an estimate of a prevalence of breast cancer in the population to be screened is required. Several estimates are available for prevalence. The screening programme detects around 6.6 cancers per 1000 women screened.⁴ In addition to this, a proportion of cancers is missed at screening (estimated to be between 10 and 25%⁵). This produces estimates of prevalence of between 0.77 and 0.88%. Other studies have suggested that the ratio of cancers to control cases in the test sets is in the order of 1:200,⁵⁹ suggesting a prevalence of 0.5%. Given the uncertainty regarding the prevalence, the model used a range of 0.5–0.8% for prevalence with a base case of 0.65% and the uncertainty in the prevalence was addressed in a sensitivity analysis. Overall, therefore, the model assumed that the sensitivity and specificity achieved during experimental conditions would be replicated in routine screening, but that the prevalence of breast cancer is 0.65%.

The decision analysis model produced estimates of additional costs and any additional costs per any additional cancers detected, based on the sensitivity and specificity estimates reported in Chapter 3. A cost per cancer detected was estimated by dividing the additional costs by the additional cancers detected.

Cost per QALY gained

Overview of modelling approach

The decision analysis produced estimates of a cost per cancer detected, but the results could not be used to judge the relative cost-effectiveness of computer prompts compared with other health technologies because they gave no indication of what was gained from detecting any extra cancers. To address this gap, a modelling approach was adopted to extrapolate from a cost per cancer detected to a cost per QALY. Modelling is increasingly being used in economic evaluation and is often required to extrapolate from short-term to long-term outcomes, since long-term

outcomes are most relevant for decision-making. An approach that is commonly used in economic evaluation to model long-term outcomes is Markov modelling. It is particularly appropriate for economic evaluation since it allows future costs and outcomes to be estimated simultaneously.⁶⁰

In the context of breast screening, Markov models have been used to estimate the cost-effectiveness of digital mammography and the cost-effectiveness of extending screening to women aged 40–49 years.⁶¹ Extrapolation from intermediate outcome measures, such as cancers detected, to final outcome measures, such as life-years gained, was used in the trial of one- and two-view mammography,⁴⁹ although the methods used were quite crude as the estimation of future costs and benefits was based on an assumption that the same percentage rate of increase in cancers detected achieved with two views would apply to the rate of mortality reduction.

In this study, a Markov model was used to estimate long-term costs and outcomes at 10 years between prompted and unprompted reading. The approach has several features. First, the model extrapolates from cancers detected to QALYs gained using prognosis at time of detection as the predictor of future survival. Since there is no evidence that computer prompting changes prognosis, the model assumes that computer prompting does not affect prognosis and effectively this means that any differences in future costs and life years between prompted and unprompted reading arise as a result of the number of cancers detected. Second, the model takes into account cancers arising from both true-positive and false-negative outcomes. Third, the model takes into account the likelihood of breast cancer recurrences and the costs and quality of life effects associated with recurrences. Finally, the model allows future costs to be estimated. This is a particularly important issue for breast screening, since screening may prevent treatment for late-stage breast cancer, with a possible reduction in treatment costs.

Prognostic index

The indicator used to predict future costs, life-years and QALYs is the Nottingham Prognostic Index (NPI).⁶² The NPI is assigned at diagnosis and incorporates three prognostic factors: tumour size, nodal status and histological grade. The NPI is derived from a Cox regression model and is estimated as follows: $(0.2 \times \text{Size} + \text{Lymph-node status} + \text{Grade})$. The NPI is a continuous index, but can be categorised into three main prognostic groups: good, moderate and poor. In addition to

TABLE 25 Prognostic groups identifiable using the NPI

Prognostic group	Score
DCIS	NA
Good	$NPI \leq 3.4$
Moderate	$3.4 < NPI \leq 5.4$
Poor	$NPI > 5.4$
NA, not applicable.	

these groups, a further group for DCIS is identifiable. Consequently, there are four potential prognostic groups (PGs) into which a woman can be classified at diagnosis (*Table 25*).

The key advantage of the NPI over other staging indicators, such as tumour, node, metastasis (TNM) staging, is that it includes grade of cancer, a factor that has been shown to correlate highly with prognosis.⁶³ The NPI has been shown to be stable⁶² and the factors that comprise the NPI have been shown to pick up the effects of screening.⁶⁴ A further advantage of using the NPI is that DCIS (non-invasive cancers that are commonly detected with breast screening) are identified separately. The NPI is increasingly being used as the surrogate end-point for breast screening trials (e.g. the breast screening frequency trial in the UK).⁶⁵

Prognosis of cancers in studies 1 and 2

Table 26 reports the PGs of the cancers in studies 1 and 2, and shows that the prognosis of cancers is similar between studies. It also shows the PGs of cancers detected at screening,⁶⁶ which found that 22% of cancers were non-invasive and 78% were invasive. Of the 4195 invasive cancers, 46% were of good prognosis, 27% were of moderate prognosis and 5% were of poor prognosis. *Table 26* suggests that the test sets differ from cancers in the screening programme in terms of the lower proportion of poor prognosis cancers being included in the test sets.

In terms of the modelling, the key issue is whether prompting with the R2 ImageChecker detects cancers with a different prognosis. Freer and Ulissey found that the additional cancers detected by computer prompting were primarily DCIS, so it is important to determine whether the ImageChecker is able to detect cancers with prognoses other than DCIS.³⁴ Examination of the cancers in study 1 shows that the ImageChecker missed one out of 14 cancers with DCIS, nine out

TABLE 26 Prognosis of cancers in studies 1 and 2

Prognostic group	Study 1 n (%)	Study 2 n (%)	Screening programme %
DCIS	14 (23)	13 (30)	22
Good	31 (52)	22 (50)	46
Moderate	13 (22)	8 (18)	27
Poor	1 (2)	0 (0)	5
Missing	1 (2)	1 (2)	0

of 31 cancers with a good prognosis and five out of 13 cancers with a moderate prognosis. R2 detected all other cases. Examination of the cancers in study 2 shows that the ImageChecker missed one out of 13 cancers with DCIS and two out of 20 cancers with good prognosis, but detected all other cases. Although the numbers are small, they suggest that the ImageChecker picks up most DCIS cancers, but is also able to detect cancers with other prognoses. In the absence of any other evidence, it can be concluded that since the ImageChecker is picking up the range of prognoses, the cancers that it detects are broadly similar to those detected by the screening programme. Hence, the conservative assumption made in the modelling is that prompted and unprompted reading detects cancers of the same prognosis, and that any differences in the cost-effectiveness of prompted and unprompted reading arise through differences in sensitivity and specificity.

Key features of the Markov model

Details of the Markov model are presented in Appendix 3. This section highlights some of the key features of the model. The model begins at the point where breast cancer has been diagnosed and NPI assigned. There are five states in the model: breast cancer diagnosed, local recurrence, regional recurrence, distant recurrence and dead. A Markov model is estimated for each PG.

Life-years and QALYs are produced by the model by entering data on survival benefits and quality of life effects. Each PG has a different probability of moving from the breast cancer diagnosis state to the different recurrence states, and has a different probability of death. DCIS is assumed to confer no survival benefit and therefore the probability of death following breast cancer is the same as other-cause mortality. Utilities are attached to each state in the model based on the literature (further details on sources are presented in Appendix 3). Life-years and QALYs are discounted at the recommended rate of 1.5%.⁶⁷

Future costs are produced by the model by entering data on the costs of treatment in each state. Each PG has a different initial treatment cost arising from different treatment protocols by PG. All treatment costs include primary treatment, recurrences and follow-up. Costs are discounted at the recommended rate of 3.5%.⁵⁶

The Markov model begins with all patients in the breast cancer-diagnosed state at the age of 50 years. A cycle length of 1 year is chosen since follow-up after breast cancer diagnosis and treatment is annual. The model is run for ten cycles to estimate costs and life-years at 10 years.

The proportion of cases entering each PG is assumed to be the same for both prompted and unprompted reading. The sensitivity and specificity of prompted and unprompted reading are used to determine the number of cancers that are true positive and false negative. The true-positive and false-negative cancers are assumed to have different prognoses, since false-negative cancers are detected at a later stage than true-positive cancers. The proportion of true-positive cases in each PG was taken from *Table 26* (22% DCIS, 46% good prognosis, 27% moderate prognosis and 5% poor prognosis). The proportion of false-negative cases by PG was taken from a study of symptomatically detected cancers,⁶⁸ which found that 3% of cancers were non-invasive and 97% were invasive. Of the 306 invasive cancers, 24% were of good prognosis, 52% were of moderate prognosis and 21% were of poor prognosis. All false-negative cases are assumed to arise in the first year of the model.

The costs, life-year and QALY estimates from the model for prompted and unprompted reading are then used to estimate additional costs, additional life-years and additional QALYs between prompted and unprompted. A cost-effectiveness ratio is calculated as the additional costs divided by the additional life-years or QALYs between prompted and unprompted.

Sensitivity analysis and presentation of results

One-way sensitivity analysis is performed on the cost estimates. In a one-way sensitivity analysis, median reading times are entered and the impact on the difference in screening costs between prompted and unprompted is observed. The estimates of the cost of the computer prompting equipment are also explored in the following way:

by increasing the discount rate to 6% (the recommended rate until 2003), decreasing it to 0% to explore the results without discounting, and assuming a 25% reduction in the list price of the computer prompting system.

In addition to the sensitivity analyses on costs, a series of one-way sensitivity analyses is performed, leaving costs and life-years undiscounted.

Rather than perform a large number of one-way sensitivity analyses, uncertainty in the cost-effectiveness estimates is explored by performing probabilistic sensitivity analysis. The aim of the probabilistic analysis is to address all uncertainties simultaneously. For example, the analysis addresses the uncertainty in the prevalence of the breast cancer as well as the uncertainty in sensitivity, specificity and costs at the same time. The probabilistic analysis uses Monte Carlo simulation methods to estimate uncertainty in the results. Details of the probabilistic analysis are presented in Appendix 3.

A cost-effectiveness acceptability curve⁶⁹ is calculated to indicate the probability of computer prompting being cost-effective at a function of what a decision-maker's threshold cost-effectiveness ratio might be. A £30,000 threshold is used, as this is often perceived to be the threshold ratio in the UK.

Results

Costs

Table 27 presents the difference in health service costs per 1000 women screened between prompted and unprompted reading. The cost of equipment is the main difference between prompted and unprompted, and the table shows that computer prompting increases the cost of equipment by £4016 per 1000 women screened. Computer prompting also increases cost through training and digitising costs. In terms of reading costs, with the exception of study 1 (simulation of double-reading), there is only a minor cost saving arising from prompted reading. This reflects the slightly shorter mean time taken to read prompted rollers. Given, however, that there was no statistically significant difference in the time taken with prompted and unprompted reading, this cost saving is uncertain. For study 1 (simulation of double-reading), savings in reading cost arise as a result of the fact that the total time of single-reading (one radiologist using the computer prompt) is less than the total time of double-

TABLE 27 Difference in health service costs of prompted and unprompted reading by study

Type of cost	Study 1	Study 1 (double-reading simulation)	Study 2	Study 2 (virtual roller)
Equipment	£4,016	£4,016	£4,016	£4,016
Digitising	£435	£435	£435	£435
Training	£773	£773	£773	£773
Reading	-£16	-£245	-£16	-£16
Assessment	£0	£5,497	-£671	-£1,361
Total difference	£5,209	£10,476	£4,538	£3,848

TABLE 28 Cost per cancer detected per 1000 women screened, by study

Additional costs, cancers, cost per cancer detected	Study 1	Study 1 (double-reading simulation)	Study 2	Study 2 (virtual roller)
Costs	£4,016	£10,476	£4,538	£3,848
Cancers	0	0.13	0.195	0.13
Cost per cancer detected	NA	£80,587	£23,272	£29,600

reading (one radiologist and one radiographer). Any savings in reading costs by single-reading are offset by the higher costs of the computer prompting equipment and the increase in assessment costs. The assessment costs reflect any change in recall rate in the studies and the prevalence of breast cancer in routine screening. In study 1, there was no change in recall rate and therefore there were no assessment cost increases or decreases. For study 1 (simulation of double-reading), there was an increase in assessment costs resulting from a higher recall rate with single-reading (one radiologist reading with CAD) (this is consistent with the lower specificity observed for single-reading) and this led to the highest costs relative to the others.

Overall, computer prompting is cost-increasing, with the minimum cost increase being £3848 per 1000 women screened. If no changes in assessment costs arise and a minor reduction in reading costs is found, then the additional cost of computer prompting is £5209 per 1000 women screened.

One-way sensitivity analysis on the costing methods showed that, if median reading times, rather than mean times, are used as the basis for estimating reading costs, the total difference in health service cost between prompted and unprompted reading rises from £5209 per 1000 women screened to £5220. When equipment costs are discounted at 6% and 0%, the total difference in equipment cost between prompted and unprompted reading changes from £4016 per 1000 women screened to £4262 and £3692, respectively. If the list price of the ImageChecker

changes from £108,000 to £81,000 then the increase in equipment costs falls from £4016 to £3375 per 1000 women screened.

Cost per cancer detected

Table 28 reports the additional cost per additional cancer detected (additional costs divided by additional cancers detected). These results are based on the costs in Table 27, as well as the estimates of sensitivity and specificity from the studies reported in Chapter 3 and the prevalence of breast cancer.

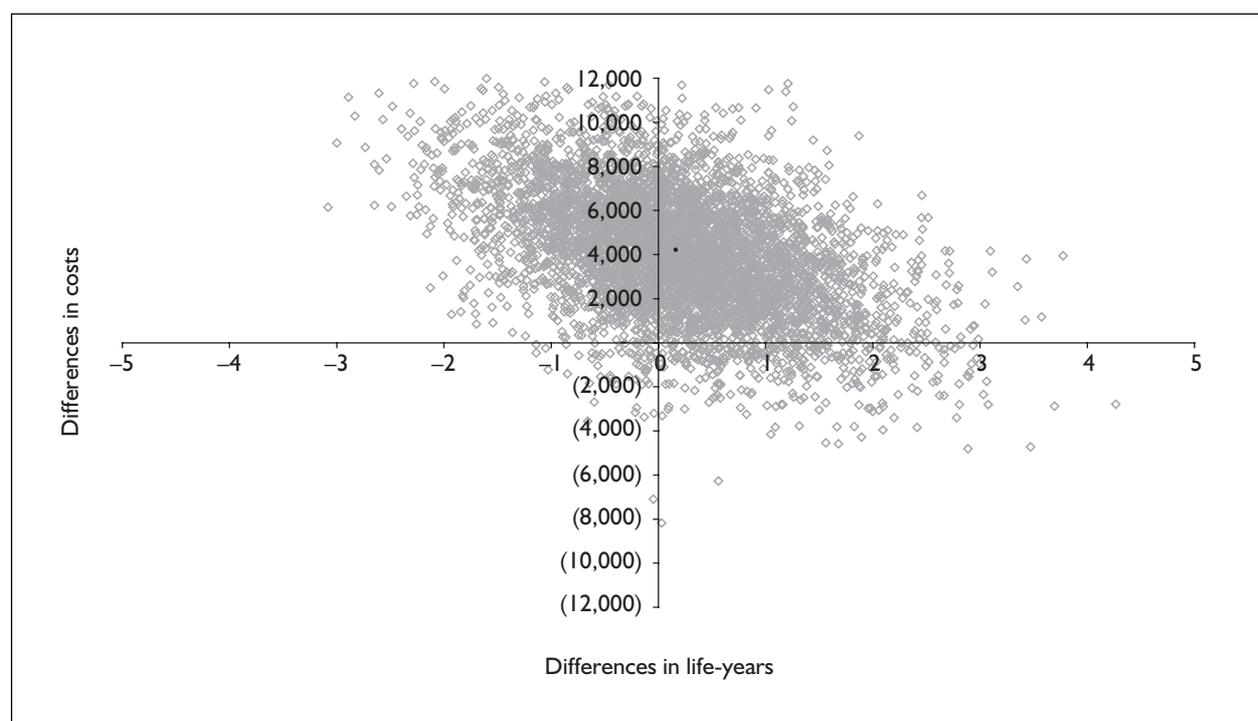
Table 28 shows that, as would be expected from the results on sensitivity and specificity reported in previous chapters, the additional number of cancers detected by computer prompting is small. Since study 1 did not detect any additional cancers, a cost per cancer detected cannot be calculated. For each of the other studies a cost per cancer detected is shown and, since the additional cancer detection rate with computer prompting is small, the cost per cancer detected ratios are high, with the highest being for study 1 (double-reading simulation) because of its higher costs. The interpretation of these figures is discussed in more detail in Chapter 5 (section 'Economic assessment', p. 37).

Cost per life-year gained and cost per QALY

Table 29 reports the estimates of costs, life-years and QALYs by study, per 1000 women screened,

TABLE 29 Cost per life-year gained and per QALY per 1000 women screened between prompted and unprompted, by study at 10 years

Additional costs, life-years, QALYs	Study 1	Study 1 (double-reading simulation)	Study 2	Study 2 (virtual roller)
Costs (undiscounted)	NA	£10,200	£4,123	£3,571
Costs (discounted)	NA	£10,232	£4,171	£3,603
Life-years (undiscounted)	NA	0.13	0.19	0.13
Life-years (discounted)	NA	0.11	0.17	0.11
QALYs (undiscounted)	NA	0.13	0.20	0.13
QALYs (discounted)	NA	0.10	0.16	0.10
Cost per life-year gained (undiscounted)	NA	£78,461	£21,700	£27,469
Cost per life-year gained (discounted)	NA	£93,018	£24,535	£32,755
Cost per QALY gained (undiscounted)	NA	£78,461	£20,615	£27,469
Cost per QALY gained (discounted)	NA	£102,320	£26,069	£36,030

**FIGURE 3** Cost-effectiveness plane (costs and life-years discounted)

10 years after detection of cancer. The cost estimates include the costs of screening and assessment up to the point of cancer detection, as well as the future costs of treating breast cancer and any recurrences. The future costs and life-years associated with study 1 are not modelled since study 1 did not detect any additional cancers.

Table 29 shows that the incremental cost per QALY gained (discounted) for study 2 is £26,069, but for study 2 (virtual roller) and study 1 (double-reading) the discounted cost per QALY ranges from £36,030 to £102,320. This suggests that,

based on the results from study 2, with a discounted cost per QALY of £26,069 there may be a possibility that computer prompting is cost-effective (since this estimate is below the perceived threshold ratio in the UK of £30,000 per QALY). The estimates shown are, however, point estimates of costs, life-years and life-years gained, and do not represent the uncertainty in the estimates. Figure 3 presents the results of the probabilistic sensitivity analysis for study 2. This addresses uncertainty in the parameters in the model and plots the results of 5000 simulations of costs and life-years (both discounted). The point estimate of cost-effectiveness is shown in black; the data

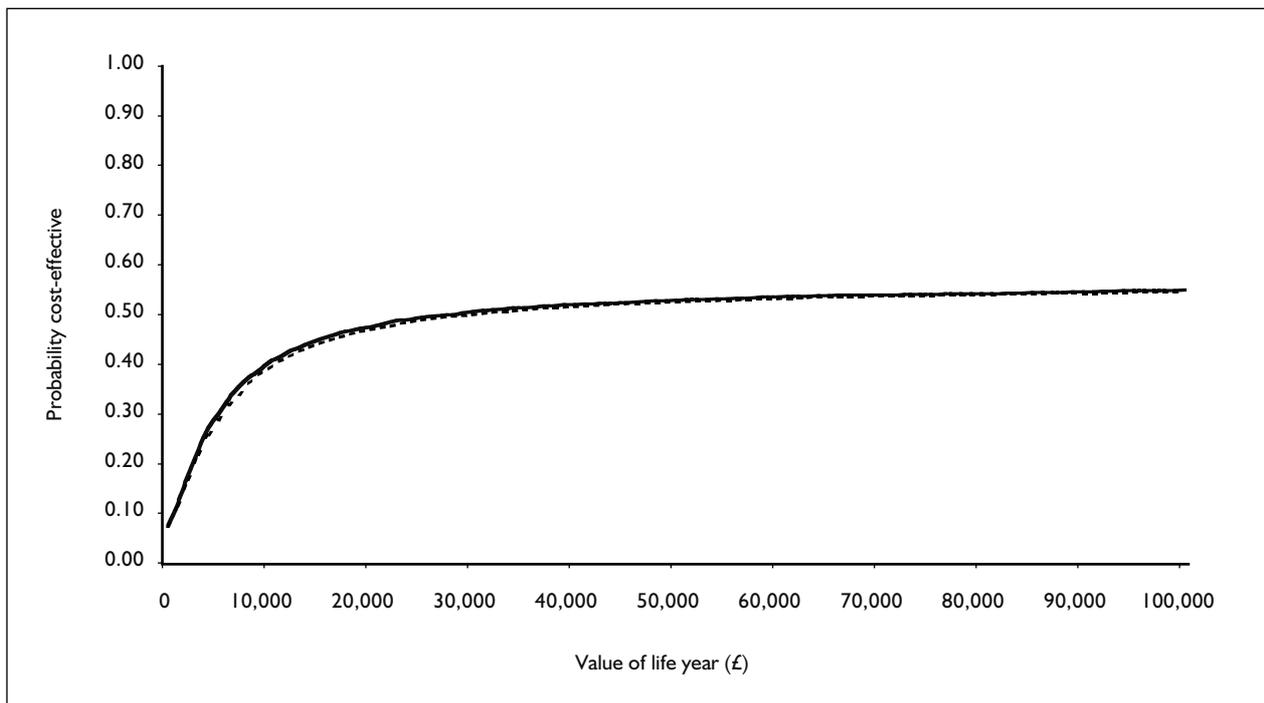


FIGURE 4 Cost-effectiveness acceptability curve

points surrounding it (shown in grey) indicate the degree of uncertainty. If there was no uncertainty about whether computer prompting was cost-effective, all data points in the figure would fall into the top-right quadrant of the figure, suggesting that computer prompting costs more but is also more effective than unprompted. The majority of data points fall in the top half of *Figure 3*, indicating, as observed earlier, that computer prompting is cost-increasing compared with unprompted. It is also important to note that, looking at the top half of the figure, a considerable number of the data points fall on either side of the y axis (cost difference), which suggests that there is no significant difference between prompted and unprompted in terms of effectiveness (based on study 2 data). Overall, *Figure 3* shows that, based on study 2

results, computer prompting is generally cost-increasing, but there is uncertainty over the direction of effectiveness between prompted and unprompted.

Figure 4 shows the cost-effectiveness acceptability curves for computer prompting for both life-years (dotted line) and QALYs (based on study 2 data). The figure shows that, if a decision-maker was prepared to pay £30,000 per QALY (the perceived threshold value of cost-effectiveness in the UK), there is a 50% probability that prompted is more cost-effective than unprompted reading. Conversely, there is a 50% probability that unprompted is more cost-effective than prompted reading. Further interpretation of these results is discussed in Chapter 5 (section 'Economic assessment', p. 37).

Chapter 5

Discussion

The aim of this project was to assess the potential of CAD as an aid for the screening programme. Specifically, the intention was to test two hypotheses: (1) that a single radiologist supported by CAD would perform as well as two radiologists double-reading; and (2) that a radiographer using CAD performed as well as a radiologist. If there was a significant impact due to the prompts such comparisons could provide valuable information that could form part of an economic assessment of the likely value of CAD. However, if there was no significant impact due to CAD, although useful data could be obtained, if the study failed to detect an impact due to CAD, no direct assessment of its impact could be made. The study could therefore be viewed as having two underlying questions, the primary question being whether or not the CAD prompts improved the film readers' sensitivity, with a secondary question being whether or not it had a differential impact on radiologists and radiographers. It therefore seemed sensible to attempt to answer these questions first, since this would maximise the study power. In the rest of this chapter the results are considered, first in relation to the primary question and then in relation to the secondary question. Some other lessons learnt about CAD and about the evaluation of software in rapidly changing fields are also noted.

Impact on readers' sensitivity and specificity

In study 1, no evidence was found that the prompts provided by the R2 ImageChecker affected readers' sensitivities or specificities. To ensure that the set was challenging, a number of false-negative interval cancers was included. The set of 60 cancers was made up of 20 false-negative interval cancers and 40 screen-detected cancers. Cancers of each category were taken at random from those available in a participating screening centre. The resulting set was felt to be an appropriate test, including both obvious and very subtle cancers. However, analysing this set of 60 cancers and the data from study 1, the 60 cases can be divided into three groups. There are 15 too-difficult cases, where there is no possibility of detecting an improvement due to prompting

because the ImageChecker failed to prompt the cancer. There are 31 too-easy cases, where there is no real scope for detecting an improvement because more than 90% of readers detected the cancer in the unprompted condition. There are, therefore, only 14 cases on which there is any real scope for detecting improvements. Looking at just these cases, the power of the study is greatly reduced (simulations suggest a power of around 50% to detect a 5% improvement in sensitivity).

Therefore, a second study was carried out using a data set of 120 cases that contained 40 different cancers that were missed by at least one radiologist at screening, but which were prompted by the R2 ImageChecker. Thirty-five film readers (all of whom had participated in study 1) read the films in this set both with and without computer prompts. The results of this study seem to confirm those of study 1. Although there was an effect, also found with cases from study 2 that met the criteria for study 1, whereby the prompts improved both sensitivity and specificity, it was small and only the impact on specificity was statistically significant. Using the data from this study to simulate double-reading showed a significant improvement of double-reading over single-reading without CAD. This suggests that the study was adequately powered.

The authors can therefore be confident that the computer-aided prompts provided by this version of the ImageChecker do not have a significant impact on film readers' sensitivity, certainly when reading these kinds of artificial roller under test conditions.

This is consistent with other results reviewed in Chapter 1 (section 'Tests of the impact of systems on radiologists' decision-making', p. 5), such as those of Brem and Schoonjans,³² but runs against the general tenor of findings in evaluations of prompting systems. Studies such as those by Funovics and colleagues²⁹ or Thurjfell and colleagues²⁸ have found in favour of the ImageChecker. These studies were, however, much less powerful than the present study and should not be regarded as demonstrations of an effect due to prompting. The present authors believe strongly that conclusions should not be drawn

from studies involving only two or three radiologists. The current negative results also contrast with some of the earlier studies conducted in university research laboratories which used ROC analysis to investigate the impact of prototype prompting systems, for example Chan and colleagues.³³ This may reflect a difference in the analysis. An increase in the area under the ROC curve may reflect a slight impact of the prompts on the confidence of the decision-maker over a range of operating points, rather than a significant change in their behaviour with respect to a real clinical decision.

The rest of this section considers some of the possible explanations for this result, focusing first on three characteristics of the tool that may have undermined performance, and then on the possible weaknesses of the research paradigm used.

Possible explanations for the negative result

Insensitivity to subtle cancers

Study 2 used cases that had been missed in the past by the first reader in normal double-checking. Checking the records for 1 April 1999 to 31 March 2000, 262 screen-detected cancer cases were found, 31 of which were missed by the first reader, but only 19 (61%) of which were prompted by R2. In other work this group again found a sensitivity of 60% to missed cancers.⁵ These missed cancers included 19 out of 35 classified as having 'minimal signs', so a low figure is perhaps unsurprising. Nevertheless, the tool needs to be able to detect the kinds of cancers that are currently missed in the screening programme if it is to be valuable.

Most of the current algorithms used in CAD analyse each mammogram separately. Yet the complexity and variability of mammographic appearances are so great that there can be no absolute basis for the detection of abnormalities. The task cannot be reduced to a simple set of pattern recognition exercises. Radiologists know this; they are taught that the most effective basis for assessing a mammogram is to compare it with another mammogram from the same woman. Radiologists compare current with previous films and look for signs of change, and compare left and right breasts and look for asymmetries. Significant progress in CAD now requires that these two comparison tasks are tackled.

There is evidence that the detection of asymmetries between left and right breasts merits particular attention. Blanks and colleagues looked

at missed cancers where a visible radiological abnormality was misclassified as benign and found that ten out of 24 (42%) misinterpreted invasive cancers were asymmetries.⁷⁰ In five out of the 15 cancers missed by R2 in study 1 the primary radiological sign was an asymmetry. In other work it was found that on ten out of 35 missed cancers the primary abnormality was an asymmetry and the R2 ImageChecker failed to prompt five out of ten asymmetries.⁵ Although asymmetries are much less common than microcalcifications and masses, they are much harder to interpret and are therefore overrepresented in cases of cancer missed at screening. Given the current state of play in CAD, they therefore represent a significant challenge for future research.

Specificity of the prompts

The authors believe that around 25% of radiologically visible cancers are missed at screening.⁵ If the number of cancers detected per 1000 cases is 6.7, the number of visible cancers would be 8.9, with the number missed being 2.2. However, two facts need to be taken into account when assessing the potential contribution of CAD to the detection of these missed cancers. First, there is evidence that at least 40% of false-negative interval cancers correspond to lesions that were seen by radiologists and misclassified as benign.⁷⁰ Low specificity prompts are unlikely to affect decision-making in these cases. Second, the present data suggest that the sensitivity of CAD on false-negative interval cancers is as low as 60%. Taking these two facts into consideration, the number of prompts for cancers that would otherwise not be detected is probably less than one prompt per 1000 cases.

A radiologist reading 5000 cases per annum will, in a year, see 10,000 prompts, of which maybe only five will be prompts for cancers that he or she had not detected. If every radiologist detected five additional cancers per annum, this would be a significant intervention, but the likelihood is that these prompts will be so swamped by false positives that they will fail to generate a recall decision.

The latest version of the ImageChecker, which only became available after this research was completed, is able to provide the sensitivity of earlier versions with 30% fewer prompts. This may enhance its impact on film readers' sensitivity.

CAD as an aid to decision-making

Comments made by participants in this study suggest that the current low specificity of the

prompts means that they are not valuable as aids to decision-making. This is significant because some authorities suggest that as many as half the cancers missed at screening are missed not because the radiologist failed to spot the abnormality, but because he or she made the wrong decision in deciding whether or not to recall it.⁷⁰

Film readers' self-reports indicate that, although they questioned the value of the tool in general, they did believe that it improved their decision-making on certain classes of cases, notably microcalcifications. The experimental data do not, however, bear this out. It was not the case that prompts for microcalcifications had a greater impact on decision-making than other prompts. This suggests that readers were aware of the impact on the prompts on a number of salient cases. These cases were not significant by comparison with the much larger number of cases where film readers were not aware of having changed their mind in between the two conditions. It is clearly important to interpret film readers' self-reports carefully.

The data from study 2 were analysed on a case-by-case basis. The analysis suggests that the prompts have a greater impact in more difficult cases. The positive impact of CAD in this study was greater on specificity than that for sensitivity, suggesting that the film readers were using the prompts to help in their decision-making, but not in the way one might expect. Taking these conclusions together with the responses of participants, it seems that film readers were more inclined to use the absence of prompts as confirmation that a case was benign than to rely on them as indicators of abnormality.

Further work on improving the role of CAD as an aid to decision-making as well as detection is clearly required. Karssemeijer and colleagues simulated combining radiologists' judgements with the output of CAD algorithms and their results suggest that improved decision-making would result, but there is a need to show that this can be done in practice.⁷¹ New implementations of CAD, such as the latest version of the ImageChecker, give the user more information about the degree of certainty associated with a prompt, and this may have an impact on their role as decision aids.

Weaknesses of the research paradigm used

The authors would like to stress the caveat that the conclusions of this work are only valid if the method is sound, and they are aware of its

limitations. Rutter and Taplin compared test and clinical performance in a generic mammographic interpretation task (not involving computer prompts) and found no correlation between accuracy in the two conditions and only weak correlations in overall preponderance to call a mammogram positive.⁷² Their study was limited by the small number of film readers involved, but clearly the present method is imperfect. If the method is fatally flawed, then a great deal of radiological research as well as quality assessment exercises such as the PERFORMS self-assessment programme are also fatally flawed.

A number of specific criticisms can be made of the study. The researchers have listed those of which they are aware and made some observations about their likely significance.

- The readers had only limited experience of using the machine. They were, however, given a tutorial based on the material that the company provides for new users and, in addition, they all completed a practice session before data collection began. The authors do not believe that their posited inexperience had a significant impact on the results. One might expect that inexperienced users might rely too much on the prompts and unnecessarily recall benign patients. This did not seem to be the case.
- The sample of films was small and heavily weighted towards cancers. The sample sizes used were justified in the power calculations made in the original proposal and the subsequent request for an extension. They are comparable to those in studies of other interventions. It is true that, as in many other studies, the sample was heavily weighted towards cancers. Not weighting samples makes this form of research extremely expensive. This weighting could have affected the expectation that a prompt was associated with a cancer. If this were true, then it would have appeared to improve the specificity of the system, since a higher proportion of cancers implies a lower proportion of normal cases and less scope for false prompts. Despite this, poor specificity was the principal difficulty that the readers had with the system.
- Using the study 1 data to simulate double-reading did not show an improvement over single-reading. The performance of the readers was surprisingly strong and consistent, with the effect that pairing readers to simulate double-reading did little to improve overall sensitivity. It was concluded from an analysis of these data that, to carry out a fair evaluation, the study

needed to look at a further 120 cases, including 40 cancers specifically selected to maximise the potential for an improvement due to CAD. Analysis of the data from study 2 showed an improvement when double-reading was simulated.

- There was not a great difference in reading times for the two conditions. It is difficult to infer too much from the timing data in this study because readers in both conditions were not just looking at films but also recording their observations on detailed data entry forms. The form was also used as the prompt sheet, in the prompted condition. One could argue that much of the additional work that readers had to perform in the prompted condition was subsumed within the data collection tasks that the experiment required for both conditions. It is true that readers in the prompted condition still had to take a second look at the image if any of prompts suggested additional areas of search. The timing data show that users did not take much longer on cases when the prompts were available. It may be that they did not spend as long as one might expect on the second look. However, the system has to be judged as it is actually used, not as the manufacturers might like it to be used. No other study of CAD has used such a large sample of readers and nothing suggests that their behaviour is atypical.
- There was also concern that the relatively small interval between readings in the test and control conditions may have reduced the effect. If this were the case one would expect that CAD would have had a greater effect on those who were randomised to complete the intervention condition after the control condition. However, analysis of the data shows that there is no significant effect attributable to the order in which the conditions were completed.

Film readers participating in this study were not reading under normal conditions. Their vigilance, concentration and decision-making thresholds would all have been affected by the knowledge that the balance of cancers and normals was artificial and that the results were being used in an experiment. The authors accept that CAD may have an effect in routine use that will not be detected in studies such as this. However, it is likely that if the impact of the prompts was as great as some previous studies have suggested, then an effect would have been observed on the data. The unavoidable conclusion of study 2 is that film readers will ignore a large percentage of correctly placed prompts.

A fuller understanding of the impact of the prompting system requires a study of a very different type. There are highly significant questions that cannot realistically be answered using archive data. These concern issues to do with the practicality of integrating the computer processing of films into the workflow of a busy screening unit, as well as issues such as that raised above about the impact on the recall rates of radiologists and radiographers using the computer-aided prompts for data sets with realistic frequencies of cancers. A further set of questions can only be answered by a study in which participants use the machine over a substantial period: is there a learning effect for the use of the machine?

Questions such as these can only, realistically, be answered by an evaluation of the impact of a CAD system in routine use. The researchers are, therefore, replicating the prospective study of Freer and Ulissey with a larger group of radiologists and radiographers using double-reading in the context of a UK screening programme.³⁴ Ethical permission for this study was obtained on the basis of the results of study 1, showing that use of the ImageChecker was unlikely to be detrimental to patient care. This is important because of the differences between the UK and US screening populations and processes, and also because the findings of the Freer study may be regarded as inconclusive given the high proportion of DCIS in the additional cancers. The initial cancer detection rate in the Freer study is very low compared with that in the NHSBSP, perhaps reflecting the lower incidence of cancer in their younger population, and the increase in recall rate found by Freer and Ulissey would have serious consequences for the UK screening workload. It is worth noting that although this will be a prospective trial, it will not be an RCT, and although it may reveal that there is an effect attributable to CAD it will not provide unambiguous evidence of the size of the effect.

Why are the prompts ignored?

On 22% of the observations of readers looking at correctly prompted cancers in study 2, the case was not marked for recall, despite the presence of an accurate prompt. Several factors may affect whether users respond to prompts. The prompts seemed to have more influence in difficult cases. It is, however, difficult to assess this if the same data that are used to assess impact are also used to assess difficulty. An average of only 54% of readers

call the most difficult ten cases correctly, and the difference between the prompted and unprompted conditions in these cases averages 9.8%. There is a clear correlation between failure to recall a prompted cancer and the confidence with which the case was dismissed in the unprompted condition.

Film readers' self-reports indicate that, although they questioned the value of the tool in general, they believed that it improved their decision-making in certain classes of case, notably microcalcifications.⁷³ The experimental data did not, however, bear this out. There is no evidence of a difference between the impact of prompts for calcifications compared with masses. One factor that seemed to affect the impact of the prompts was the use of a circled prompt to indicate increased confidence. Although these emphasised prompts were not always correct, readers were much less likely to ignore them.

Impact on professional groups

No significant difference in recall rate was found between different professional groups of reader ($p = 0.2$). This is consistent with other studies, such as those of Haiart and Henderson¹⁰ and Cowley and Gale, which have showed that trained radiographers perform as well as radiologists.¹² Pauli and colleagues also found that radiologist/radiographer double-reading gave similar sensitivities to double-reading by radiologists.¹¹

The readers were divided into two groups based on their average sensitivity scores, to see whether CAD had a greater effect on readers with lower sensitivities. No evidence was found that use of the ImageChecker affected more and less able readers differently.

Cowley and Gale reported that both breast clinicians and radiographers are slower than radiologists at reporting films.¹² This was also found in the present study. However, no significant difference was found in time taken to read films in either the prompted or the unprompted condition ($p = 0.6$).

Overall, the results provide strong supporting evidence for the claim, now quite widely accepted, that appropriately trained radiographers can play a valuable role in interpreting screening films. Manufacturers of CAD systems have in the past been wary of attempts to use CAD as an aid for untrained film readers. The authors' experience is

that the grounds for this anxiety do not apply to trained radiographers working in the UK screening programme.

Modelling the impact of CAD on the screening programme

The data from study 1 were used to run a simulation of double-reading with arbitration. The results of double-reading without CAD were then compared with single-reading with and without CAD. The original intention had been to present this simulation as a test of the hypothesis that single-reading with CAD was equivalent to double-reading without CAD. In practice, the results of the simulation are undermined by the fact that no difference was found between reading with and without CAD. Therefore, although the simulation appears to show that double-reading is no better than single-reading with CAD, at least in terms of sensitivity, one should not be too trusting of this result given that the simulation also appears to show that double-reading is no more sensitive than single-reading. The conclusion may be drawn that the real differences between the conditions are not revealed in these data since the performance of the film readers was less variable than in real life, an effect that will tend to erode the benefit due to double-reading.

Repeating the exercise with the study 2 data, an impact due to double-reading was shown, which was greater than that for single-reading with CAD. It is difficult to interpret the results of this simulation, however. Looking at the three conditions, single-reading, single-reading with CAD and double-reading, there is only one difference that is statistically significant: that between single-reading and double-reading. It does not follow from this that single-reading with CAD is equivalent to double-reading, although it should be noted that the figures obtained for sensitivity in these two conditions were close.

Economic assessment

There are several limitations with the methodological approach adopted. The estimates of time costs associated with computer prompting are based on experimental conditions and, were computer prompting to be adopted in routine screening, the time costs may be very different. However, reading time costs do not represent a large proportion of the change in total cost between prompted and unprompted reading.

TABLE 30 League table of cost-effectiveness estimates of changing the number of views/reading policies in breast screening in the UK

Study (options compared)	Additional costs per 1000 women screened (2001/02)	Additional cancers per 1000 women screened	Additional costs per additional cancer detected (2001/02)
Non-consensus double-reading versus single-reading ⁵³	£2,168	1	£2,168
Non-consensus double-reading versus single-reading ⁵²	£1,862	0.66	£3,387
Two view versus one view ⁵⁰	£2,796	0.5	£5,523
Two view versus one view ⁴⁹	£9,532	1.32	£7,221
Two view double-reading versus one view double-reading ⁵¹	£4,796	0.6	£7,993
Prompted versus unprompted (study 2)	£4,538	0.195	£23,272
Prompted versus unprompted (study 2 virtual roller)	£3,848	0.3	£29,600

The study 1 simulation of double-reading found that the costs of double-reading (one radiologist and one radiographer) were lower than those of single-reading (one radiologist with CAD). If the double-reading option using one radiologist and one radiographer could improve its levels of sensitivity and specificity relative to single-reading, then having radiographers as second readers may provide a better alternative than computer prompting and be able to address the shortage in radiologists.

Although there are important quality of life effects associated with screening, these were not included in the model of cost per cancer detected. Given, however, that the cost per cancer detected is based on a 12-month duration, inclusion of quality of life effects in the model would have a very small impact on the overall outcome.

The model of cost-effectiveness did incorporate quality of life effects, but it has a number of limitations. One is that transition probabilities between types of recurrence and between recurrence and death were not available by PG and were assumed to be constant across PGs. A further limitation is that patient-specific resource use (and hence costs) could not be estimated and consequently treatment cost estimates had to be based on average treatment protocols. This meant that variation in cost within PGs could not be addressed. By using probabilistic sensitivity analysis, however, variation in costs and transitions was addressed.

The model assumed that all false-negative cases arise in the first year of the 10-year model. This assumption was made because there is no other information on the timing of false negatives. Although the timing of detection was not addressed in the model, the differences in the

number of false negatives between prompted and unprompted were taken into account.

Despite these limitations, however, the modelling approach used here is an advance over previous studies that have either not modelled life-years gained or used crude approaches.

Comparisons with cost-effectiveness of other technologies

The relative cost-effectiveness of computer prompting can be judged by comparing the cost per cancer detected with computer prompting with the cost per cancer detected estimated in other breast screening studies in the UK. *Table 30* presents a league table of estimates of cost-effectiveness from a number of studies exploring the cost-effectiveness of the number of views/reading policies in breast screening in the UK. The league table presents the additional costs, additional cancers and the additional cost per additional cancer detected (cost-effectiveness). Cost estimates from the original studies have been updated using the Hospital and Community Health Services index (as reported in Netten and Curtis⁷⁴). Notwithstanding the fact that the studies have used different methodologies to estimate costs and cancers detected, the league table provides a useful comparison of results across studies. For comparison purposes, *Table 30* also shows the cost-effectiveness of computer prompting estimated from this study.

Comparing the previous estimates of cost per cancer detected with the results from this study shows that computer prompting has a higher cost per additional cancer detected than previous studies (which ranged from £2168 to £7993 per additional cancer detected). The number of additional cancers detected with computer prompting is, however, much lower than the

number detected in previous studies. Computer prompting detects at best 0.195 cancers per 1000 women screened and at worst zero additional cancers detected, compared with a range of 0.5–1.32 cancers detected per 1000 women screened in previous studies. This suggests that the high cost per cancer detected from computer prompting arises from the fact that computer prompting does not detect many additional cancers, rather than from its additional costs. A high cost per cancer detected compared with other studies suggests that computer prompting does not represent good value for money.

The relative cost-effectiveness of computer prompting can also be judged by estimating the probability of computer prompting being cost-effective relative to threshold cost-effectiveness ratios. Long-term cost-effectiveness at 10 years for study 2 found that the probability of computer prompting being cost-effective was 0.50 at a threshold ratio of £30,000 per QALY. This can be compared with a recent estimate of the probability of cost-effectiveness for a screening intervention (aortic aneurysms) which reported screening for aneurysms to have a 0.55 probability of being cost-effective at 4 years, but at 10 years the cost-effectiveness was significantly greater.⁷⁵ Even if a decision-maker were willing to pay twice as much as perceived thresholds of cost per QALY, which they may be given the shortage of radiologists, the probability of computer prompting being cost-effective only rises to 0.53 as a result of the uncertainty surrounding its effectiveness relative to unprompted.

Overall, the cost and cost-effectiveness analysis has shown that computer prompting is cost-increasing and the cost-effectiveness is uncertain. For study 1, no difference in sensitivity and specificity was detected, suggesting that computer prompting is cost-increasing and provides no additional benefits, and is therefore not cost-effective. For study 2, as a result of the small but non-significant difference in sensitivity and specificity between prompted and unprompted reading, additional cancers were detected, but even the magnitude of these was small compared with other mammographic reading interventions. For study 2, the cost per QALY gained of computer prompting relative to unprompted is uncertain, reflecting the non-significant differences in effectiveness between prompted and unprompted reading. It can be concluded, therefore, that until computer prompting can demonstrate a significant improvement in sensitivity, the cost-effectiveness of computer prompting will remain uncertain.

Usability of CAD

No significant problems relating to the digitisation of the films or their display occurred during the evaluation. Neither work done by the study team nor the ergonomic evaluation performed by the team from Edinburgh revealed any problems suggesting that the digitisation and display of films for use with CAD could not be integrated into workflow.

The project revealed two issues that may pose significant threats to the acceptability of CAD. The first is the lack of reproducibility. The nature of this problem needs to be set out with some care. Each time a film is put through the digitiser a new digital image is created. If the same film is put through twice the resulting digital images will appear identical but will be different, since the alignment of fine detail with the digitising array will have changed. The fact that the algorithms will respond to the digital images differently is surprising to radiologists and causes concern. The problem is not that the lack of reproducibility is evidence that the device is performing less well than the manufacturer's claim. There is no reason to doubt the figure of 86% sensitivity. The problem is, rather, that it shows that the behaviour of the system will always be unpredictable (one can never learn which cancers will be the missed 14%) and that this represents a challenge to those working with it. In the authors' judgement this is a difficulty, but not a conclusive argument against the use of the machine.

The second point concerns the users' understanding of the basis for the system's operation. There are numerous occasions in modern medicine where clinicians have to use tools of such complexity that they must be regarded more or less as black boxes. This can be problematic if the output from the system is in effect a measurement that has only statistical validity and which then has to be used by the clinician in a decision that is made in part on the basis of clinical judgement. Users were uncomfortable with the fact that they could not explain some of the system's prompts. Readers need to have an idea of the ways that the system behaves and what it has prompted and why. The training given to users must provide the basis for this understanding, but needs to be grounded in the radiologists' way of working and not an attempt to explain the technology. A working understanding, however, is only likely to emerge through familiarity with the machine. Research on users who are not well acquainted with the tool may not, therefore, be valid.

Evaluation of rapidly changing technology

The proposal for this research was written almost exactly 5 years before the submission of the final report. That delay is clearly inappropriate in a field where technology changes suddenly and dramatically. One of the most robust findings of this work is that the low specificity of the prompts means that they fail to have the impact one might expect on readers' sensitivities. On 13 March 2003 a new version of the R2 ImageChecker was installed at St George's NHS Trust that generates 30% fewer false prompts than the version of the software evaluated in this study. Clearly, this will go some way to rectifying the problem, but to delay the report while a further study was designed, performed, analysed and written up would have been inappropriate.

The evaluation of rapidly changing technologies requires (1) short, focused studies that answer specific questions about specific systems, (2) thorough experimental studies to answer underlying questions about the use of such tools, and (3) much greater use of modelling exercises to establish the key performance parameters. The existing funding mechanisms of the HTA programme seem inappropriate for studies of type (1). These kinds of evaluation would perhaps be best supported by some other mechanism. In the case of CAD the obvious approach would seem to be to allow evaluations to be funded directly by the NHSBSP.

The prospects for CAD

It is clear that the performance of the evaluated CAD tool does not allow its recommendation for any kind of role within the NHSBSP on the basis of these results. That should not, however, be the end of the story. The North American market for CAD is likely to remain buoyant. CAD manufacturers will therefore continue to develop and market their products, and these will be continually reassessed for their potential as aids to the NHSBSP. Three factors seem particularly important to consider in assessing the future for CAD in the UK.

Further research

The authors are well aware of the limitations of their work and that several recent prospective trials in the USA have shown some improvement attributable to the use of CAD. These trials are not full RCTs and it is difficult to assess the size of the effect. As explained above, their methods may exaggerate the effect due to CAD and their results

may not be applicable to the UK setting. Nevertheless, there is a real need to replicate these studies in the UK. This would show whether use of CAD over a certain period with realistic frequencies of cancers would allow users to take more advantage of the prompts. The research team is carrying out one such study and their colleagues intend to carry out other studies.

Full-field digital mammography (FFDM)

FFDM machines are now available from a number of manufacturers, notably GE and Lorad. These machines are currently five to six times as expensive as conventional analogue machines and their purchase is therefore difficult to justify on financial grounds. The image quality is, however, felt by many to be superior. The spatial resolution is not as high as film (although the difference is relatively small in the case of the Lorad machine), but the other characteristics of the image are better: detector quantum efficiency is higher, allowing better signal-to-noise ratio, improved resolution of low-contrast detail and lower doses. The running costs of the machines are lower if cases are read on monitors, avoiding the costs of processing and printing. Patient throughput is also better. Fewer films are lost and fewer repeats required. These advantages however, have to be set against significantly higher capital costs, and the technology is too recent to allow a fair assessment of the total costs over the lifetime of the equipment. The NHSBSP has recently made a substantial investment in analogue machines that will have a lifespan of at least 10 years, suggesting that the take-up of digital will be faster in symptomatic clinics than in screening centres.

Many of these centres will choose to purchase CAD modules as part of the FFDM reading workstations. The ergonomics of CAD in an FFDM environment are very different (films do not need to be digitised). CAD helps to overcome the difficulties of detecting subtle calcifications in images with a lower spatial resolution. The higher cancer detection rate in symptomatic clinics means that the low specificity of the prompts is less of an issue.

This experience with the use of CAD is likely to help to overcome some of the weaknesses identified earlier in this report.

Improvements to CAD

The CAD algorithms continue to be developed and improved. The latest version of the R2 ImageChecker has significantly fewer false prompts than the previous one. There are still interesting and challenging areas of research to be

tackled that will allow further improvement, notably temporal registration. Since the regulation of software as a medical device is much weaker in the European Union than in the USA, it is possible that European centres will gain faster access to new releases of software than their US colleagues. Researchers in the UK should be encouraged to make sure that the early evaluations of new releases are performed in NHSBSP centres.

Conclusions and recommendations for future research

This section presents the final conclusions of this work, and a series of recommendations for further research is made. The authors indicate, for each recommendation, whether they believe that the work is of immediate importance or of longer term significance.

The paradox of CAD is that prompts seem not to have a strong impact on film readers' performance, despite their high sensitivity. The authors believe that this is due to their low specificity, their relatively poor sensitivity for subtle cancers and the fact they cannot serve as aids to decision-making.

Improvements are needed in the overall specificity of the prompts and in the sensitivity of the prompts to subtle signs, such as asymmetries. Improving the specificity of CAD systems is likely to have a more immediate impact than research aimed at the detection of subtle lesions, which will only contribute to improvements in performance if the resulting algorithms can be incorporated without any overall loss in specificity. The accuracy of the algorithms is likely to improve as a result of incremental changes made by the manufacturers on the basis of experience currently being accumulated. One research goal that may help in this is the registration of pairs of mammograms (current and previous, left and right breast, craniocaudal versus mediolateral oblique) in a way that will allow automated comparison of images, for example to detect temporal change or asymmetry. Solving this problem is likely to enhance the accuracy of CAD substantially.

- **Research recommendation (longer term significance): the improvement of sensitivity to subtle cancers, for example through the registration of temporal sequences of mammograms.** Perhaps the most significant weakness of CAD is that the prompts do not provide a guide to decision-making. They are

intended to act as prompts, to guide film readers' attention, not as pieces of evidence to be taken into account in making decisions. However, a high percentage of the cancers 'missed' at screening are actually detected but are not recalled because incorrect decisions are made on the basis of ambiguous evidence. Future research could investigate the decision processes of film readers. It would be interesting to see whether film readers could make effective use of the information that image analysis algorithms could provide about the likelihood of abnormality at image locations. Such research could include investigation of the range of individual differences and of the extent to which algorithms can be tuned to reflect the needs of different individuals.

- **Research recommendation (longer term significance): the assessment of the potential of image analysis as an aid to clinical decision-making.** Although the case for CAD as an element of the NHSBSP is not made at the current time, further evaluative research is still required. The authors are aware of the limitations of their results. It may be that readers would be better able to make use of the prompts if they had longer to become accustomed to working with them. The prompts may have an impact in routine use that is not detectable in an experimental setting. Evaluations of the impact of CAD tools in routine use are already under way and their results should be given careful attention. In considering these conclusions it is also worth noting that the algorithms used in the ImageChecker are periodically revised and the machines upgraded. Improvements in the detection software are likely to have a significant impact on the usefulness of the device. It is also clear that the market for CAD will be significantly enhanced by the progressive move towards digital mammography, a move that has already begun in symptomatic clinics.
- **Research recommendation (immediate significance): that new releases of CAD software be assessed for their value as aids in breast cancer screening. Future evaluations should consider prospective designs.** One further recommendation may be made: the NHS should also consider carefully its approach towards the assessment of technologies such as CAD. There should be a clearer route to funding of rapid evaluations. Other studies should attempt to make more lasting claims answering underlying questions affecting the impact of tools. Greater use should be made of mathematical models and simulations.



Acknowledgements

We would like to acknowledge the contributions of all the film readers who participated. The work was funded by the NHS HTA programme. The support of Dr Donna Christiansen, Dr Regina Pauli and Dr Julie Cooke, who all participated as members of a Steering Group, is also gratefully acknowledged.

The authors are grateful to the British Institute of Radiology for permission to use Tables 1–5 from the following publication: Taylor PM, Champness J, Given-Wilson RM, Potts HW, Johnston K. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. *Br J Radiol* 2004;**77**:21–7.

Contributions of authors

Dr Paul Taylor (Senior Lecturer) was the principal investigator and guarantor of the overall integrity of the study. He had overall responsibility for design and management of the experimental studies, and was the primary author of the final report.

Ms Jo Champness (Service Improvement Facilitator) was the project coordinator, with

responsibility for recruiting and training participants, running the experimental studies and data collection.

Dr Rosalind Given-Wilson (Consultant Radiologist) was the lead clinician. She shared responsibility for design and management of the studies and had the primary responsibility of collating and annotating materials. She guaranteed the security of patient data, liaised with ethics committees and contributed to the writing of the report.

Dr Katharine Johnston (Health Economist) was responsible for economic evaluation. She also contributed to the design of experiments, advised on running the project and contributed to the writing of the report.

Dr Henry Potts (Statistician) was responsible for statistical analysis. He also contributed to the design of experiments, advised on running the project and contributed to the writing of the report.



References

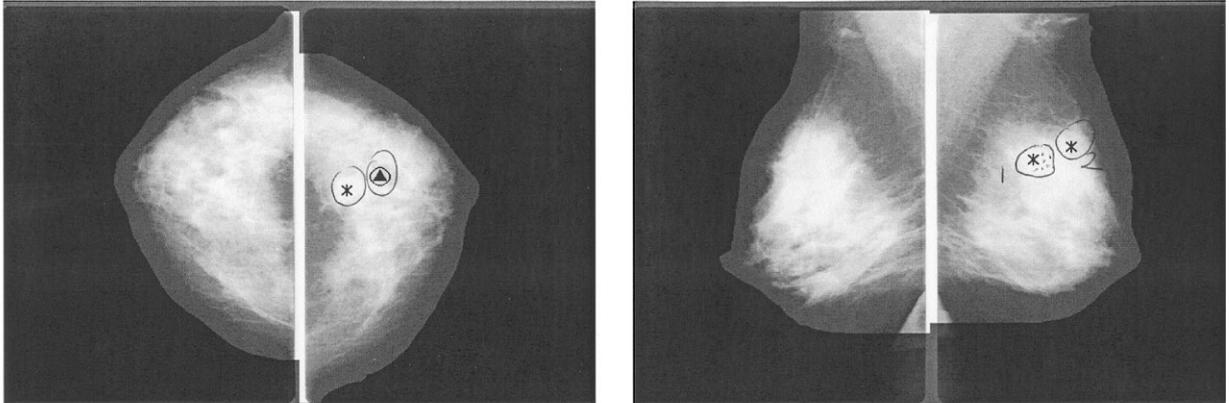
1. Forrest P. *Breast cancer screening. Report to the Health Ministers of England, Wales, Scotland and Northern Ireland*. London: HMSO; 1986.
2. NHS Cancer Screening Programmes. The NHS Breast Screening Programme. URL: <http://www.cancerscreening.nhs.uk/breastscreen/>. Accessed 20 March 2003.
3. Blanks RG, Moss SM, McGahan CE, Quinn MJ, Babb PJ. Effect of NHS breast screening programme on mortality from breast cancer in England and Wales, 1990–8: comparison of observed with predicted mortality. *BMJ* 2000; **321**:665–9.
4. NHS Cancer Screening Programmes. NHS Breast Screening, 2001 Review. URL: <http://www.cancerscreening.nhs.uk/breastscreen/publications/2001review.html>. Accessed 20 March 2003.
5. Taylor P. Computer aids for detection and diagnosis in mammography. *Imaging* 2002; **14**:472–7.
6. BBC News, Report on breast cancer screening delays. URL: <http://news.bbc.co.uk/1/hi/health/2240815.stm>. Accessed 20 March 2003.
7. Field S. UK Radiologist Workforce Survey – Breast Imaging Service. *Royal College of Radiologists Newsletter* 1998; **54**:12–14.
8. Department of Health. *The NHS Plan*. London: HMSO; 2000.
9. Department of Health. *The NHS Cancer Plan*. London: HMSO; 2000.
10. Haiart D, Henderson J. A comparison of interpretation of screening mammograms by a radiographer, a doctor and a radiologist: results and implications. *British Journal of Clinical Practice* 1991; **45**:43–5.
11. Pauli R, Hammond S, Ansell J. Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening. *J Med Screen* 1996; **3**:18–22.
12. Cowley H, Gale A. *PERFORMS and mammographic film reading performance: radiographers, breast physicians and radiologists*. A report for the Workforce Issues in the Breast Screening Programme Meeting. Derby: Institute of Behavioural Sciences, University of Derby; 1999.
13. Winsburg F, Elkin M, Macy J, Bordaz V, Weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology* 1967; **89**:211–15.
14. Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. *Eur J Radiol* 1999; **31**:97–109.
15. Ciatto S, Del Turco MR, Risso G, Catarzi S, Bonardi R, Viterbo V, *et al.* Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography. *Eur J Radiol* 2003; **45**:135–8.
16. URL: <http://www.r2tech.com/new/030408.html>. Accessed 11 April 2003.
17. Highnam R, Brady M. *Mammographic image analysis*. Dordrecht: Kluwer Academic; 1999.
18. URL: <http://www.r2tech.com/new/030408.html>. Accessed 11 April 2003.
19. URL: <http://www.r2tech.com>. Accessed 11 April 2003.
20. URL: <http://www.cadxmed.com/>. Accessed 11 April 2003.
21. URL: www.cadvisionmt.com/home.html. Accessed 11 April 2003.
22. URL: www.icadmed.com. Accessed 11 April 2003.
23. Warren Burhenne LJ, Wood SA, D’Orsi CJ, Feig SA, Kopans DB, O’Shaughnessy KF, *et al.* Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000; **215**:554–62.
24. URL: <http://www.r2tech.com/prf/prf001.html>. Accessed 11 April 2003.
25. URL: http://www.cadxmed.com/clinical_data/trial_summary.html. Accessed 11 April 2003.
26. Nelson M, Lechner M. Comparison of three commercially available FDA approved computer-aided detection (CAD) systems. *Proceedings of RSNA 2002*. URL: <http://shows.rsna.org/rsna2002/V40/conference/session.cvn?eID=3103179>. Accessed 28 March 2003.
27. Destounis SV, DiNitto P, Logan-Young W, Bonaccio E, Zuley ML, Willison KM. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. *Radiology* 2004; **232**:578–84.
28. Thurfjell E, Thurfjell G, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. *Acta Radiol* 1998; **39**:384–8.
29. Funovics M, Schamp S, Lackner B, Wunderbaldinger P, Lechner G, Wolf G. Computer assisted diagnosis in mammography: the R2

- ImageChecker in detection of spiculated lesions. *Wien Med Wochenschr* 1998;**148**:321–4.
30. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computed assisted detection of interval breast cancers. *Eur J Radiol* 2001;**39**:104–10.
 31. Marx C, Malich A, Grebenstein U, Kaiser WA. Are unnecessary follow-up procedures induced by computer-aided diagnosis (CAD) in mammography? Comparison of mammographic diagnosis with and without use of CAD. URL: <http://shows.rsna.org/rsna2002/V40/conference/session.cvn?eID=3105590>. Accessed 28 March 2003.
 32. Brem RF, Schoonjans JM. Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. *Clin Radiol* 2001;**56**:150–4.
 33. Chan H, Doi K, Vyborny C, Schmidt RA, Metz CE, Lam KL, et al. Improvements in radiologists' detection of clustered microcalcifications in mammograms. The potential of computer-aided diagnosis. *Invest Radiol* 1990;**25**:1102–10.
 34. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001;**220**:781–6.
 35. Morton M, Whaley D, Brandt K, Amrami K. The effects of computer-aided detection (CAD) on a local/regional screening mammography program: prospective evaluation of 12,646 patients. URL: <http://shows.rsna.org/rsna2002/V40/conference/session.cvn?eID=3108646>. Accessed 28 March 2003.
 36. Young W, Destounis S, Bonaccio E, Zuley M. Computer-aided detection in screening mammography: can it replace the second reader in an independent double read? Preliminary results of a prospective double blinded study. *Proceedings of RSNA 2002*. URL: <http://shows.rsna.org/rsna2002/V40/conference/session.cvn?eID=3108329>. Accessed 28 March 2003.
 37. Pepe M, Urban N, Rutter C, Longton G. Design of study to improve accuracy of reading mammograms. *J Clin Epidemiol* 1997;**12**:1327–38.
 38. Beam CA, Layde M, Sullivan D. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. *Arch Intern Med* 1996;**156**:209–13.
 39. Anscombe FJ. On estimating binomial response relations. *Biometrika* 1956;**43**:461–4.
 40. Van der Maas P, de Koning H, van Ineveld BM, van Oortmarssen GJ, Habbema JF, Lubbe TK, et al. The cost-effectiveness of breast cancer screening. *Int J Cancer* 1989;**43**:1055–60.
 41. Hall J, Gerard K, Salkeld G, Richardson J. A cost utility analysis of mammography screening in Australia. *Soc Sci Med* 1992;**34**:993–1004.
 42. Wait SH, Allemand HM. The French breast cancer screening programme: epidemiological and economic results from the first round of screening. *Eur J Public Health* 1996;**6**:43–8.
 43. Beemsterboer PM, de Koning HJ, Warmerdam PG, Boer R, Swart E, Dierks ML, Robra BP. Prediction of the effects and costs of breast-cancer screening in Germany. *Int J Cancer* 1994;**58**:623–8.
 44. Garuz R, Forcen T, Cabases J, Antonanzas F, Trinxet C, Rovira J, Anton F. Economic evaluation of a mammography-based breast cancer screening programme in Spain. *Eur J Public Health* 1997;**7**:68–76.
 45. Lindfors KK, Rosenquist CJ. The cost-effectiveness of mammographic screening strategies. *JAMA* 1995;**274**:881–4.
 46. Salzmann P, Kerlikowske K, Phillips K. Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. *Ann Intern Med* 1997;**127**:955–65.
 47. Okubo I, Glick H, Frumkin H, Eisenberg JM. Cost-effectiveness analysis of mass screening for breast cancer in Japan. *Cancer* 1991;**67**:2021–9.
 48. Nields MW, Galaty RR. Digital mammography: a model for assessing cost-effectiveness. *Acad Radiol* 1998;**5**:S310–13.
 49. Wald NJ, Murphy P, Major P, Parkes C, Townsend J, Frost C. UKCCCR multicentre randomised controlled trial of one and two view mammography in breast cancer screening. *BMJ* 1995;**311**:1189–93.
 50. Bryan S, Brown J, Warren R. Mammography screening: an incremental cost effectiveness analysis of two view versus one view procedures in London. *J Epidemiol Community Health* 1995;**49**:70–8.
 51. Johnston K, Brown J. Two view mammography at incident screens: cost effectiveness analysis of policy options. *BMJ* 1999;**319**:1097–102.
 52. Cairns J, van der Pol M. Cost-effectiveness of non-consensus double reading. *Breast* 1999;**7**:243–6.
 53. Brown J, Bryan S, Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ* 1996;**312**:809–12.
 54. Briggs AH, O'Brien BJ. The death of cost-minimization analysis? *Health Econ* 2001;**10**:179–84.
 55. HM Treasury. *The green book: appraisal and evaluation in central government: treasury guidance*. London: TSO; April 2003.
 56. Gerard K, Brown J, Johnston K. UK breast screening programme: how does it reflect the Forrest recommendations? *J Med Screen* 1997;**4**:10–15.
 57. NHSBSP. *Computer aided detection in mammography: Working party of the Radiologists Quality Assurance Coordinating Group*. Sheffield: NHS Cancer Screening Programmes; 2001.
 58. Johnston K, Gerard K, Morton A, Brown J. *NHS costs for the breast screening frequency and age trials*.

- Health Economics Research Group Discussion Paper No. 16. Uxbridge: Brunel University; 1996.
59. Blanks R, Given-Wilson RM, Moss SM. Efficiency of cancer detection during routine screening: two view versus one view mammography. *J Med Screen* 1998; **5**:141–5.
 60. Briggs A, Sculpher M. An introduction to Markov modelling for economic evaluation. *Pharmacoeconomics* 1999; **13**:397–409.
 61. Rosenquist CJ, Lindfors KK. Screening mammography beginning at age 40 years: a reappraisal of cost-effectiveness. *Cancer* 1998; **82**:2235–40.
 62. Todd JH, Dowle C, Williams MR, Elston CW, Ellis IO, Hinton RW, *et al.* Confirmation of a prognostic index in primary breast cancer. *Br J Cancer* 1987; **56**:489–92.
 63. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. 1. The value of histological grade in breast cancer: experience from a large study with long term follow up. *Histopathology* 1991; **19**:403–10.
 64. Tabar L, Fagerberg CJ, Gad A, Baldetorp L, Holmberg LH, Grontoft O, *et al.* 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* 1985; **i**:829–32.
 65. Breast Screening Frequency Trial Group. The frequency of breast cancer screening: results from the UKCCCR randomised trial. *Eur J Cancer* 2002; **38**:1458–64.
 66. NHSBSP, Association of Breast Surgery at BASO. *An audit of screen detected breast cancers for the year of screening April 2001 to March 2002*. Sheffield: NHS Cancer Screening Programmes; 2003.
 67. National Institute for Clinical Excellence. *Interim guidance for manufacturers and sponsors*. 2000. URL: www.nice.org.uk. Accessed 28 March 2003.
 68. Blamey RW, Wilson ARM, Patnick J. Screening for breast cancer. *BMJ* 2000; **321**:689–93.
 69. van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects and C/E ratios alongside a clinical trial. *Health Econ* 1994; **3**:309–19.
 70. Blanks RG, Wallis MG, Given-Wilson RM. Observer variability in cancer detection during routine repeat (incident) mammographic screening in a study of two versus one view mammography. *J Med Screen* 1999; **6**:152–8.
 71. Karssemeijer N, Otten JD, Verbeek AL, Groenewoud JH, De Koning HJ, Hendriks JH. Computer-aided detection versus independent double reading of masses on mammograms. *Radiology* 2003; **227**:192–200.
 72. Rutter CM, Taplin S. Assessing mammographers' accuracy. A comparison of clinical and test performance. *J Clin Epidemiol* 2000; **53**:443–50.
 73. Hartswood M, Procter R, Rouncefield M, Slack R, Soutter J, Voss J. 'Repairing' the machine: a case study of the evaluation of computer-aided detection tools in breast screening. In Kuutti K, Karsten EH, Fitzpatrick G, Dourish P, Kjeld K, editors. *ECSCW 2003: Proc. of the European Conference on Computer Supported Co-operative Work*. Helsinki 2003. pp. 375–94.
 74. Netten A, Curtis L. *Unit costs of health and social care, 2002*. Canterbury: PSSRU; 2003.
 75. Multicentre Aneurysm Screening Group. Multicentre Aneurysm Screening Study (MASS): cost effectiveness analysis of screening of aortic abdominal aortic aneurysms based on four year results from a randomised controlled trial. *BMJ* 2002; **325**:1135–42.
 76. Johnston K. Modelling the future costs of breast screening. *Eur J Cancer* 2001; **37**:1752–8.
 77. Blamey RW. The design and clinical use of the Nottingham Prognostic Index in breast cancer. *Breast* 1996; **5**:156–7.
 78. Karnon J. Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. *Health Econ* 2003; **12**:837–48.
 79. Hillner BE, Smith TJ. Efficacy and cost effectiveness of adjuvant chemotherapy in women with node-negative breast cancer: a decision analysis model. *N Engl J Med* 1991; **324**:160–8.
 80. O'Rourke, S, Galea MH, Morgan D, Euhus D, Pinder S, Ellis IO, *et al.* Local recurrence after simple mastectomy. *Br J Surg* 1994; **84**:386–9.
 81. Sibbering DM, Galea MH, Morgan DAL, Elston CW, Ellis IO, Robertson JFR, Blamey RW. Safe selection criteria for breast conservation without radical excision in primary operable invasive breast cancer. *Eur J Cancer* 1995; **31**:2191–5.
 82. Robertson JFR, Whynes DK, Dixon A, Blamey RW. Potential for cost economies in guiding therapy in patients with metastatic breast cancer. *Br J Cancer* 1995; **72**:174–7.
 83. Jansen S, Stiggelbout A, Wakker P, Vlieland T, Leer J, Nooy M. Patients' utilities for cancer treatments: a study of the chained procedure for the standard gamble and time trade off. *Med Decis Making* 1998; **18**:391–9.
 84. Doubilet P, Begg CB, Weinstein MC, Braun P, McNeil BJ. Probabilistic sensitivity analysis using Monte Carlo simulation. *Med Decis Making* 1985; **5**:157–77.
 85. Briggs AH, Ades AE, Price MJ. Probabilistic sensitivity analysis for decision trees with multiple branches: use of the Dirichlet distribution in a Bayesian framework. *Med Decis Making* 2003; **23**:341–50.

Appendix I

Data collection form



Please mark and number any areas of abnormality on the images. For each feature please indicate the type of abnormality with degree of Suspicion. (1 – no significant lesion 2 – Benign 3 – Indeterminate 4 – Probably Malignant 5 – definitely malignant)

	0000525640b Irregular Mass	0000625640a Round Mass	0000525640f Calcification	Asymmetry	0000525640e Other
Example			4		
Feature 1	4		3		
Feature 2	3				
Feature 3					
Feature 4					
Feature 5					
Feature 6					

Outcome: Recall Discuss but probably recall Discuss but probably no recall No Recall

FIGURE 5 Example of a data collection form used in the study

Appendix 2

Model for a test of reading accuracy

The mathematical model used in the simulation program for calculating sample size is based on that proposed by Pepe and colleagues.³⁷ The probability of a radiologist r correctly identifying a positive film i in the control (or preintervention) condition is represented as:

$$\text{Pre}_{ri}^D = \exp\{x_{\text{pre}}^D + y_i^D + z_r^D\} / (1 - \exp\{x_{\text{pre}}^D + y_i^D + z_r^D\})$$

The probability of a radiologist r correctly identifying a positive film i in the postintervention condition is represented as:

$$\text{Post}_{ri}^D = \exp\{x_{\text{post}}^D + y_i^D + z_r^D\} / (1 - \exp\{x_{\text{post}}^D + y_i^D + z_r^D\})$$

The x parameter is derived from the sensitivity of the average radiologist on the average film and calculated for the control and intervention conditions as follows:

$$x_{\text{pre}}^D = \ln\{S_{\text{pre}}^D / (1 - S_{\text{pre}}^D)\}$$

$$x_{\text{post}}^D = \ln\{S_{\text{post}}^D / (1 - S_{\text{post}}^D)\}$$

where S_{pre}^D is sensitivity of the average radiologist on the average film in the control group (or before the intervention), and S_{post}^D is the sensitivity after the intervention.

The y parameter is used to represent the variation in the difficulty of different images and is defined as:

$$y_i^D = \ln\{U_i^D / (1 - U_i^D)\} - x_{\text{pre}}^D$$

where U_i^D is a random variable with a uniform distribution in the range $(S_{\text{pre}}^D - a^D, S_{\text{pre}}^D + a^D)$.

The z parameter represents the variation between the readers and is defined as:

$$z_r^D = \ln\{U_r^D / (1 - U_r^D)\} - x_{\text{pre}}^D$$

where U_r^D is a random variable with a uniform distribution in the range $(S_{\text{pre}}^D - b^D, S_{\text{pre}}^D + b^D)$.

The above equations are used to calculate, for each image as read by each observer, the probability of a correct interpretation. The simulation generates a random number between 0 and 1 for each observation, and tests it against the calculated probability to determine how it would be interpreted. After running the simulation for the total number of observations in both pre and post conditions, the mean and standard error of the change in sensitivities can be calculated for each reader and a t -test used to decide whether the simulation provides a positive result. Running 400 simulations gives an estimate of the power of the design with different sample sizes.

Appendix 3

Modelling the cost-effectiveness of computer prompting

K Johnston¹ and A Davies²

¹ Scottish Executive, Edinburgh, UK

² Medtap International Inc., London, UK

Long-term cost-effectiveness is estimated using a Markov model. The model is an adaptation and extension of an earlier model.⁷⁶ The main adaptation is a shortening of the time horizon from lifetime to 10 years and consequent changes to some transition probabilities. The main extension to the model is improved representation of parameter uncertainty. Data supplied by the Professorial Unit of Surgery at the City Hospital, Nottingham, UK, for the earlier study are gratefully acknowledged.

The purpose of the model is to estimate the future costs, life-years and QALYs arising from an increase in the number of cancers detected. The Markov model estimates costs, life-years and QALYs for each prognostic group (PG). The number of true-positive and false-negative cancers is estimated from the sensitivity and specificity of prompted and unprompted readings. The proportion of cancers in each PG is then applied to the number of cancers for prompted and unprompted readings. Finally, the costs, life-years and QALYs of each PG are applied to the weighted cancers for both prompted and unprompted readings.

Figure 6 shows the Markov state diagram used for each PG, with the arrows representing allowable

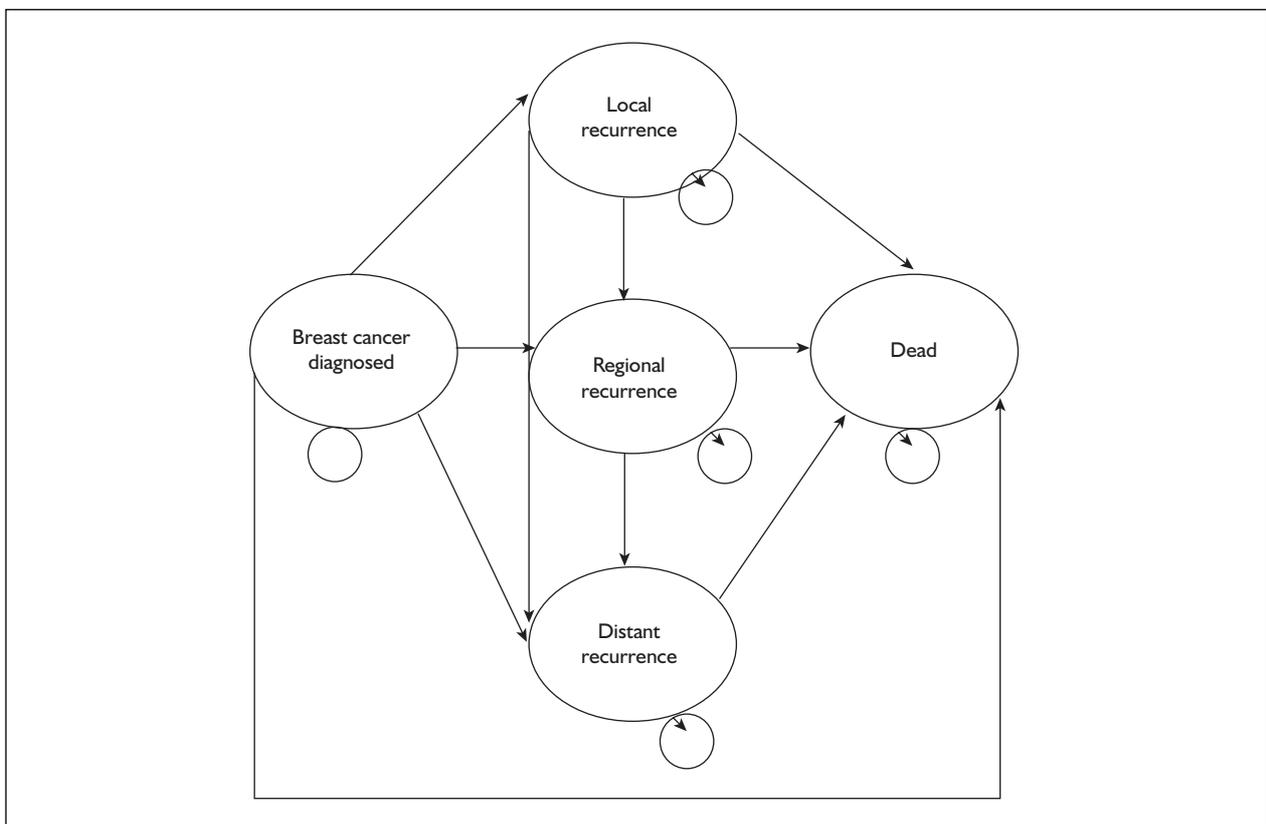


FIGURE 6 Markov state diagram

TABLE 31 Annual transition probabilities differing by prognostic group

Transition/PG	DCIS	Good	Moderate	Poor
BCD to LR	0.0038	0.0096	0.0109	0.0280
BCD to RR	0.0054	0.0068	0.0155	0.0257
BCD to DR	0	0.0049	0.0155	0.0764
BCD to dead	0	0.0039	0.0200	0.0567

BCD, breast cancer diagnosed; LR, local recurrence; RR, regional recurrence; DR, distant recurrence; DCIS, ductal carcinoma *in situ*.

transitions. The model begins at the point where breast cancer has been diagnosed and the NPI and PG have been assigned. There are five states in the model: breast cancer diagnosed, local recurrence, regional recurrence, distant recurrence and dead.

The three types of recurrence refer to the following: local recurrence is recurrence in the ipsilateral breast or mastectomy flaps, regional recurrence is recurrence in the regional lymph nodes (i.e. internal mammary, axillary and/or intraclavicular nodes), and distant recurrence is spread of disease beyond the above sites.

The first state in the model refers to breast cancer diagnosis and the primary treatment received following the breast cancer diagnosis. It is not possible to return to the breast cancer diagnosed state after movements to any of the recurrence states. Once in any of the recurrence states, it is possible to move to more severe recurrence states, but it is not possible to move backwards to less severe recurrence states. The transitions allow patients to be in remission from breast cancer, equivalent to remaining in any of the recurrence states. Although patients may progress through the different states of recurrence, the Markovian assumption means that transitions are independent of what happened in any previous cycle. For example, the probability of transiting to the dead state is independent of the number of regional recurrences that have occurred.

Probabilities

The probabilities of moving from the breast cancer diagnosed state to the different types of recurrences differ by PG and are derived from a database from the Professorial Unit of Surgery at the City Hospital, Nottingham. The database contains follow-up data on 1264 women by PG for

8 years and has information on both recurrence and survival. The database was used to estimate annual transition probabilities. Analysis of breast cancer survival by PG was conducted using a life-table analysis up to 8 years. This provided estimates of the number of women alive at the end of each interval. The estimates were extrapolated from 8 to 10 years using an exponential function (i.e. assuming the hazard function is constant with time for the additional 2 years). The estimates were then compared with other published evidence on survival by NPI group for other cohorts of women⁷⁷ to assess validity.

As the average mortality in the cohort included the additional risks due to recurrence, the baseline cancer mortality without recurrence was adjusted to maintain consistency with the cohort survival when relative risks from the literature for mortality associated with recurrence were applied.⁷⁸ The annual probabilities of dying were then calculated for each PG before recurrence. For the DCIS PG, no survival benefit was assumed and therefore the transition probability of death following breast cancer was equivalent to the rate of other-cause mortality.

Table 31 reports the transition probabilities to dead from breast cancer diagnosis for each PG. A transition probability of zero indicates that it is not an allowable transition.

The database was not able to provide information on probabilities of transitions between different types of recurrence by PG. In the absence of any other information, transitions linking recurrence and survival are assumed to be the same for each PG and are obtained from the literature.^{78,79} Data on probabilities of death by other causes are derived from life tables of deaths of females in England and Wales in 1995 from the Office of National Statistics. *Table 32* reports the transition probabilities assumed to be the same by PG.

TABLE 32 Annual transition probabilities for recurrences and death

Transition	Transition probabilities
LR to RR	0.0400
LR to DR	0.2258
LR to death	0.2152
RR to DR	0.2258
RR to death	0.2438
DR to death	0.7450
Other-cause death aged 50–59 years	0.0091
Other-cause death aged 60–69 years	0.0266
Other-cause death aged 70–79 years	0.0423
Other-cause death aged 80–84 years	0.0719
Other-cause death aged ≥ 85 years	0.1166

Costs and utilities entered into the model

The costs of treating breast cancer following a breast cancer diagnosis were estimated from the City Hospital, Nottingham. The Nottingham database does not record patient-specific resource use and only has information on the type of primary treatment for breast cancer, but no details of the treatment. Consequently, standard treatment protocols for the City Hospital were used to estimate treatment costs. Costs were estimated for primary treatment, recurrences and follow-up. Annual costs were estimated and attached to the appropriate Markov states and transitions. *Table 33* reports these costs. Initial state costs accrue at the beginning of the model, incremental state costs accrue each time a state is entered, and transition costs accrue each time the movement between states occurs. The source of the unit costs of surgery and outpatient visits is the Nottingham City Hospital finance department. The source of the unit costs of adjuvant hormone treatment and chemotherapy treatment is the British National Formulary.

Utilities (required to estimate QALYs) were derived from the literature. The utility for the breast cancer diagnosed state was 0.94.⁸³ The utilities for recurrences were sourced from Hillner and Smith:⁷⁹ 0.7 for local recurrence, 0.5 for regional recurrence and 0.3 for distant recurrence. *Table 34* summarises the methods used by the studies to estimate utilities.

Costs are discounted at 3.5% in the model⁵⁵ and life-years are discounted at 1.5%.⁶⁷

Solving the model for PG-specific costs, life-years and QALYs

The Markov model begins with all patients in the breast cancer diagnosed state at the age of 50 years. A cycle length of 1 year was chosen since follow-up after breast cancer diagnosis and treatment is annual. For transition events, half-cycle corrections were applied. The model was run for ten cycles to estimate 10-year costs (10 years was chosen as a period over which transition probabilities were most valid). The Markov model was built using DATA Professional (version 7) software and replicated in Excel. The cost-effectiveness model was evaluated using a cohort analysis and second order Monte Carlo simulation to address uncertainty (described further below).

Table 35 presents the costs, life-years and QALYs for each PG. Costs are presented in 2001/02 prices. The figures show the expected pattern, namely, that costs increase as severity of prognosis increases with life-years, and QALY showing the opposite pattern.

Probabilistic sensitivity analysis

Probabilistic sensitivity analysis⁸⁴ was used to estimate parameter uncertainty in the model and thereby quantify uncertainty surrounding the cost per life-year and cost per QALY figures. The analysis imposed distributions on parameters and performed 5000 simulations (generating 5000 average costs, life-years and QALYs for both screening strategies). The differences between prompted and non-prompted in costs, life-years and QALYs were calculated and were then plotted

TABLE 33 Costs entered in the model

State	Type of cost	Annual cost (2001/02)
Primary treatment^a		
Primary treatment of DCIS	Initial state cost	£3331
Primary treatment of good prognosis	Initial state cost	£3355
Primary treatment of moderate prognosis	Initial state cost	£4117
Primary treatment of poor prognosis	Initial state cost	£4167
Follow-up^b		
Follow-up after primary	Incremental state cost	£81
Treatment for recurrence^c		
Treatment for LR	Transition cost	£2855
Treatment for RR	Transition cost	£3797
Treatment for DR	Transition cost	£5989
Follow-up after recurrence^d		
Follow up after LR and RR	Incremental state cost	£186
Follow up after DR	Incremental state cost	£4948
Palliative care		
Palliative care	Transition cost	£3158

^a Primary treatment costs are based on proportions having different treatment combinations of surgery and adjuvant treatments. Surgery costs include the costs of operation (anaesthesia time, theatre time, overheads and consumables) and ward costs, and are as follows: £3033 for mastectomy, £2930 for lumpectomy and £2994 for subcutaneous mastectomy. Adjuvant hormone treatment is tamoxifen 20 mg per day for 5 years and four outpatient visits per year (£264, discounted). Adjuvant chemotherapy treatment is based on CMF (cyclophosphamide, methotrexate and 5-fluorouracil) repeated at 28-day intervals for six cycles (£1117) with six outpatient visits per year. Adjuvant radiotherapy consisted of radiotherapy to the breast (50 Gy, 25#, for breast conservation and simple mastectomy, cost of £1585 and 45 Gy, 15# for subcutaneous mastectomy cost of £868) with four outpatient visits per year.

^b Annual outpatient visit £81.

^c The costs of treating recurrence are based on the proportions having different investigative procedures and treatments.⁸⁰⁻⁸² Investigative procedures for local and regional recurrence are core biopsy (£111) or open biopsy (£1039). Investigative procedures for distant recurrence are bone scan (£99), liver ultrasound (£25), computed tomographic scan (£87), chest X-ray (£19), biochemistry tests (£8), skeletal survey (£32), blood count (£6) and magnetic resonance imaging (£52). Costs of treatment for local and regional recurrence are based on surgery and adjuvant treatment costs above. Costs of distant recurrence are based on proportions having first line chemotherapy (£2399) and second line chemotherapy (£5162).⁸²

^d Follow-up after local and regional recurrence is based on two outpatient visits per year and one mammogram (total cost of £194); follow-up of distant recurrence involves second line chemotherapy (£5162).⁸²

TABLE 34 Summary of methods used by studies to estimate utilities

Health state	Source	Mean value	Method of estimation
BCD	Jansen <i>et al.</i> , 1998 ⁸³	0.94	Estimated using standard gamble and time trade-off methods Derived from survey of 61 patients Measured on a scale 0 (dead) to 1 (good health)
LR	Hillner and Smith, 1991 ⁷⁹	0.7	Estimated using visual analogue methods Derived from a survey of oncologists and nurses. Sample size not stated Measured on a scale 0 (dead) to 1 (well)
RR	Hillner and Smith, 1991 ⁷⁹	0.5	Estimated using visual analogue methods Derived from a survey of oncologists and nurses. Sample size not stated Measured on a scale 0 (dead) to 1 (well)
DR	Hillner and Smith, 1991 ⁷⁹	0.3	Estimated using visual analogue methods Derived from a survey of oncologists and nurses. Sample size not stated Measured on a scale 0 (dead) to 1 (well)

TABLE 35 Costs, life-years and QALYs at 10 years by prognostic group

	DCIS	Good	Moderate	Poor
Costs (undiscounted)	£4,981	£6,298	£9,055	£13,707
Costs (discounted)	£4,689	£5,785	£8,236	£12,390
Life-years (undiscounted)	9.32	8.82	7.68	5.15
Life-years (discounted)	8.62	8.16	7.14	4.85
QALYs (undiscounted)	8.69	8.13	6.95	4.33
QALYs (discounted)	7.26	6.83	5.89	3.80

TABLE 36 Distributions imposed on probabilities

Parameter	Expected value	Distribution
death from DR	0.75	Beta
death from LR	0.22	Beta
death after RR	0.24	Beta
BCD to death	–	Fixed
BCD to death	0.00	Beta
BCD to death	0.02	Beta
BCD to death	0.06	Beta
of DR after LR	0.23	Dirichlet
of move to DR after RR	0.23	Dirichlet
of DR in DCIS group	–	Dirichlet
of DR in good group	0.00	Dirichlet
of DR in moderate group	0.02	Dirichlet
of DR in poor group	0.08	Dirichlet
of LR in DCIS group	0.00	Dirichlet
of LR in good group	0.01	Dirichlet
of LR in moderate group	0.01	Dirichlet
of LR in poor group	0.03	Dirichlet
of RR after LR	0.04	Dirichlet
of RR in DCIS group	0.01	Dirichlet
of RR in good group	0.01	Dirichlet
of RR in moderate group	0.02	Dirichlet
of RR in poor group	0.03	Dirichlet

on a cost-effectiveness plane (where the y axis is the cost difference and the x axis is the life-year difference).

Tables 36–38 summarise the distributions imposed on parameters in the model (probabilities, costs and outcome-related parameters). In many cases the distribution imposed is the Dirichlet distribution, which is used because there are multiple branches.⁸⁵ As each reader's sensitivities and specificities are likely to be negatively correlated,³⁷ correlation between sensitivity and specificity under both strategies is accounted for in the simulation, based on the observed correlations in study 2 data.

Cost-effectiveness acceptability curve

In addition to uncertainty arising from parameter uncertainty, a further source of uncertainty in cost-effectiveness figures concerns the maximum (or ceiling) cost-effectiveness ratio that a decision-maker is willing to pay. This uncertainty was represented by plotting cost-effectiveness acceptability curves over a range of ceiling cost-effectiveness ratios, to indicate the probability that the computer prompting strategy is cost-effective at a given ceiling of willingness to pay per unit of health benefit, i.e. life-years or QALYs.

TABLE 37 Distributions imposed on cost parameters

Parameter	Expected value	Distribution
Cost of one cycle in DR	5989.00	Gamma
Cost of 1 year on DR follow-up	4948.00	Gamma
Cost of one cycle in LR	2855.00	Gamma
Cost of recurrence follow-up	186.00	Gamma
Discount rate for costs	3.50%	Fixed
Cost of palliation	3158.00	Gamma
Cost of primary treatment DCIS	3331.00	Gamma
Cost of primary follow-up in BCD	81.00	Gamma
Cost of primary in good	3355.00	Gamma
Cost of primary in moderate	4117.00	Gamma
Cost of primary in poor	4167.00	Gamma
Cost of one cycle in RR	3797.00	Gamma

TABLE 38 Distributions imposed on outcome parameters

Parameter	Expected value	Distribution
Discount rate for outcomes	1.50%	Fixed
Utility of breast cancer diagnosis	0.94	Beta
Utility of DR	0.30	Beta
Utility of LR	0.70	Beta
Utility of RR	0.50	Beta
Age-specific all-cause mortality	0.01	Beta
Age-specific all-cause mortality	0.03	Beta
Age-specific all-cause mortality	0.04	Beta
Age-specific all-cause mortality	0.07	Beta
Age-specific all-cause mortality	0.12	Beta
Prevalence	0.65%	Uniform
Sensitivity of unprompted	0.78	Beta
Specificity of unprompted	0.86	Beta
Sensitivity of prompted	0.81	Beta
Specificity of prompted	0.87	Beta



Health Technology Assessment Programme

Prioritisation Strategy Group

Members

Chair, Professor Tom Walley, Director, NHS HTA Programme, Department of Pharmacology & Therapeutics, University of Liverpool	Professor Bruce Campbell, Consultant Vascular & General Surgeon, Royal Devon & Exeter Hospital	Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Radcliffe Hospital, Oxford
	Professor Shah Ebrahim, Professor in Epidemiology of Ageing, University of Bristol	Dr Ron Zimmern, Director, Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

HTA Commissioning Board

Members

Programme Director, Professor Tom Walley, Director, NHS HTA Programme, Department of Pharmacology & Therapeutics, University of Liverpool	Professor John Brazier, Director of Health Economics, Sheffield Health Economics Group, School of Health & Related Research, University of Sheffield	Professor Peter Jones, Head of Department, University Department of Psychiatry, University of Cambridge	Professor Mark Sculpher, Professor of Health Economics, Centre for Health Economics, Institute for Research in the Social Services, University of York
Chair, Professor Shah Ebrahim, Professor in Epidemiology of Ageing, Department of Social Medicine, University of Bristol	Dr Andrew Briggs, Public Health Career Scientist, Health Economics Research Centre, University of Oxford	Professor Sallie Lamb, Research Professor in Physiotherapy/Co- Director, Interdisciplinary Research Centre in Health, Coventry University	Professor Martin Severs, Professor in Elderly Health Care, Portsmouth Institute of Medicine
Deputy Chair, Professor Jenny Hewison, Professor of Health Care Psychology, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds School of Medicine	Professor Nicky Cullum, Director of Centre for Evidence Based Nursing, Department of Health Sciences, University of York	Professor Julian Little, Professor of Epidemiology, Department of Medicine and Therapeutics, University of Aberdeen	Dr Jonathan Shapiro, Senior Fellow, Health Services Management Centre, Birmingham
Dr Jeffrey Aronson Reader in Clinical Pharmacology, Department of Clinical Pharmacology, Radcliffe Infirmary, Oxford	Dr Andrew Farmer, Senior Lecturer in General Practice, Department of Primary Health Care, University of Oxford	Professor Stuart Logan, Director of Health & Social Care Research, The Peninsula Medical School, Universities of Exeter & Plymouth	Ms Kate Thomas, Deputy Director, Medical Care Research Unit, University of Sheffield
Professor Ann Bowling, Professor of Health Services Research, Primary Care and Population Studies, University College London	Professor Fiona J Gilbert, Professor of Radiology, Department of Radiology, University of Aberdeen	Professor Tim Peters, Professor of Primary Care Health Services Research, Division of Primary Health Care, University of Bristol	Professor Simon G Thompson, Director, MRC Biostatistics Unit, Institute of Public Health, Cambridge
Professor Andrew Bradbury, Professor of Vascular Surgery, Department of Vascular Surgery, Birmingham Heartlands Hospital	Professor Adrian Grant, Director, Health Services Research Unit, University of Aberdeen	Professor Ian Roberts, Professor of Epidemiology & Public Health, Intervention Research Unit, London School of Hygiene and Tropical Medicine	Ms Sue Ziebland, Senior Research Fellow, Cancer Research UK, University of Oxford
	Professor F D Richard Hobbs, Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham	Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh	

Diagnostic Technologies & Screening Panel

Members

<p>Chair, Dr Ron Zimmern, Director of the Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge</p>	<p>Professor Adrian K Dixon, Professor of Radiology, Addenbrooke's Hospital, Cambridge</p>	<p>Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London</p>	<p>Dr Margaret Somerville, Director of Public Health, Teignbridge Primary Care Trust</p>
<p>Ms Norma Armston, Freelance Consumer Advocate, Bolton</p>	<p>Dr David Elliman, Consultant in Community Child Health, London</p>	<p>Dr Edmund Jessop, Medical Adviser, National Specialist Commissioning Advisory Group (NSCAG), Department of Health, London</p>	<p>Professor Lindsay Wilson Turnbull, Scientific Director, Centre for MR Investigations & YCR Professor of Radiology, University of Hull</p>
<p>Professor Max Bachmann, Professor Health Care Interfaces, Department of Health Policy and Practice, University of East Anglia</p>	<p>Professor Glyn Elwyn, Primary Medical Care Research Group, Swansea Clinical School, University of Wales Swansea</p>	<p>Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford</p>	<p>Professor Martin J Whittle, Head of Division of Reproductive & Child Health, University of Birmingham</p>
<p>Professor Rudy Bilous, Professor of Clinical Medicine & Consultant Physician, The Academic Centre, South Tees Hospitals NHS Trust</p>	<p>Dr John Fielding, Consultant Radiologist, Radiology Department, Royal Shrewsbury Hospital</p>	<p>Dr Susanne M Ludgate, Medical Director, Medical Devices Agency, London</p>	<p>Dr Dennis Wright, Consultant Biochemist & Clinical Director, Pathology & The Kennedy Galton Centre, Northwick Park & St Mark's Hospitals, Harrow</p>
<p>Dr Paul Cockcroft, Consultant Medical Microbiologist/Laboratory Director, Public Health Laboratory, St Mary's Hospital, Portsmouth</p>	<p>Dr Karen N Foster, Clinical Lecturer, Dept of General Practice & Primary Care, University of Aberdeen</p>	<p>Dr William Rosenberg, Senior Lecturer and Consultant in Medicine, University of Southampton</p>	
	<p>Professor Antony J Franks, Deputy Medical Director, The Leeds Teaching Hospitals NHS Trust</p>	<p>Dr Susan Schonfield, CPHM Specialised Services Commissioning, Croydon Primary Care Trust</p>	

Pharmaceuticals Panel

Members

<p>Chair, Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Oxford Radcliffe Hospital</p>	<p>Dr Christopher Cates, GP and Cochrane Editor, Bushey Health Centre</p>	<p>Mrs Sharon Hart, Managing Editor, <i>Drug & Therapeutics Bulletin</i>, London</p>	<p>Professor Jan Scott, Professor of Psychological Treatments, Institute of Psychiatry, University of London</p>
<p>Professor Tony Avery, Professor of Primary Health Care, University of Nottingham</p>	<p>Professor Imti Choonara, Professor in Child Health, University of Nottingham, Derbyshire Children's Hospital</p>	<p>Dr Christine Hine, Consultant in Public Health Medicine, Bristol South & West Primary Care Trust</p>	<p>Mrs Katrina Simister, New Products Manager, National Prescribing Centre, Liverpool</p>
<p>Professor Stirling Bryan, Professor of Health Economics, Health Services Management Centre, University of Birmingham</p>	<p>Mr Charles Dobson, Special Projects Adviser, Department of Health</p>	<p>Professor Stan Kaye, Professor of Medical Oncology, Consultant in Medical Oncology/Drug Development, The Royal Marsden Hospital</p>	<p>Dr Richard Tiner, Medical Director, Association of the British Pharmaceutical Industry</p>
<p>Mr Peter Cardy, Chief Executive, Macmillan Cancer Relief, London</p>	<p>Dr Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham</p>	<p>Ms Barbara Meredith, Project Manager Clinical Guidelines, Patient Involvement Unit, NICE</p>	<p>Dr Helen Williams, Consultant Microbiologist, Norfolk & Norwich University Hospital NHS Trust</p>
	<p>Dr Karen A Fitzgerald, Pharmaceutical Adviser, Bro Taf Health Authority, Cardiff</p>	<p>Dr Frances Rotblat, CPMP Delegate, Medicines Control Agency, London</p>	

Therapeutic Procedures Panel

Members

Chair,

Professor Bruce Campbell,
Consultant Vascular and
General Surgeon, Royal Devon
& Exeter Hospital

Dr Mahmood Adil, Head of
Clinical Support & Health
Protection, Directorate of
Health and Social Care (North),
Department of Health,
Manchester

Dr Aileen Clarke,
Reader in Health Services
Research, Public Health &
Policy Research Unit,
Barts & the London School of
Medicine & Dentistry,
Institute of Community Health
Sciences, Queen Mary,
University of London

Mr Matthew William Cooke,
Senior Clinical Lecturer and
Honorary Consultant,
Emergency Department,
University of Warwick, Coventry
& Warwickshire NHS Trust,
Division of Health in the
Community, Centre for Primary
Health Care Studies, Coventry

Dr Carl E Counsell, Senior
Lecturer in Neurology,
University of Aberdeen

Dr Keith Dodd, Consultant
Paediatrician, Derbyshire
Children's Hospital

Professor Gene Feder, Professor
of Primary Care R&D, Barts &
the London, Queen Mary's
School of Medicine and
Dentistry, University of London

Professor Paul Gregg,
Professor of Orthopaedic
Surgical Science, Department of
Orthopaedic Surgery,
South Tees Hospital NHS Trust

Ms Bec Hanley, Freelance
Consumer Advocate,
Hurstpierpoint

Ms Maryann L. Hardy,
Lecturer,
Division of Radiography,
University of Bradford

Professor Alan Horwich,
Director of Clinical R&D, The
Institute of Cancer Research,
London

Dr Phillip Leech, Principal
Medical Officer for Primary
Care, Department of Health,
London

Dr Simon de Lusignan,
Senior Lecturer, Primary Care
Informatics, Department of
Community Health Sciences,
St George's Hospital Medical
School, London

Dr Mike McGovern, Senior
Medical Officer, Heart Team,
Department of Health, London

Professor James Neilson,
Professor of Obstetrics and
Gynaecology, Dept of Obstetrics
and Gynaecology,
University of Liverpool,
Liverpool Women's Hospital

Dr John C Pounsford,
Consultant Physician, North
Bristol NHS Trust

Dr Vimal Sharma,
Consultant Psychiatrist & Hon
Snr Lecturer,
Mental Health Resource Centre,
Victoria Central Hospital,
Wirrall

Dr L David Smith, Consultant
Cardiologist, Royal Devon &
Exeter Hospital

Professor Norman Waugh,
Professor of Public Health,
University of Aberdeen

Expert Advisory Network

Members

Professor Douglas Altman,
Director of CSM & Cancer
Research UK Med Stat Gp,
Centre for Statistics in
Medicine, University of Oxford,
Institute of Health Sciences,
Headington, Oxford

Professor John Bond,
Director, Centre for Health
Services Research,
University of Newcastle upon
Tyne, School of Population &
Health Sciences,
Newcastle upon Tyne

Mr Shaun Brogan,
Chief Executive, Ridgeway
Primary Care Group, Aylesbury

Mrs Stella Burnside OBE,
Chief Executive,
Office of the Chief Executive.
Trust Headquarters,
Altnagelvin Hospitals Health &
Social Services Trust,
Altnagelvin Area Hospital,
Londonderry

Ms Tracy Bury,
Project Manager, World
Confederation for Physical
Therapy, London

Mr John A Cairns,
Professor of Health Economics,
Health Economics Research
Unit, University of Aberdeen

Professor Iain T Cameron,
Professor of Obstetrics and
Gynaecology and Head of the
School of Medicine,
University of Southampton

Dr Christine Clark,
Medical Writer & Consultant
Pharmacist, Rossendale

Professor Collette Mary Clifford,
Professor of Nursing & Head of
Research, School of Health
Sciences, University of
Birmingham, Edgbaston,
Birmingham

Professor Barry Cookson,
Director,
Laboratory of Healthcare
Associated Infection,
Health Protection Agency,
London

Professor Howard Stephen Cuckle,
Professor of Reproductive
Epidemiology, Department of
Paediatrics, Obstetrics &
Gynaecology, University of
Leeds

Professor Nicky Cullum,
Director of Centre for Evidence
Based Nursing, University of York

Dr Katherine Darton,
Information Unit, MIND – The
Mental Health Charity, London

Professor Carol Dezateux,
Professor of Paediatric
Epidemiology, London

Mr John Dunning,
Consultant Cardiothoracic
Surgeon, Cardiothoracic
Surgical Unit, Papworth
Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw,
Consultant Vascular Surgeon,
Gloucestershire Royal Hospital,
Gloucester

Professor Martin Eccles,
Professor of Clinical
Effectiveness, Centre for Health
Services Research, University of
Newcastle upon Tyne

Professor Pam Enderby,
Professor of Community
Rehabilitation, Institute of
General Practice and Primary
Care, University of Sheffield

Mr Leonard R Fenwick,
Chief Executive, Newcastle
upon Tyne Hospitals NHS Trust

Professor David Field,
Professor of Neonatal Medicine,
Child Health, The Leicester
Royal Infirmary NHS Trust

Mrs Gillian Fletcher,
Antenatal Teacher & Tutor and
President, National Childbirth
Trust, Henfield

Professor Jayne Franklyn,
Professor of Medicine,
Department of Medicine,
University of Birmingham,
Queen Elizabeth Hospital,
Edgbaston, Birmingham

Ms Grace Gibbs,
Deputy Chief Executive,
Director for Nursing, Midwifery
& Clinical Support Servs,
West Middlesex University
Hospital, Isleworth

Dr Neville Goodman,
Consultant Anaesthetist,
Southmead Hospital, Bristol

Professor Alastair Gray,
Professor of Health Economics,
Department of Public Health,
University of Oxford

Professor Robert E Hawkins,
CRC Professor and Director of
Medical Oncology, Christie CRC
Research Centre, Christie
Hospital NHS Trust, Manchester

Professor F D Richard Hobbs,
Professor of Primary Care &
General Practice, Department of
Primary Care & General
Practice, University of
Birmingham

Professor Allen Hutchinson,
Director of Public Health &
Deputy Dean of SCHARR,
Department of Public Health,
University of Sheffield

Dr Duncan Keeley,
General Practitioner (Dr Burch
& Ptnrs), The Health Centre,
Thame

Dr Donna Lamping,
Research Degrees Programme
Director & Reader in Psychology,
Health Services Research Unit,
London School of Hygiene and
Tropical Medicine, London

Mr George Levvy,
Chief Executive, Motor
Neurone Disease Association,
Northampton

Professor James Lindesay,
Professor of Psychiatry for the
Elderly, University of Leicester,
Leicester General Hospital

Professor Rajan Madhok,
Medical Director & Director of
Public Health, Directorate of
Clinical Strategy & Public
Health, North & East Yorkshire
& Northern Lincolnshire Health
Authority, York

Professor David Mant,
Professor of General Practice,
Department of Primary Care,
University of Oxford

Professor Alexander Markham,
Director, Molecular Medicine
Unit, St James's University
Hospital, Leeds

Dr Chris McCall,
General Practitioner,
The Hadleigh Practice,
Castle Mullen

Professor Alistair McGuire,
Professor of Health Economics,
London School of Economics

Dr Peter Moore,
Freelance Science Writer,
Ashtead

Dr Andrew Mortimore,
Consultant in Public Health
Medicine, Southampton City
Primary Care Trust

Dr Sue Moss,
Associate Director, Cancer
Screening Evaluation Unit,
Institute of Cancer Research,
Sutton

Professor Jon Nicholl,
Director of Medical Care
Research Unit, School of Health
and Related Research,
University of Sheffield

Mrs Julietta Patnick,
National Co-ordinator, NHS
Cancer Screening Programmes,
Sheffield

Professor Robert Peveler,
Professor of Liaison Psychiatry,
University Mental Health
Group, Royal South Hants
Hospital, Southampton

Professor Chris Price,
Visiting Chair – Oxford,
Clinical Research, Bayer
Diagnostics Europe,
Cirencester

Ms Marianne Rigge,
Director, College of Health,
London

Dr Eamonn Sheridan,
Consultant in Clinical Genetics,
Genetics Department,
St James's University Hospital,
Leeds

Dr Ken Stein,
Senior Clinical Lecturer in
Public Health, Director,
Peninsula Technology
Assessment Group,
University of Exeter

Professor Sarah Stewart-Brown,
Director HSRU/Honorary
Consultant in PH Medicine,
Department of Public Health,
University of Oxford

Professor Ala Szczepura,
Professor of Health Service
Research, Centre for Health
Services Studies, University of
Warwick

Dr Ross Taylor,
Senior Lecturer,
Department of General Practice
and Primary Care,
University of Aberdeen

Mrs Joan Webster,
Consumer member, HTA –
Expert Advisory Network

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (<http://www.nchta.org>) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.