**TITLE:**
The use of Computer Aided Detection tools in screening mammography: A multidisciplinary investigation

**AUTHORS:**

Eugenio Alberdi, PhD[1]
Andrey A. Povyakalo, PhD[1]
Lorenzo Strigini, M.Eng[1]
Peter Ayton, PhD[2]
Mark Hartswood, PhD[3]
Rob Procter, PhD[3]
Roger Slack, PhD[3]

**ADDRESSES:**

[1] Centre for Software Reliability, City University, Northampton Square, London, EC1V 0HB, UK
[2] Psychology Department, City University, Northampton Square, London, EC1V 0HB, UK
[3] Institute for Communicating and Collaborative Systems School of Informatics, The University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, UK

**SHORTENED VERSION OF THE TITLE:**

Multidisciplinary study of CAD use in mammography

**FUNDING:**
The work described in this paper has been partly funded by the UK Engineering and Physical Sciences Research Council (EPSRC) through DIRC, the Dependability Interdisciplinary Research Collaboration, a project investigating the dependability of computer based systems.

**KEYWORDS:**
Breast cancer screening, digital imaging, mammography

**ABSTRACT**

We summarise a set of analyses and studies we have conducted to assess the effects of the use of a Computer Aided Detection (CAD) tool in breast screening. We have used an interdisciplinary approach which combines: a) statistical analyses inspired by reliability modelling in engineering, b) experimental studies of mammography experts' decisions using the tool, interpreted in the light of human factors psychology; and c) ethnographic observations of the use of the tool both in trial conditions and in everyday screening practice. Our investigations have shown patterns of human behaviour and effects of computer-based advice that would not have been revealed by a standard clinical trial approach. For example, we found that the negligible measured effect of CAD could be explained by a range of effects on experts' decisions, beneficial in some cases and detrimental in others. There is some evidence of the latter effects being due to the experts using the computer tool differently from the developers' intentions. We integrate insights from the different pieces of evidence and highlight their implications for the design, evaluation and deployment of this sort of computer tool.

## 1. Introduction

In this paper we summarise the results of a variety of studies and analyses we have conducted to investigate the use of a Computer Aided Detection (CAD) tool in mammography. This work was done as part of DIRC (http://www.dirc.org.uk), a collaborative, multidisciplinary project dealing with the dependability of computer based systems: systems encompassing computer software and hardware together with their human users and the social context of computer use. The use of computer aids in breast screening is a good example of this sort of 'human-computer system'.

As is common in the radiological literature, we will use the abbreviation 'CAD' to mean the computer tool, whenever the context does not create ambiguities with its literal meaning 'detection activity aided by a computer'.

CAD is used to alert (prompt) a human expert (typically a radiologist) to areas of a mammogram (an X-ray image of the breast) where computerised image analysis suggests that abnormalities may be found. Typically the CAD tool processes a digitised version of a mammogram and marks it with 'prompts' to highlight mammographic features that the reader should examine. The design goal for CAD is to aid human experts (hereafter referred to as 'readers') to notice features in a mammogram that might indicate cancer but that they may otherwise miss.

Our interest in the human-CAD system originated from a clinical trial funded by the UK Health Technology Assessment (HTA) programme [1]. The goal was to assess the impact of a particular CAD tool, R2 ImageChecker M100 (from R2 Technology, Inc., Los Altos, California) [2], in breast screening. The trial was designed to test whether CAD increases readers' sensitivity (the proportion of cancers recalled out of all cancers) without adversely affecting their specificity (the proportion

of normal cases *not* recalled out of all normal cases). In Section 2 of this paper we outline aspects of the HTA trial (reported in detail in [1]) that are relevant to our research.

Our approach is multidisciplinary and combines insights from various disciplines: reliability engineering, computing, psychology, human factors and sociology. We were granted access to the HTA trial data and conducted supplementary analyses to investigate in more detail the effects of CAD on readers' decisions. Our analyses were influenced by reliability modelling in engineering [3-4] and focused on how, *inter alia*, the effects of CAD vary between cases or depend on whether the tool prompts mammograms correctly or not. Section 3 summarises some results of these analyses, which have been partly reported in [5]. We also conducted two follow-up studies, outlined in Section 4, to investigate how readers react to incorrect CAD output [6]. The HTA trial and the follow-up studies were complemented by ethnographic studies of CAD use (see Section 5), which have been partly reported in [7].

The discussion in Section 6 attempts to integrate the different findings and highlights their potential implications for the design, evaluation and use of CAD. This demands that we consider how we may extrapolate from data obtained in trial conditions to the real world use of CAD in breast screening; we reflect on the potential benefits and as yet unresolved questions of using a multidisciplinary approach to address this important issue.

## 2. The HTA Trial

The goal of the HTA trial was to assess the impact of a CAD tool (R2 ImageChecker M1000 [2]) in breast screening. This CAD tool had been shown to have a high

sensitivity, that is, it prompts a high proportion of breast cancers. Castellino and colleagues report an overall sensitivity of up to 90% [8]. However, high sensitivity comes at the expense of low specificity, that is, it generates many false prompts, with an average of 2.06 prompts per case in one study.

The trial was run with 50 readers, experienced in breast screening, and used 180 cases (a mixture of 60 cancers and 120 normal cases) distributed in 3 sets of 60. All participating readers saw all the cases in two different experimental conditions: a) 'unprompted condition': without CAD; b) 'prompted condition': with CAD. The order of conditions was randomised across the participants. In both conditions, the participants saw two versions of each case: 1) the mammograms positioned on a standard viewing roller; and 2) a digitised version of the mammograms printed on paper. In the prompted condition, the printouts contained the prompts generated by CAD. Participants were also asked to make their decisions as whether a case should be recalled for further tests as if they were viewing the mammograms as single readers in the screening programme. More details of the procedures can be found in [1].

Analysis of the results showed no statistically significant impact of CAD (no improvement and no reduction) on the readers' sensitivity and specificity [1].

## 3. Supplementary analyses of the data from the HTA trial

The analyses we describe here were inspired by studies of the dependability of software systems with diverse redundancy [3]. These studies have shown that variation and co-variation of difficulty of input cases for different system components affect the dependability of the overall system substantially. They also indicate that

focusing on the average probabilities of the failures of the components, or assuming statistical independence among their failures, can be misleading. We explored the possibility that conclusions from those studies could also apply to the system formed by CAD and the human reader. We focused our analyses on the role of the correctness of CAD output and evaluating its interaction with *case difficulty*, which we define as a probability of a reader, randomly selected from the given set, reaching a wrong decision about a given case.

Throughout this paper, we will be talking about 'errors' of human (reader) or machine (CAD). We clarify here how we define these errors and present the terminology we will be using:

- 'error by a human' (or 'incorrect decision') means that she/he recalls a normal case or fails to recall a cancer. 'Correct decision' means that she/he fails to recall a normal case or recalls a cancer.
- we talk of 'error by CAD', or 'incorrect output':
  - for a cancer, if CAD 'misses' the cancer (i.e. it gives a 'false negative'), which it can do in two ways: a) by failing to place any prompt on the mammogram ('unmarked' cancer), or b) by placing prompts in areas other than the actual location of the cancer ('incorrectly marked' cancer: all prompts are 'false'). We say that CAD provides 'correct output' for a cancer (i.e. it processes it correctly; true positive) if it prompts the area on the mammogram where the cancer is located ('correctly marked' cancer), even if CAD may also have prompted other areas of the mammogram.
  - for a normal case, if CAD places any prompt on the mammogram ('incorrectly marked' normal case; false positive). We say that CAD

provides ‘correct output’ for a normal case if it places no prompts (‘unmarked’ normal; a ‘true negative’).

*3.1 Potential of CAD for improving readers’ sensitivity*

It is interesting to estimate how much CAD could potentially improve readers' sensitivity for the test set in the trial.

In the HTA trial, readers in the unprompted condition (without CAD), made 2994 decisions about cancers (50 readers * 60 cancers - 6 missed data points). 741 of these decisions were incorrect. Thus, the estimated probability of human error for cancers, in the unprompted condition, is 741/2994=0.247 with 95% confidence interval (0.232, 0.263). Of these 741 errors, 314 occurred on cases correctly prompted by CAD.

If we assume (optimistically) that readers in the prompted condition will make the best possible use of the CAD, i.e.:

1. will recognise and recall all the cancers correctly prompted by CAD;
2. will still recall all cancers that they recalled in the unprompted conditions, even with incorrect output from CAD,

then the expected number of errors for cancers in the prompted condition is 741 – 314 = 427 (out of all 2994 decisions), i.e. the probability of failure in the prompted condition is estimated to be 427/2994 = 0.143 with 95% confidence interval (0.130, 0.156). Considering only the mean estimates, the estimated potential for improving readers’ sensitivity equals 314/2994 = 0.105 (10.5%). Using a similar approach in another study, Warren Burhenne and colleagues [9] reported “...CAD prompting could have potentially helped reduce...” the initial rate of FN errors (0.21) by 77%,

i.e., they estimated the potential for improving readers' sensitivity as 0.21 ? 77% = 16.2% (0.162).

The analysis of the HTA trial data [1] shows that the estimated potential 10.5% improvement in sensitivity does not seem to be realistic, at least for this experimental setting, and that the assumptions for the above calculation were not verified. In the prompted condition, readers missed correctly prompted cancers; they also made errors they did not make in the unprompted condition, as shown in Table 1 for different categories of cancers.

*3.2 Correctness of CAD output affects readers' decisions*

In this section, we are looking at situations where readers' decisions about the same case were different in the prompted and unprompted conditions, using an approach similar to [10].

We ignore those outcomes in which the two decisions were either both correct or both wrong and consider only the following outcomes:

- the decision by a reader for a case was incorrect in the unprompted condition and correct in the prompted condition ('prompted decision better');
- the reader's decision was correct in the prompted condition and incorrect in the unprompted condition ('prompted decision worse').

If there were statistical independence between the outcomes 'prompted decision better/worse' and 'CAD output correct/incorrect', then we could conclude that the correctness of CAD output does not affect readers' decisions. We tested for independence using Fisher's exact test [11] for the contingency tables given in Table 2. Independence would mean that the probability of the prompted decision being

better rather than worse would be the same for cases correctly processed by CAD as for cases incorrectly processed by CAD.

One can see in Table 2 that for cases (either cancers or normal) that were incorrectly processed by CAD, readers' decisions in the prompted condition were more often worse (than their decisions in the unprompted condition) rather than better. On the other hand, for correctly processed cases, readers' decisions in the prompted condition were more often better than worse. For example, the odds that the prompted decision for a cancer is worse equal 89/73 ? 1.22 for incorrectly processed cancers and 113/147 ? 0.73 for correctly processed cancers. The odds ratio equals 1.22/0.73 ? 1.58. The test indicates that this ratio is significantly different from 1 with the p-value = 0.0274.

The results indicate that correct CAD output is likely to help in reaching a correct decision and that incorrect CAD output makes it more difficult.

*3.3 Error rates of readers and CAD for different populations of cases in different conditions*

To investigate the effect of correctness of CAD output, we estimated the probabilities of human error in both the prompted and the unprompted conditions, for cases categorised according to case type (normal or cancer) and the output of CAD (correct or incorrect).

These estimates are shown in Table 3. Both for cancers and for normal cases, all the estimated probabilities of error do not differ significantly between the prompted and unprompted conditions, *except* for those normal cases that CAD did not prompt.

For this category of cases, the readers' rate of false recalls in the prompted condition was smaller than in the unprompted condition by 0.06.

As seen in Table 3, even in the unprompted condition, readers were more likely to make incorrect decisions for cases which CAD also processed incorrectly: reader and CAD errors are strongly correlated. In other words, for cases that are more difficult for humans to interpret, CAD is less likely to give useful output.

We tested the correlation between errors made by readers and CAD for significance with the exact Fisher's test [11] applied to the contingency tables for correct and wrong decisions by readers, on cancers and on normal cases in both conditions (see Table 4). The test indicated significant correlation for all 4 contingency tables (p-value < 0.05).

*3.4 Correctness of CAD output and case difficulty*

We define *difficulty* of a case as the fraction of readers who produced an incorrect decision about that case. We denote difficulty in the prompted condition as $d_p$ and difficulty in the unprompted condition as $d$.

We call *non-obvious* a case on which at least one reader made an incorrect decision, either in the prompted or unprompted condition.

ANOVA indicated that the value: $d - d_p$ is significantly different for groups of cases processed by CAD correctly and incorrectly:

- all cancers: F(1,58)=4.1899, p-value= 0.0452;
- non-obvious cancers: F(1,44)= 4.0676, p-value=0.0498;
- normal cases: F(1,118)=9.3384, p-value=0.00278;
- non-obvious normal cases: F(1,114)= 9.0746, p-value=0.00319;

In [5] we applied logistic regression to highlight general patterns in the effect of CAD. It appears that CAD tends to make cancers which are relatively easy (i.e. with $d<0.6$) less difficult (i.e., $d_p<d$) and cases which are relatively difficult (i.e. with $d>0.6$) even more difficult (i.e., $d_p>d$).

The plot in Figure 1 illustrates this effect. The horizontal axis represents the unprompted difficulty $d$. The vertical axis shows the differences $(d_p - d)$. So, a point below the horizontal line indicates a cancer for which CAD appears to reduce the rate of reader errors for cancers. Points marked 'w' and 'c' indicate the observed values of $d$ and $(d_p - d)$ for non-obvious cancers, divided into those with correct CAD output (symbols 'c') and with wrong CAD output (symbols 'w'). The curves show the regression estimate for the mean value of $(d_p - d)$ for cases with different difficulty $d$. The dashed curve corresponds to incorrectly prompted cancers, the dotted-dashed curve to the correctly prompted cancers and the solid curve to all cancers together. For more details of our regression analyses, see [5].

## 4. Follow-up studies

We conducted two studies as a follow up to the HTA trial, investigating in more detail the effects of incorrect computer output on human decision making [6].

The test sets in the HTA trial did not contain enough examples of cancers incorrectly processed by CAD (in particular, 'unmarked' cancers; see definitions in Section 3) to allow us to draw statistically significant estimates of their effects on the readers' decisions. We ran a follow-up experiment (Study 1) with a new test set (60

cases) containing a larger proportion of cancers missed by CAD (20 out of the 30 cancers in the set). Nine of the false negatives in our test set were ‘unmarked cancers’. We kept all other characteristics of the test set as similar as possible to the sets used in the HTA trial as we wanted the readers to perceive this study as a natural extension to the original trial and to behave in a comparable way.

The participants in Study 1 were twenty readers who had participated in the original trial. We used essentially the same procedures as used in the HTA trial, except that readers in Study 1 saw all the cases only once: always with the benefit of CAD (‘prompted condition’).

At that stage, we were not interested in comparing readers’ performance with and without CAD; our goal was to estimate the probability of reader error when the output from CAD was incorrect. However, the results turned out to be highly unexpected: the average reader sensitivity was surprisingly low (52%) and this decrease was particularly strong for the ‘unmarked’ cancers. This led us to suspect that CAD errors may have had a significant negative impact on readers’ decisions. On the other hand, we could not exclude the alternative explanation that the cases in our study had characteristics that made them particularly difficult (perhaps mammographically undetectable) for both human readers and the CAD tool.

As a ‘control’ for Study 1, we ran Study 2, where readers saw the same test set without CAD (‘unprompted condition’). We used nineteen readers from three different UK screening centres , none of whom had participated in Study 1 but who were equivalent to the group used in Study 1 in terms of years of experience and professional background. We used the same procedures as in Study 1 except that readers did not see the CAD output.

Additionally, we conducted a new test with six of the more experienced participants in Study 2 to get a better understanding of the 'difficulty' of the cancers in our test set. These participants were presented again with the 30 mammograms containing cancer and were asked to rank them according to various criteria of case 'difficulty'. The responses of this subset of readers, as well as the performance of all participants in Study 2, strongly indicated that six of the cases in our test set were probably 'occult' cancers, undetectable via mammography and so we eliminated these cases from our analyses (see more details in[6]).

Readers' average sensitivity for the remaining 24 cancers was 61% for those who saw the cases with CAD (Study 1) and 73% for those who saw them without CAD (Study 2). The difference in average sensitivity between the two sets of readers was statistically significant. In contrast, the average specificity in Study 2 was lower than in Study 1 (86% vs. 90%) but the difference was not statistically significant.

Table 5 shows the proportions of incorrect human decisions in Study 1 and Study 2 for the 54 analysed cases, categorised according to case type (normal or cancer) and output of CAD ('unmarked', 'correctly marked', 'incorrectly marked'). ANOVA showed statistically significant differences between Study 1 and Study 2 for: the 'unmarked cancers' ($p<0.001$), the 'incorrectly marked' cancers ($p<0.05$) and the 'unmarked' normal cases ($p<0.05$).

These findings strongly suggest that, at least for some categories of cases, incorrect CAD output had a significant detrimental effect on human decisions in our studies.

## 5. Ethnographic studies of CAD use

The aim of the ethnographic studies was to examine how readers used the CAD tool and the ways in which they came to understand and explain its prompts and general behaviour. Readers' behaviour was observed both in trial conditions [7] and *in vivo* in everyday breast screening practice. Observations took place over a total period of about one month, and were videoed and reviewed so as to examine in fine grained detail the work of reading.

Using a 'think aloud' protocol in which participants vocalised their thought processes to the observer as they read cases, we attempted to explore the sense readers made of prompts in the context of the mammograms on which they occurred. Subsequently, we discussed with readers cases they identified as problematic (especially cases on which they had spent a substantial amount of time) to clarify how they dealt with these 'difficult' cases.

A significant finding was the importance readers attached to ascertaining what a prompt 'meant': how it could be explained in the context of the mammogram. In some cases the accounts were of the order "I don't know why it's prompted that", in others readers saw the features prompted as, for example, composite shadows and gave an account of why they thought the CAD tool had prompted the feature together with what they saw it as actually being (e.g., benign, because it could be 'picked apart').

Although readers were advised to use CAD as an attention cue, and to use their own judgement to decide whether a prompted feature required recall, we observed that they sometimes used prompts to inform their decisions. For example, one reader commented: "This is a case where without the prompt I'd probably let it go … but seeing the prompt I'll probably recall … it doesn't look like a mass but she's got quite

difficult dense breasts … I'd probably recall." In other instances, we observed readers using the absence of a prompt in as evidence for 'no recall'.

The explanation for this may be that the uses to which prompts are put are contingent on the specific problems posed by individual cases. For example, reading dense, feature rich breasts poses demands very different from those of lucent or uncomplicated breasts, and the reader's comment above demonstrates how she marshals the 'evidence' of the prompt in making a decision under these specific circumstances.

Over time, readers acquired a 'biography' for the CAD tool: they came to believe they knew what features it would and would not prompt and they read with this putative biography as a factor in their work. For example, readers would often remark that they had anticipated that the system would prompt for a particular feature within the breast, sometimes then dismissing the prompt as they already had judged the feature to be benign. This does not mean that they would ignore the prompt, or that they would not pay serious attention to it, but that they thought the prompt could be expected given their emerging understanding of what the CAD tool could and did do.

Post trial discussions with readers indicated some of the strengths they attributed to the CAD tool:

- an ability to draw their attention to subtle signs they might have missed otherwise; in some instances, readers reported that a prompt had led them to recall a case that they would have 'let go' had the prompt not been there;
- its perceived consistency; prompts were seen as useful by compensating in some consistent way for individual human weaknesses.

On the other hand, readers noticed the following weaknesses of the tool:

- the many false prompts: the tool was seen to prompt the wrong mammographic features; in many instances, readers explicitly dismissed prompts that they interpreted as artefacts, noise or obvious benign appearances;
- the distracting effect of too many prompts;
- the fact that the tool missed obvious features that readers thought should have been prompted.

## 6. Discussion

### *6.1 Impact of CAD on readers' sensitivity*

CAD in mammography has the potential to improve readers' ability to detect cancers. As we saw in Section 3.1, if readers were always to react appropriately to the prompts placed correctly by CAD, their sensitivity in the HTA trial could have increased by up to ~10%. However, the trial showed no statistically significant effects of CAD on readers' *average* sensitivity. This result is actually consistent with most experimental measurements of the impact of R2 ImageChecker and other similar CAD tools on mammogram reading (e.g. [12-16]). To improve the effectiveness of computer aids, it is desirable to explain why CAD apparently had no effect in these studies.

A simple conjecture would be that readers tend to ignore CAD outputs, possibly because the high number of false prompts creates excessive load. However, our statistical analyses and ethnographic studies do suggest systematic effects of the use of CAD, which are positive or negative depending on aspects of the cases and on CAD output. Readers' reports suggested that the presence of CAD prompts had at

times alerted them to relevant mammographic features that they would have missed otherwise as well as affecting their recall decision for features that they had already noticed. These reports are corroborated by our supplementary analyses of the HTA trial data, which indicate that for a subset of cases CAD did have beneficial effects on readers' decisions (see Table 2 & Fig 1). At the same time, readers often dismissed explicitly many of the prompts as they considered them false. Analysis of the data indicates that some of these were, in fact, correct prompts: readers sometimes had difficulties distinguishing between correct and incorrect prompts.

Our statistical analyses also show that CAD output could have detrimental effects on readers' sensitivity for a subset of cases ('difficult' cancers), especially when the output of CAD was incorrect.

We argue that CAD, rather than having too little effect on readers' decisions to produce a measurable impact, had both beneficial and detrimental effects on readers' performance, but in the trial these effects compensated for each other, resulting in no significant impact on average sensitivity.

### 6.2 *Effects of the absence of prompts*

By choosing in our follow-up studies a test set with a large proportion of cancers missed by CAD, we managed to isolate (somewhat serendipitously) the potential detrimental effects of CAD on reader sensitivity. Participants who read our test set with the benefit of CAD (Study 1) showed a significantly lower sensitivity than those who saw the same cases without CAD (Study 2).

This effect was particularly marked for those cancers which CAD did not prompt and led us to conjecture that the absence of prompts may have a much bigger impact

on readers' decisions than anticipated. Our analyses of the data from both the HTA trial and our follow-up studies strongly suggest that readers may have used the absence of prompts on a mammogram as a sort of reassurance for their 'no recall' decisions for normal cases. It appears that, based on their experience with the tool, readers tended to (correctly) assume that the absence of prompting was a strong indication that a case was normal. The participants (both in the HTA trial and in follow-up Study 1) were very unlikely to recall cases for which CAD had issued no prompts. Again, these findings are broadly corroborated by our ethnographic observations.

One could argue that this is a rational approach. Readers perceived many of the prompts as distracting. As most mammograms contained prompts, the absence of prompts was more informative than their presence (especially if detailed analysis of every prompt was too demanding and thus practically infeasible). This can be beneficial when dealing with equivocal normal cases (see Table 2). However, as the results from our follow-up studies indicate, this can have damaging effects on readers' decisions for difficult-to-detect cancers that CAD does not prompt.

To our knowledge, such detrimental effects have not been reported before in the radiological/CAD literature. Earlier human factors studies of the effects of computer failures on human behaviour have shown that the failure of a computer aid to detect and warn of a target event could make users less likely to make the right decision for the event [17-18]. However, the participants in such studies are typically students working in artificial laboratory settings, while our participants were experts working in relatively realistic settings relevant to their area of expertise.

One plausible mechanism to explain these effects is that the absence of prompts made readers revise their decisions for ambiguous abnormalities they *had* already

detected. In other words, they may have used the absence of prompts as a reassurance for a 'no recall' decision when dealing with features they found difficult to interpret. It is possible that readers were using whatever evidence was available to resolve uncertainty. The implication is that CAD was being used not only as a detection aid but also as a classification or diagnostic aid, which is not what the tool is designed for. This is consistent with our observations (section 5) and also with earlier studies of CAD tools [19]. But we cannot exclude alternative mechanisms. For example, the absence of prompts may have caused readers to pay less attention to a case and, as a result, they may have failed to detect signs they would not have missed otherwise (as proposed by studies of 'automation bias' or 'over-reliance' on computer advice [17]). Although this is a plausible scenario, we have not found evidence to support its occurrence in the HTA trial or our follow-up studies.

*6.3 Difficulties of extrapolating results to real world practice*

Most of the findings reported in this paper are based on studies and analyses of readers' behaviour in trial conditions. Although there was an attempt to make the experimental settings reasonably realistic, many artificialities and simplifications were unavoidable. One must be careful, therefore, when extrapolating from the behaviours observed to effects in the field. We highlight here some important differences between the trial(s) and everyday practice:

a) A common criticism is that clinical trials are conducted with test sets containing unrealistically high proportions of pathological cases (so as to achieve sufficient statistical power with manageable numbers of cases). Evidence that radiologists do behave differently when faced with case samples containing different

prevalence of disease has been reported [20-21]. These effects have not received sufficient attention to date and are worth exploring further.

b) In everyday breast screening, readers have access to many other sources of information in addition to CAD (e.g. earlier mammograms, medical records, etc.); these sources of information may make readers interpret CAD output in different ways from how they did in the trials.

c) In the NHS Breast Screening Programme, reading is essentially a collaborative activity [22]: double reading is common practice and the final decision on a case is often the result of group discussion. In contrast, trial participants acted as though they were reading alone, which may have influenced how they interpreted the prompts and made their decisions.

d) In the trial, we saw how readers attempted to make sense of the tool's behaviour. In everyday practice, readers would have a better opportunity to gain a progressive understanding of how the tool works and to adjust their interpretation of its behaviour accordingly.

Despite these difficulties, many of the considerations derived from the clinical trial are relevant and useful. The statistical analyses, in conjunction with observations of humans in this and similar tasks, indicate plausible mechanisms that would cause the effects we observed, but do not allow us to decide which of these mechanisms will be active to a perceptible extent in a given activity. However, finding evidence of them in practice, even in partially artificial environments, is *prima facie* evidence of their being likely to occur. Designers of tasks and the computer tools to support them should consider how these behaviours may arise, and how to adapt tools and procedures to reduce those that are considered negative; assessors should consider them as factors that may change between the clinical trial and clinical use.

*6.4 Methodological implications for the evaluation of computer aids*

The use of standard clinical trial regime for evaluating new healthcare technologies has been subject to much criticism in recent years [23-26]. One problem that has been highlighted is that trial designs inevitably ignore the contextual nature of the work being supported, raising doubts about extrapolating trial results to real settings of use. Our decision to employ a multidisciplinary approach to CAD evaluation was partly intended as an exploration of how criticisms of clinical trials might be addressed.

Our experience of using ethnographic methods to complement statistical anlysis has shown some promise. Ethnography aims at understanding the 'situatedness' of work practices: i.e., how activities take place in their real-world context. The value of ethnographic methods has already become quite widely recognised as a way of informing requirements so that IT systems are designed appropriately for their actual circumstances of use [27]. In our evaluation of CAD, we also found that ethnographic methods can be valuable in addressing the 'ecological validity' of clinical trials by helping the interpretation of trial results to take into account differences between the context of the trial and a more realistic context of use.

First, it is only through ethnographic studies that we have been able to gain an understanding in detail of the character of everyday screening work and the context in which a CAD tool would be used. This has helped us identify possible mismatches between the tool as designed and readers' requirements [7]. As an example, we observed readers' perceived need to explain the behaviour of the CAD tool in order to use it properly and how they struggled to do this with the tool as designed. Second, ethnographic studies helped to reveal aspects of how readers actually used the CAD

tool in the trial. We found evidence of readers not adhering to the trial protocol of using the tool as an attention cue and, instead, using it at times as a decision aid. Ethnography and statistical analysis thus corroborated each other's conclusions. Third, ethnographic findings also influenced the choice of probabilistic models [4] by avoiding unrealistic assumptions that would invalidate results.

From a statistical viewpoint, a first consideration is simply that averages may hide substantial variations between sub-populations. Our statistical analyses, motivated by the 'diversity modelling' approach (reported in [3-4]) and its emphasis on how performance varies across classes of cases, have proven very useful here.

We found both beneficial and detrimental systematic effects of the use of CAD that just happen to cancel out in the trial (though the detrimental effects appear acutely in our follow-up studies). If these effects were to be present in practical clinical use, with different mixes of cases and readers from those in the trial, the net overall effect might be positive, negative or null, in addition to some possible transfer of risk between categories of patients. It might still be possible to estimate the net effect in future clinical use from the results in the trial [4], if further studies showed that the {case, reader} pairs can be classified by variables that can be estimated in both situations before introducing CAD and that are sufficiently predictive.

*6.5 Implications for CAD design and deployment*

We consider here the implications of four of our main sets of findings:

a) *Limited diversity between human and machine errors*. In the HTA trial, CAD errors were heavily correlated with the 'difficulty' of cases (see Sect. 3.3, Table 3). To increase the 'potential' advantage we estimated in Sect. 3.1, CAD should

prompt correctly those cases where unaided readers would tend to fail: its error pattern should be as 'diverse' as possible from the readers' (cf the mathematical models in [3-4]). Some improvement in CAD effectiveness could be sought by increasing this diversity, even without improving the average sensitivity or specificity of CAD. The tool could be tuned to be more sensitive for classes of cases on which readers tend to be less effective, as these cases are natural candidates for CAD to make a difference. Perhaps CAD thus tuned to be 'more diverse' from its users may improve the latter's performance more than CAD simply tuned to be very good (in terms of sensitivity, specificity or any weighted combination of the two). We believe that this possibility is worth exploring although how much gain (if any) it would produce depends on the details of the specific CAD algorithms and the degrees of freedom in tuning them. It might even be desirable to adapt CAD tuning differently for each individual reader, automatically or manually.

b) *Evidence of systematic positive and negative effects of CAD.* Specific human behaviours that are considered undesirable may be targeted with methods for either avoiding them or correcting their negative effects ('fault tolerance'). We could think of many solutions, whose acceptability in the specific working environment would need to be checked before they are adopted. For example, the cognitive load on readers could be reduced by not repeating prompts on features that readers have already noticed and marked. Similarly, the self-calibration of readers, and their reliance on the CAD outputs at times as a decision aid instead of an attention cue, could be corrected by including in their normal workload fictitious cases with incorrect CAD outputs, rarely but yet frequently enough to refresh their memory of types of possible errors by CAD. Changes could also be

made to the CAD tool so that it makes readers' violations of prescribed protocols of use impractical or harder to justify.

c) *Readers' evident concern to explain the presence or absence of prompts.* One approach would be to explore how the CAD tool could provide explanations of its behaviour on demand. The challenge would be to produce the sorts of accounts that would be useful to a reader, which calls for an understanding of the sorts of explanations where confusion may arise. The question also exists whether such exacting specifications could be implemented reliably enough not to make the explanation facility a source of further problems.

d) *Readers seemed to be over-influenced by the absence of prompts.* When using CAD, readers are required to pay attention to prompts and not to their absence but, in reality, they seemed to be doing the opposite. Design changes might be implemented to combat this problem (e.g. by introducing mechanisms that make readers aware of the risks of this behaviour). However, one could argue that the requirement that readers not be influenced at all by the absence of prompts is possibly psychologically impossible to satisfy as well as normatively incorrect, if the absence of prompts has indeed informative value; in this case, there could be merit in seeking to provide readers with a heuristic procedure to follow that would give absence of prompts approximately the right weight in decisions, rather than attempting to have them ignore it altogether.

## 7. Concluding remarks

The overall aim of our work is to examine more closely the effects of decision aids on human judgement and the factors that affect the dependable performance of human-

computer systems. We have used CAD as a case study for exploring some of the subtler interactions between human decisions and computer-generated evidence. Our multidisciplinary approach combining quantitative and qualitative methods has helped to detail and understand some of the possible factors underlying the apparent lack of effects of CAD in some trials, and to identify areas for further development. Statistical analyses identified some possibly detrimental effects of CAD on reader performance. These findings were corroborated by the ethnographic studies that provided important insights into the manner in which readers came to use the tool as a part of their work, and by follow-up experiments. In this way, the various methods used helped to orient each other and improve confidence in our findings.

Our multidisciplinary approach has helped to address some of the recognised limitations of the clinical trial as an evaluation methodology for IT-based healthcare interventions. The use of ethnographic observation, in particular, has helped to put the clinical trial results into a real world context of use and thereby gain a better understanding of their implications for the everyday work of reading mammograms. We would note also that the introduction of a new technology such as CAD may change work practices and thus, in an iterative process, make new demands on the technology. Evaluation methodologies that do not take into account the influence of adaptation and learning may fail to deliver meaningful results for users and designers alike and to contribute to improving the technology. Again, our multidisciplinary approach has enabled us to gain some perspectives on how readers learn and adapt to the behaviour of the CAD tool and suggested some promising lines of enquiry for how the tool and the procedures for using it may themselves be adapted.

In conclusion, our studies have provided some insight about the effects of CAD on readers' performance, as well as some methodological indications. Both results

may be of interest for a larger class of tasks and computer aids than that of CAD for mammography.

## ACKNOWLEDGMENTS

The work described in this paper has been partly funded by the U.K.'s Engineering and Physical Sciences Research Council (EPSRC) through DIRC, the Dependability Interdisciplinary Research Collaboration, a project investigating the dependability of computer based systems. We would like to thank: DIRC collaborator Mark Rouncefield for his substantial contributions to the ethnographic work summarised here; R2 Technologies Inc. (and very especially Gek Lim, Jimmy Roehrig and Julian Marshall) for their support in obtaining the data samples for our follow-up studies; Paul Taylor and Jo Champness (from UCL) for granting us access to their data, facilitating the follow-up studies and helping run them; all the readers who volunteered to take part in the follow-up studies; and, last but by no means least, David Wright for his invaluable feedback on earlier drafts of this paper.

**REFERENCES**


1. Taylor PM, Champness J, Given-Wilson RM, Potts HWW, Johnston K. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. Br J of Radiol 2004; 77(913): 21-27.
2. US Food and Drug Administration. Pre-market approval decision. Application P970058. June 26, 1998, http://www.fda.gov/cdrh/pdf/p970058.pdf
3. Littlewood B, Popov P, Strigini L. Modelling software design diversity – a review. ACM Computing Surveys 2001; 33(2): 177-208.
4. Strigini L, Povyakalo AA, Alberdi E. Human-machine diversity in the use of computerised advisory systems: a case study. In: Proceedings of DSN 2003, International Conference on Dependable Systems and Networks, San Francisco, 2003; 249-258.
5. Povyakalo AA, Alberdi E, Strigini L, Ayton P. Evaluating ``Human + Advisory computer" system: A case study. In Proceedings of The 18th British HCI Group Annual Conference. Leeds, UK, 6-10 September 2004, Vol. 2, in press.
6. Alberdi E, Povyakalo AA, Strigini L, Ayton P. Effects of incorrect CAD output on human decision making in mammography. Acad Radiol, in press.
7. Hartswood M, Procter R, Rouncefield M, Slack R, Voss A. ‘Repairing’ the Machine: A Case Study of the Evaluation of Computer Aided Detection Tools in Breast Screening. In Dourish, P and Fitzpatrick, G. (Eds.) Proceedings of the European Conference on Computer Supported Cooperative Work, Helsinki, September 14th-17th, 2003.
8. Castellino R, Roehrig J, Zhang W. Improved computer aided detection (CAD) algorithms for screening mammography. Radiology 2000; 217(P) 400.

9. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology. 2000;215:554-562.
10. Hartswood M, Procter R, Williams L. Prompting in mammography: Computer-aided detection or computer-aided diagnosis? In Proceedings of Medical Image Understanding and Analysis, MIUA 98, Leeds, UK, 6-7 July 1998, http://www.robots.ox.ac.uk/~mvl/miua98/proceedings.html
11. Agresti A. Categorical data analysis. New York: Wiley. 59-66 (1990)
12. Thurfjell E, Thurfjell G, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. Acta Radiologica 1998; 39: 384-388.
13. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computed assisted detection of interval breast cancers. Eur J Radiol. 2001 Aug;39(2):104-10
14. Marx C, Malich A, Grebenstein U et al. Are Unnecessary Follow-up Procedures Induced by Computer-aided Diagnosis (CAD) in Mammography? Comparison of Mammographic Diagnosis with and without Use of CAD. In Proceedings of the 88th Scientific Assembly and Annual Meeting of Radiological Society of North America (RSNA 2002), Chicago, USA, December, 2002.
15. Brem RF, Schoonjans JM Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. Clin Radiol 2001 Feb; 56(2):150-4.
16. Ciatto S, Del Turco MR, Risso G, et al.Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography. Eur J Radiol. 2003 Feb;45(2):135-8.

17. Skitka LJ. Accountability and automation bias. Int. J. Human-Computer Studies 2000; 52: 701-717.
18. Meyer J. Effects of warning validity and proximity on responses to warnings. Human Factors 2001; 43(4): 563-572.
19. Hartswood M, Procter R. Computer-Aided Mammography: A Case Study of Error Management in a Skilled Decision-making Task. Journal of Topics in Health Information Management 2000; 20(4): 38-54.
20. Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. Journal of the American Medical Association 1996; 76(21): 1752-1755.
21. Rutter CM, Taplin S. Assessing mammographers' accuracy: A comparison of clinical and test performance. Journal of Clinical Epidemiology 2000; 53: 443-450.
22. Hartswood M, Procter R, Rouncefield M, Slack R. Performance Management in Breast Screening: A Case Study of Professional Vision and Ecologies of Practice. In Johnson, C. (Ed.), special edition on Human Error and Medical Work, Journal of Cognition, Technology and Work 2002; 4(2): 91-100.
23. Heathfield, H., Wyatt, J. Philosophies for the design and development of clinical decision support systems, Methods of Information in Medicine 1993; 32(1): 1-8.
24. Heathfield H, Buchan I. Letters: Current evaluations of information technology in health care are often inadequate, BMJ 1996; 313: 1008.
25. Heathfield H, Pitty D, Hanka R. Evaluating information technology in health care: barriers and challenges. BMJ 1998; 316:1959-1961.
26. Sullivan FM, Pagliari HC, Mitchell ED. Health Informatics. London, RCGP, 2002.

27. Hughes, J., King, V., Rodden, T. and Anderson, R. (1994). Moving Out from the Control Room: Ethnography and Systems Design. In Proceedings of the ACM Conference on Computer-Supported Cooperative Work, ACM Press, pp. 429-439.

# TABLES & FIGURE

***Table 1. Two categories of human error for cancers, in the prompted condition***

| **Sub-populations of cancers** | **Errors on cases correctly processed by CAD** | **Errors that readers did not make in the unprompted condition** |
|---|---|---|
| detected by all readers in the unprompted condition | 4 | 4 |
| detected by more than 50% of readers in the unprompted condition | 144 | 52 |
| all cases | 279 | 97 |

***Table 2 Contingency tables for correctness of CAD output and change of reader's decision between two conditions****

| | **Reader's decisions for Cancers**** | | **Reader's decisions for Normal cases***** | |
|---|---|---|---|---|
| | prompted decision worse | prompted decision better | prompted decision worse | prompted decision better |
| incorrect CAD output | 89 | 73 | 495 | 443 |
| correct CAD output | 113 | 147 | 42 | 103 |
| Fisher's Exact Test for Count Data | p-value=0.0274<br>estimated odds ratio=1.58 | | p-value=9.86 $10^{-8}$<br>estimated odds ratio=2.74 | |

* Complete count data about readers' decisions are shown in Table 4.

** Total number of complete pairs of decisions for cancers =2988 (12 data points missing)

*** Total number of complete pairs of decisions for normal cases =5986 (14 data points missing)

### *Table 3 Probabilities of readers' errors* for different sub-populations of cases***

| **Sub-populations of cases** | **Cancers** | | **Normal cases** | |
|---|---|---|---|---|
| | Prompted condition | Unprompted condition | Prompted condition | Unprompted condition |
| Correct CAD output | 0.124 (0.111, 0.139)** | 0.140 (0.126, 0.155) | **0.097*** (0.080,0.117)** | **0.155*** (0.134,0.179)** |
| Incorrect CAD output | 0.594 (0.558, 0.630) | 0.571 (0.534, 0.607) | 0.200 (0.189,0.211) | 0.189 (0.179,0.201) |
| All cases | 0.242 (0.227, 0.258) | 0.247 (0.232, 0.263) | 0.182 (0.172,0.192) | 0.183 (0.174,0.193) |

* Note that sensitivity and specificity can be computed as follows: sensitivity = 1 - "probability of error for cancers"; specificity = 1 - "probability of error for normal cases"
** Complete count data for estimating these probabilities are shown in Table 4
***95% confidence interval
**** Statistically significant difference

### *Table 4 Contingency tables for correctness of readers' decisions and CAD's output for different categories of cases in different conditions**

| | **Cancers** | | **Normal cases** | |
|---|---|---|---|---|
| | Correct CAD output | Incorrect CAD output | Correct CAD output | Incorrect CAD output |
| **With CAD** | | | | |
| Correct reader decisions | 1932 | 321 | 948 | 3956 |
| Incorrect reader decisions | 314 | 427 | 102 | 987 |
| Exact Fisher's Test | estimated odds ratio=8.176 p-value < $2.2\ 10^{-16}$ | | estimated odds ratio = 2.319 p-value < $2.2\ 10^{-16}$ | |
| **Without CAD** | | | | |
| Correct reader decisions | 1966 | 304 | 887 | 4007 |
| Incorrect reader decisions | 279 | 445 | 163 | 936 |
| Exact Fisher's Test | estimated odds ratio = 10.300 p-value < $2.2\ 10^{-16}$ | | estimated odds ratio = 1.271 p-value = 0.009542 | |

* Estimated probabilities of readers' errors with 95% confidence intervals are shown in Table 3.

### *Table 5. Proportions of incorrect human decisions in follow-up Studies 1 and 2*

| | **Correctly marked by CAD** | | **Incorrectly marked by CAD** | | **Unmarked by CAD** | |
|---|---|---|---|---|---|---|
| | **Study 1** (with CAD) | **Study 2** (without CAD) | **Study 1** (with CAD) | **Study 2** (without CAD) | **Study 1** (with CAD) | **Study 2** (without CAD) |
| **Cancer** | 0.19 | 0.10 | **0.45** | **0.33** | **0.67** | **0.46** |
| **Normal** | n/a | n/a | 0.08 | 0.13 | **0.06** | **0.12** |

**Note:** The proportions have been calculated out of the total number of 'recall/no recall' decisions generated by the participants (20 in Study 1 and 19 in Study 2) in each case category. Bold type indicates that there was a statistically significant difference between the proportions in Study 1 and those in Study 2 for the same sub-population of cases (ANOVA).

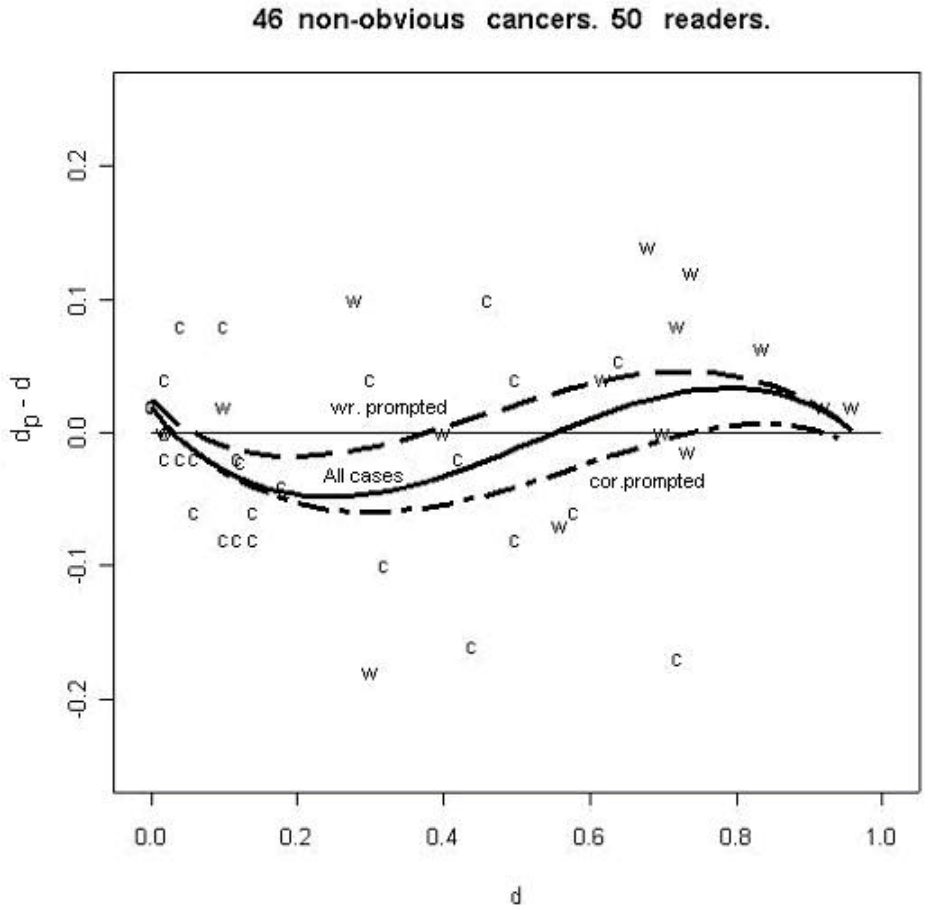


**Figure 1.** Effect of correctness of CAD output on case difficulty (rate of FN errors): $d_p$: prompted difficulty; $d$: unprompted difficulty; $w$: cancer that was wrongly processed by CAD; $c$: cancer that was correctly processed by CAD; dashed curve: logistic regression curve for wrongly processed cancers; dashed-dotted curve: logistic regression curve for correctly processed cancers; solid curve: logistic regression curve for all cancers.