JCARS/CARS 2008 Submission

Title:

CAD in mammography: lesion-level vs. case-level analysis of the effects of prompts on human decisions

Authors: Eugenio Alberdi[1], Andrey A. Povyakalo[1], Lorenzo Strigini[1], Peter Ayton[2], Rosalind Given-Wilson[3]

Affiliations:
[1]Centre for Software Reliability, City University, London, UK
[2] Psychology Department, City University, London, UK
[3] Department of Radiology, St. George's Hospital, London, UK

Communicating author:

Eugenio Alberdi, Ph. D.

Address: Centre for Software Reliability, Northampton Square, City University, London, EC1V 0HB, United Kingdom

Email: e.alberdi@csr.city.ac.uk

Telephone: +44 20 7040 8424

Fax: +44 20 7040 8585

Word count – abstract: ~200

Word count – main text: ~5100

Number of figures: 1

Number of tables: 3

Number of references: 18

Abstract

OBJECT:

To understand decision processes in CAD-supported breast screening by analysing how prompts affect readers' judgements of individual mammographic features (lesions). To this end we analysed hitherto unexamined details of reports completed by mammogram readers in an earlier evaluation of a CAD tool.

MATERIAL AND METHODS:

Assessments of lesions were extracted from 2889 reports for 59 cancer cases. Statistical analyses of these data focused on what features readers considered when recalling a cancer case and how readers reacted to CAD prompts.

RESULTS:

13.5% of recall decisions were found to be caused by responses to features other than those indicating actual cancer. Effects of CAD: lesions were more likely to be examined if prompted; the presence of a prompt on a cancer increased the probability of both detection and recall especially for less accurate readers in more subtle cases; lack of prompts made cancer features less likely to be detected; false prompts made non-cancer features more likely to be classified as cancer.

CONCLUSION:

The apparent lack of impact reported for CAD in some studies is plausibly due to CAD systematically affecting readers' identification of individual features, in a beneficial way for certain combinations of readers and features and a damaging way for others. Mammogram readers do not ignore prompts. Methodologically, assessing CAD by numbers of recalled cancer cases may be misleading.

Keywords: computer-assisted diagnosis, mammography, screening, computer systems evaluation, decision making

Introduction

The effectiveness of CAD (Computer Aided Detection) in assisting mammogram readers in early cancer detection (e.g. in breast screening) has been the subject of intense debate in the literature. At best, the available evidence indicates highly variable results between different settings, with some controversy (see reviews in [1] and [2]). Whereas some studies show that the introduction of CAD results in a significant increase in cancer detection [3, 4], other studies show negligible effects [5, 6], inferior to the benefits of double reading [7]. A recent study has shown harmful effects of CAD. Its use was found to be associated with reduced accuracy of interpretation of screening mammograms. This was reflected in reduced specificity of CAD-supported readers [8]. Additionally, our previous studies [9] have shown a strong association between the absence of correct prompts on cancer areas and reduced reader sensitivity (also supported by [10]).

A number of possible explanations can be proposed to explain the harmful (or negligible) effects of CAD. For example: readers may ignore the true prompts on cancer areas; they may unduly respond to false prompts; or they may be falsely reassured by the absence of prompts on a subtle cancer [11]. Readers might also start to use CAD prompts as the primary cue for suspecting cancer, rather than as the useful aid they are intended to be for reducing the risk of overlooking features.

CAD tools are continually evolving in an attempt to improve their effectiveness (measured mostly by false positive and false negative rates). We need a more detailed understanding of how the technology affects human decisions. Without this understanding, it is not possible to either resolve the controversy about the effectiveness of the tools or to improve their design and assessment.

In this paper we present the preliminary results of a research project aimed at a better understanding of the effects of CAD on the decision making. This work continues exploratory statistical analyses that we reported previously [12], in which we used the raw data from an earlier controlled study (run by other investigators) [6]. We investigated the effects of the variation in sensitivity among mammogram readers combined with variation in difficulty of detection among cancers. A previous, theoretical study [13] had shown that such variation may have marked effects on the average effect of a CAD tool on reader proficiency. Our exploratory analyses strongly suggested that CAD improved the sensitivity of less proficient readers when dealing with relative easy cases and *reduced* that of highly proficient readers when dealing with difficult cases.

However, in our earlier exploratory analyses (in common with much prior literature about the effectiveness of CAD) we used data that were collected at the level of *cases* rather than *mammographic features*. Reader decisions were classified simply as recall/no recall decisions (characterised on a 1 to 4 scale), without further details. This omitted much information from the readers' reports that may shed light on how exactly the CAD prompts affected readers' decisions. For instance, suppose that a reader recalled a particular case when using CAD but did not when not using CAD. It might seem natural to ascribe this effect either to the CAD prompts or to natural variability in a reader's sensitivity. But suppose that the reader's report indicated that the recall decision was based on a mammographic feature that CAD had *not* prompted (a piece of information that could not be analysed prior to the current project). This would suggest either that CAD had no role in the recall decision, or that it did affect it but in a different way from that intended by designers of CAD tools (e.g. CAD may have caused people to be differentially sensitive to mammographic features that were not prompted by CAD).

In the work we report here we first entered into an electronic data base detailed information about several thousand mammography reports, and then analysed these data both to check and to refine our earlier conclusions. We used the rich, more detailed, information entered in the new electronic database to elucidate the specific mammographic features that had been

identified by the readers and how the judgements about those features were affected by CAD. In other words we analysed readers' decisions (e.g. their sensitivity) and CAD effects at the level of mammographic features rather than merely at the level of cases.

We now briefly describe the two studies on which the current analyses are based: the "HTA Study" and the "DIRC analyses".

The HTA study on CAD effectiveness

The previous study whose raw data we used (funded by the UK's Health Technology Assessment program, henceforth, the *HTA study*), was one of the largest controlled retrospective studies ever conducted to evaluate the effectiveness of CAD for breast screening. It used fifty readers, who looked at mammograms of 180 patients twice (once with and once without CAD) to decide whether or not the case should be recalled for further investigation. These cases included 60 biopsy proven cancers, which were a mixture of screen detected and prior films of false negative interval cancers. These were chosen to provide cases with a range of mammographic appearances and difficulty of detection. The CAD system was the R2 ImageChecker, applied to digitised analogue films. Readers were all experienced readers (radiologists and advanced practitioners) working in the NHSBSP (the UK's National Health Service Breast Screening Programme) and reading a minimum of 5000 mammograms annually. For each case read they recorded data on a paper pro-forma with a pictogram of the mammograms. Data included a mark indicating the position of each feature on the films, type of feature and level of suspicion.

The HTA study reported a lack of statistically significant effects of CAD on either the sensitivity or specificity of readers. Various explanations are possible, most simply that readers' decisions were largely unaffected by CAD: readers "ignored" CAD prompts. Our subsequent work, however, casts doubt on this interpretation.

The DIRC exploratory analyses

Given our interest in the effect of computer support on the reliability of human decisions, the investigators of the HTA study kindly made their raw data available to us, and we performed a finer-grained, exploratory statistical analysis (as part of the UK-based research project, DIRC – Interdisciplinary Research Collaboration on Dependability) [12, 14]. Our analyses (henceforth the *DIRC analyses*) suggested that, although there was no difference on average between readers' performance with and without CAD, the readers were actually affected by the computer prompts. The negligible *average* effects hid significant systematic effects (beneficial for some subsets of decision types and harmful for others) which compensated for each other, leading to an apparent lack of effect [12]. Computer support appeared to *reduce* the sensitivity of some readers, mostly those who were the most sensitive without CAD. This may imply over-reliance on computer outputs [9], as documented by other studies of computer-supported decision-making [15]. However, many alternative explanations were possible [14], for instance in terms of increased cognitive load from processing the prompts, or of readers instinctively readjusting their decision thresholds when using CAD.

Material and Methods

Data Entry

The purpose of the data entry phase was to produce electronic files in which detailed information (about mammographic cases and readers' decisions) is encoded in a format suitable for statistical analysis. The resulting files identify all the different visual features (areas of interest) which had one or more of these characteristics: a) identified by an expert radiologist as cancer regions, based on biopsy reports; b) marked by CAD ("prompted") with a *prompt* designating the type of abnormality (either calcification or irregular mass); and c) marked by readers on the pictogram of the mammogram on the pro-forma. Henceforth we will refer to these areas as Marked Areas of Interest (or MAOIs).

Additionally, the electronic files include other information that readers entered on the pro-forma: the degree of suspicion and type of abnormality the reader assigned to each MAOI s/he marked, plus other possible notes, and a recall/no recall decision about the whole case.

We expected the data entry phase to present difficulties, the main one being that we wished to record whether each MAOI identified by a reader was also prompted by CAD and was or was not a cancer area. It can be difficult to establish whether two marks by different readers are actually intended to identify the same feature of the image due to subtle differences in the position of marks on pictograms. Not knowing in advance how frequent this difficulty would arise, we planned a data entry procedure that would record any such ambiguous situation as two different MAOIs, with information about their relative position. It would then be possible at a later stage either to resolve the uncertainty through advice from mammography experts, or to perform the statistical analyses assuming each one of the alternative explanations, and thereby ensure that any conclusions would be robust against any such ambiguities of interpretation of the mammography reports.

The other major concern was the risk of human error in the data entry procedure. Data were entered into a Microsoft Excel spreadsheet that featured pre-set menus for most fields in the spreadsheet to reduce, for example, the risk of "nonsensical" entries from typing errors.

The data entry procedure was as follows. For each cancer case, first the CAD prompts and the retrospective identification of the cancer areas (by the expert) were entered in the spreadsheet. This information was also copied (by manually tracing the prompts and the expert's marks) to a transparency (a standard A4 "acetate" film for overhead projection). This transparency was then superimposed on the images of the same mammograms in each pro-forma in turn. If an area already traced on the transparency coincided with a mark by a reader in the report, the inference was that the reader had identified that same area and so no other trace was created on the transparency; although a record was entered in the spreadsheet indicating that the reader had marked that area. Any other (new) mark made by a reader was traced on the transparency and identified and listed in the Excel spreadsheets.

As noted, an obvious concern with such a labour intensive activity (requiring high levels of concentration for a very large set of data), was the risk of data entry errors. In addition to the precautions above, we subjected the data files to an automatic search for inconsistencies through a software tool (developed by an MSc student at City University's Department of Computing). This checked for the satisfaction of several "consistency rules". For example, if MAOI A had been marked by reader 1, the record for this MAOI should also contain information about the type of abnormality (mass or calcification) and the level of suspicion (1-5) assigned by the reader to that MAOI. The violation of this "consistency rule" would indicate incomplete or erroneous data entry. Other consistency rules checked for duplicated data. These automatic checks flagged inconsistencies in a small proportion (about 2.5%) of the electronic records. All flagged inconsistencies were checked by a member of the team who examined the original paper reports to determine the cause of the detected inconsistency and then corrected the electronic database accordingly.

In previous analyses [12] we had identified 14 cases as "obvious" cancer cases; "obvious" because they had been recalled by all 50 readers in both experimental conditions (with and without CAD). These cases were entered in a terser format than the other 45 "non obvious" cancer cases (those that at least one participant in the HTA study did not recall). The initial analyses (described next) concern the "non obvious" cases.

Analysis 1: Recalls for non-cancer areas (false-target recalls)

A confusing factor in the DIRC analyses (i.e., before we constructed the new database) was that one could not judge how many of the decisions to recall a cancer case – the decisions that appeared to be correct – were actually based on the "correct" process, i.e., due to the reader correctly interpreting the cancer features in the mammogram. To illustrate how an apparently correct recall may be due to the wrong process, we can imagine a mammogram where the

cancer features were very difficult to identify, and on which CAD would only add "false" prompts, i.e., on non-cancer areas. It is conceivable that a reader who did not recall the patient when not using CAD, might then recall it when using CAD, after spuriously judging a non-cancer area as suspicious. Such a recall decision might or might not lead to the cancer being detected by biopsy, since it would be accompanied by initially misleading indications about the results of mammography. Hence, it is difficult to judge whether such recall decisions are a desired outcome of screening. Nonetheless, it is clear that these cases should be seen as different from other recalls of cancer cases when studying how CAD affects readers' decisions, since, with these recalls, CAD might be the cause of the recall *despite not achieving its primary goal* of helping readers to detect cancerous features.

Anecdotal evidence of such "false-target recalls" was found during the DIRC analyses. Accordingly, it is conceivable that some of the interesting CAD effects on readers' sensitivity highlighted by the DIRC analyses [14] were actually effects on the probability of "false-target recalls" - not of correct identification of cancer. To check for this possibility, we carefully classified recalls *of cancer cases* into "true-target" and "false-target" depending on whether the reader identified at least one cancer feature as suspicious (or at least as the most suspicious in the mammogram). Table 1 summarises the results.

We also compared the readers' frequency of false-target recalls with and without CAD.

Analysis 2: Effect of prompts on readers' classification of a mammographic feature linked to cancer

In this analysis, we compare the "per feature sensitivity" – the probability of marking a cancer area as suspicious – of readers when using CAD *vs.* when not using it. In the 45 "non obvious" mammograms of cancer cases, the expert radiologist identified 82 individual mammographic features representing malignancy. Of these, CAD prompted 48 with the correct prompt, prompted one with a wrong prompt (i.e. suggesting the wrong kind of lesion), and did not prompt the remaining 33.

Using the richer data now available, we looked at all features associated with cancer (cancer MAOIs). For each such MAOI, a reader may have marked it, and indicated a degree of suspicion (between 1 and 5), or not marked it (which we conventionally classify as a 0 level of suspicion). Levels of suspicion of 3 or more indicate that the reader believes the feature to be either cancer or "indeterminate", that is, is likely to cause a recall decision; accordingly we considered that the reader had identified the feature as "suspicious".

We investigated whether the presence of a correct prompt on such a cancer MAOI made the reader any more likely to mark it when using CAD than when not using it, and to deem it suspicious.

Analysis 3: Readers' reactions to areas marked by CAD

One of the provisional conclusions of the DIRC analyses was that the lack of significant effects of CAD in the HTA study could not be explained as the effect of readers simply "ignoring" CAD prompts (consciously or unconsciously, e.g. because using them would require too much extra effort, or because the abundance of false prompts caused readers to regard all prompts with suspicion). This was because we observed that CAD appeared to have opposite effects on sensitivity for different groups of readers and cases. However, a doubt remained whether this was due to the prompts per se, or to a different attitude of some readers to their task when assisted by CAD.

We classified all CAD prompts as "true positives" (indicating cancer areas) or "false positives" (prompting non-cancer areas).

Similarly to Analysis 2, we compare the probability of one of the MAOIs prompted by CAD being marked or identified as suspicious by readers using CAD (i.e., when the readers see the

prompt) *vs.* readers not using CAD (i.e., they see the same mammogram but without the prompts).

Analysis 4: Exploratory regression analysis of readers' reactions to prompts

We also performed an exploratory non-linear regression analysis, similar to the DIRC analyses [14, 16]. We estimated the probabilities of reader judgement errors about prompted cancer MAOIs as functions of two independent variables:

1. the MAOI's "difficulty", measured as the proportion of readers who erroneously assessed that MAOI as non-cancer (i.e., marked it with suspicion <3 or did not mark it at all), over all reading sessions in the HTA study (in both experimental conditions, with CAD & without CAD), and
2. the reader's average "effectiveness" or "discrimination ability", a measure of the reader's ability to discriminate between cancer and non-cancer cases, calculated as the difference between the proportion of recall decisions for cancers (i.e. reader sensitivity) and the proportion of recall decisions for normal cases (i.e. 1 minus reader specificity), over all reading sessions in the HTA study (in both experimental conditions, with CAD & without CAD).

Separate regression estimates of the probability of reader error were calculated for the "with CAD" condition and for the "without CAD" condition. The goal was to estimate CAD's "impact", which we calculated as the difference between the two regression estimates, that is, an estimate of the possible effect of seeing CAD prompts on the reader's assessment of a certain MAOI. The statistical procedures used in these analyses are described in more detail in [12]. The independent variables were calculated over all reading sessions in *both* experimental conditions to avoid estimation biases related to regression towards the mean.

Results

Resulting Database

The electronic database produced contains full detailed information from a total of 2889 reading sessions for 59 cancer cases by 50 readers. The sample in the HTA study included 60 cancers, but one of these had to be excluded from processing in the current project because its documentation was no longer available.

Results of Analysis 1: Recalls for non-cancer areas ("false-target recalls")

Decision	without CAD	with CAD
true-target recall	1242	1254
false-target recall	196	196
no recall	704	692
Total	2142	2142

Table 1 Classification of decisions, highlighting "false-target" recalls, i.e., recall decisions without correct identification of the cancer areas (with and without CAD)

Table 1 shows the numbers of the various kinds of decisions for the 45 "non obvious" cancers, for those reading sessions for which reports were available from both the "with CAD" and "without CAD" conditions.

Interestingly, about 13.5% of the recall decisions appear to be "false-target recalls". Surprisingly, the number is identical (196) in the two conditions (with and without CAD), although only 78 decisions (reader-cancer pairs) were "false-target recalls" in *both* conditions. Decisions that in one condition were "false-target recall" ended up, in the other conditions, as

"true-target recall", "false-target recall" or "no recall" in proportions of approximately 25-40-35."

Indeed, the lack of effect of CAD on readers' propensity towards false-target recalls remains when we look at how an individual reader's decision about a particular cancer changes between the two conditions (which we tested with McNemar's Chi-squared test, showing no statistically significant effect; p-value 0.88).

Results of analysis 2: Effect of prompts on readers' classification of a mammographic feature linked to cancer

	changes with use of CAD in the frequency of readers marking a cancer area (any level of suspicion)	changes with use of CAD in the frequency of readers identifying a cancer area as suspicious (suspicion level ≥3)
Correctly prompted features	0.0163 (0.0074)	0.0202 (10^{-4})
Unprompted/wrongly prompted features	-0.047 (-0.0193)	-0.0372 (-0.012)
All cancer features	-0.0099 (no significant effect)	-0.0036 (no significant effect)

Table 2: Effects of CAD on probability of detection and identification of cancer features (in parentheses: 95% one-sided confidence bound on change in probability)

Table 2 summarises the results of Analysis 2. To test the hypothesis that seeing the CAD prompts has an effect on readers' detection of, or judgement about, a cancer MAOI, we calculated one-sided, lower confidence bounds on the apparent effect.

The bottom row of Table 2 shows that using CAD did not significantly change, on average, the probability of correct interpretation of a cancer feature. However, the *lack* of a correct CAD prompt significantly decreased both the probability of a reader marking a cancer area, and the probability of a reader indicating the area as suspicious. The *presence* of a prompt had significant positive effects on both measures.

Results of analysis 3: Readers' reactions to areas prompted by CAD

	changes with use of CAD in the frequency of readers marking a "prompted" area (any level of suspicion)	changes with use of CAD in the frequency of readers identifying a "prompted" area as suspicious (suspicion level ≥3)
True positive prompts	0.028 (0.004)	0.028 (0.008)
False positive prompts	0.123 (0.084)	0.049 (0.034)

Table 3: Effects of CAD prompts on readers - all features prompted by CAD (in parentheses: 95% one-sided confidence bound on change in probability)

Table 3 summarises the results of Analysis 3. The left-hand column shows a clear and significant increase in the probability of a reader marking a cancer feature (a true positive prompt) and an even greater increase for *non-cancer* features (lower bound: 8.4% increase) if the reader can see that the feature is prompted.

The right-hand columns shows that seeing the CAD prompts appeared to slightly increase (lower confidence bound slightly less than 1%) the probability of a reader marking an actual cancer feature as suspicious; and also to increase to a larger extent the probability of a reader marking a *non-cancer* area as suspicious.

Results of analysis 4: Exploratory regression analysis of readers' reactions to prompts

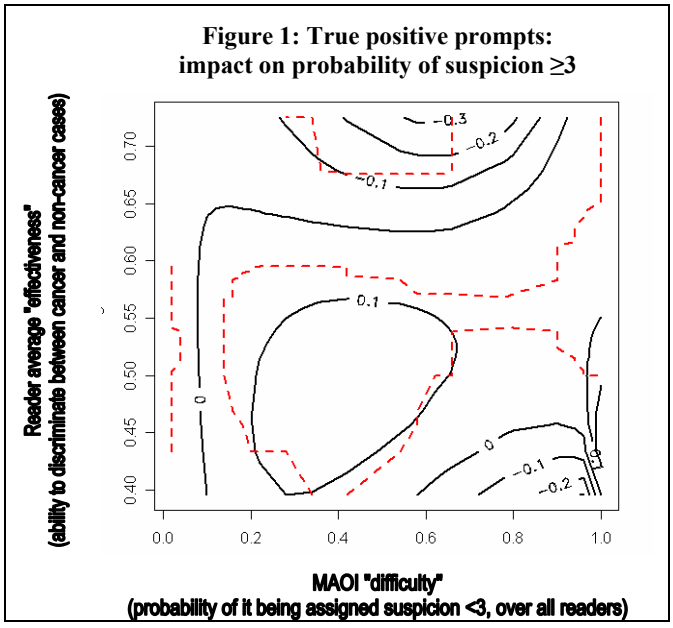

Figure 1 shows the results of the regression analysis. The areas marked in the figure represent different degrees of estimated CAD "impact" (as defined earlier). It can be seen that the estimated impact varies form up to 0.2 in the irregular ellipsoid in the central-lower part of the graph, to negative values near the bottom-right corner (high MAOI "difficulty" and low reader "effectiveness", top centre (medium MAOI "difficulty" and high reader "effectiveness" and on the left (low MAOI "difficulty").

The dashed lines bound areas where the impact is statistically significant (95% confidence level): with negative values near the left and near the top margin of the figure; with positive value for a large area that extends from the middle to the right edge of the figure. For example, for readers of "effectiveness" 0.50 when judging correctly prompted cancer MAOIs of 40% difficulty, the estimated CAD impact is between 0.1 and 0.2 (i.e., they fall within the big ellipsoid area in the centre). Since this ellipsoid area is mostly bound by a dashed outline, this estimated impact is deemed statistically significant.

The estimates suggest that a CAD prompt significantly reduces (by up to 10%) the probability of false negative error for the less effective readers (those whose discrimination ability is less than 60%: more than 2/3 of our sample), and mostly on MAOIs whose difficulty (measured across all readers) is moderate.

Another statistically significant pattern appears at the top of the figure. For the 10-20% most effective readers there is a significant *negative* estimated effect, i.e., an estimated *increase* in the probability of false negative errors about features of moderate difficulty (around 50%).

Discussion

One of the novel outcomes of the analyses just described is that "false-target recalls" (i.e., recalls of cancer cases without correct identification of the cancer features) occurred in 13.5% of cases with or without CAD (Table 1). This result should be treated with caution – our criterion for deciding whether a reader had correctly identified the cancer features is just one of several plausible different criteria, some of which we plan to apply for comparison – but the evidence that these situations can occur so frequently has important methodological implications for the evaluation of CAD. Although these tools are designed to help in the

identification of individual features, their effectiveness is most frequently measured in terms of readers' recall of the whole case. However, the presence of these false-target recalls does not affect the validity of the other findings we discuss below.

As for how CAD affects readers' decisions, the results presented here support the general conclusions drawn from our previous (DIRC) analyses [12, 14] that CAD has a mixture of subtle effects on readers' decisions and refute alternative explanations of the results of the HTA study that CAD has no impact or that readers ignore CAD prompts.

The preliminary analyses we present here indicate that the lack of significant effect of CAD use on readers' sensitivity, observed at the level of whole mammograms (recall decisions) in the HTA study [6], is also present, on average, at the level of individual cancer features. Critically however, this outcome is the result of *significant* effects *(*in opposite directions*)* of *individual prompts*. For cancer areas that CAD prompted, there is an increase in the probability of readers correctly identifying them. For non prompted cancer areas, there is a (more marked) decrease in the probability of readers identifying them. The alternative conjecture that readers simply tended to "ignore" CAD prompts is refuted. It is safe to conclude that readers *did see* and *reacted* to prompts. For instance, for some feature that they would have noticed without a CAD prompt, and mentally tagged as irrelevant, a prompt may have caused them to mark it in the pro-forma; or for some cancer feature that they would have missed, a prompt may have led them to notice and analyse the feature (the intended role of CAD). CAD prompts, when correct, did help readers to identify more cancer features. However, using CAD also reduced the readers' ability to identify those cancer features that CAD did not prompt (Table 2). This confirms our earlier conjecture [14] that using CAD makes readers partially dependent on CAD prompts for their judgement about a feature. Similar interpretations have been proposed by [17] and [10]. As we have discussed elsewhere [14], this apparent dependence should not be seen as mere "complacency", an unfortunate term used in the "automation bias" literature [18]. This cognitive effect may even be an inevitable consequence of the combination of the task demands on readers and the high false positive rate for CAD prompts (leading to a high - though far from perfect - predictive validity for non-recall decisions from unprompted features). This effect may only be controllable by changes in the design of CAD tools.

False prompts increased the probability of readers erroneously classifying features as cancerous to a greater extent than true prompts increased the probability of readers correctly classifying cancer features (Table 3). This is an unintended and undesired effect, which however may be due in part to the desired effect of reducing detection failures. Readers with CAD probably do miss fewer features that they should examine, but some of these will be features on which they are liable to make false positive errors in classification.

More detailed, exploratory analyses via regression methods also highlight that these effects were greater on the readers who in the trial were less "effective" (in discriminating between cancer and non-cancer cases). These accounted for the majority (60%) of readers. There are also indications of some negative effects on the performance of the more effective readers, which deserve further investigation. These make up 10-20% of readers (Figure 1). This might support the hypothesis that, for these more effective readers, prompts are essentially a distraction. Also, the high number of false positive prompts may cause these readers to develop an instinctive rule whereby a prompt on a marginally suspicious feature tends to *decrease* their level of suspicion [1, 14].

It seems likely that these findings apply not only to the HTA study but also to other studies showing lack of impact from CAD.

Conclusions

The study reported here was motivated by the methodological consideration that assessing CAD by numbers of recalled cancer cases (*case-level* analysis) may be misleading. A case-level analysis does not measure how well the tools achieves its intended *direct* effects (to help

in the identification of individual features), and ignores potential cues for improvement. The large proportion of “false-target recalls” that we found in our study strongly supports this view.

Our results also strongly suggest that the apparent lack of impact reported for CAD in some studies is *not* caused by readers simply ignoring the prompts. These readers did react to prompts: CAD systematically affected readers’ identification of individual features, in a beneficial way for certain combinations of readers and features and in a damaging way for others (confirming results we previously obtained [14] at the level of patient recall decisions rather than individual features). Methodologically, this confirms that the effects of computer-based decision aids should be analysed by stratifying the sample of decisions appropriately (as proposed e.g. in [14]), since how CAD affects a reader will vary markedly between different patients or different mammographic features).

Further, our results support the conclusion that CAD performs its intended function for detection, that is, alerting readers’ to mammographic features they may otherwise fail to notice. But they also provide further support for the conjecture that readers use prompts in other unintended ways. In particular, readers apparently used the prompts to assess the level of malignancy of features they had already detected, using CAD as a diagnostic aid (in contrast to the manufacturers' recommendation). This has important implications for the design of these tools and user training. It may simply support the advice to read films carefully prior to viewing CAD prompts, but it may indicate that CAD *inevitably* changes a reader's decision process, so that algorithms must be calibrated to optimise these inescapable changes; or even that CAD may be inadvisable for certain readers [14] .

As noted earlier, some of these preliminary results must be treated with caution. For example, the calculated proportion of "false-target recalls" could have been different (smaller or indeed larger) had we chosen different criteria to determine whether a reader had correctly identified a cancer feature. Other potential limitations of these results are related to our analysing only "non-obvious cancers" from the HTA study, and using only one expert judge to determine where cancer features were located (although we believe this is unlikely to strongly affect the results), among other considerations. Since we have only analysed data produced in a single previous study, our results may be affected by the circumstances in which the original data were obtained (e.g. with a single model of CAD machine, well-experienced readers, all working in the U.K. and without experience in CAD). We hope to address some of these limitations in our planned future work, as described next.

The data set now available in electronic form is a mine of useful information for tracing the detailed mechanisms through which computer aid affected the readers in this experiment, and studying which parameters affected the desirable effect and the undesirable side-effects. We plan to continue our analyses in this direction. Future work we envisage includes: applying alternative criteria to identify “false-target recalls”; conducting further analyses to understand better the impact of false CAD prompts on readers' decisions; entering new data into the electronic database (e.g. detailed information about normal cases in the HTA study, and data from other studies, such as [9]).

A main concern is to test alternative conjectures explaining the differential effects observed, in particular, the *negative* effects of CAD, and to investigate whether these effects can be controlled by changes in the calibration of CAD tools (e.g. their sensitivity and specificity), or whether they require intervention on other aspects of CAD design and use.

Acknowledgements

The work described in this paper was supported in part by Cancer Research UK (Population & Behavioural Science Committee) via grant C22515/A7339, by the U.K. Engineering and Physical Sciences Research Council via project INDEED (EP/E000517/1), and by the European Union Framework Program 6 via the ReSIST Network of Excellence, contract IST-4-026764-NOE.

We would like to thank Paul Taylor for his help in making available the HTA data and for several discussions of this set of studies; Carla Nunes, who contributed to designing the data entry procedures and performed the data entry; and Vikas Jawa who provided the consistency-checking software for the database.

References

1. Astley SM, Gilbert FJ (2004) Computer-aided detection in mammography. Clin Radiol 59:390-399

2. Bazzocchi M, Mazzarella F, Del Frate C, Girometti R, Zuiani C (2007) CAD systems for mammography: a real opportunity? A review of the literature. Radiologia Medica 112:329-353

3. Freer TW, Ulissey MJ (2001) Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 220:781-786

4. Cupples TE, Cunningham JE, Reynolds JC (2005) Impact of Computer-Aided Detection in a Regional Screening Mammography Program. Am J Roentgenol 185:944-950

5. Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim C, Hardesty L, Poller WR, Shah R, Wallace L (2004) Changes in Breast Cancer Detection and Mammography Recall Rates After the Introduction of a Computer-Aided Detection System. J Natl Cancer Inst 96:185-190

6. Taylor PM, Champness J, Given-Wilson RM, Potts HWE, Johnston K (2004) An Evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. Brit J Radiol 77:21-27

7. Khoo LAL, Taylor P, Given-Wilson RM (2005) Computer-aided Detection in the United Kingdom National Breast Screening Programme: Prospective Study. Radiology 237:444-449

8. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, Berns EA, Cutter G, Hendrick RE, Barlow WE, Elmore JG (2007) Influence of Computer-Aided Detection on Performance of Screening Mammography. N Engl J Med 356:1399-1409

9. Alberdi E, Povyakalo AA, Strigini L, Ayton P (2004) Effects of incorrect CAD output on human decision making in mammography. Acad Radiol 11:909-918

10. Taplin SH, Rutter CM, Lehman CD (2006) Testing the Effect of Computer-Assisted Detection on Interpretive Performance in Screening Mammography. Am J Roentgenol 187:1475-1482

11. Houssami N, Given-Wilson R (2007) Incorporating new technologies into clinical practice without evidence of effectiveness in prospective studies: Computer-aided detection (CAD) in breast screening reinforces the need for better initial evaluation. Breast 16:219-221

12. Povyakalo AA, Alberdi E, Strigini L, Ayton P (2004) Evaluating 'Human + Advisory computer' system: A case study. In: Dearden A, Watts L (eds) HCI2004,18th British HCI Group Annual Conf. British HCI Group, Leeds, pp 93-96

13. Strigini L, Povyakalo AA, Alberdi E (2003) Human-machine diversity in the use of computerised advisory systems: a case study. In: 2003 Int. Conf. on Dependable Systems and Networks (DSN'03). IEEE, San Francisco

14. Alberdi E, Povyakalo AA, Strigini L, Ayton P, Hartswood M, Procter R, Slack R (2005) Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation. Br J Radiol 78:S31-40

15. Skitka LJ, Mosier K, Burdick MD (1999) Does automation bias decision making? International Journal of Human-Computer Studies 51:991-1006

16. Povyakalo AA, Alberdi E, Strigini L, Ayton P (2004) Evaluating 'Human + Advisory computer' systems: A case study. In: Watts ADaL (ed) HCI2004,18th British HCI Group Annual Conference. British HCI Group, Leeds, UK, pp 93-96

17. Hartswood M, Procter R, Rouncefield M, Slack R, Soutter J, Voss A (2003) 'Repairing' the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening. In: Schmidt KKEHKGFPDK (ed) Eighth European Conference on Computer Supported Cooperative Work (ECSCW 2003). Helsinki, Finland

18. Parasuraman R, Molloy R, Singh IL (1993) Performance consequences of automation-induced "complacency". International Journal of Aviation Psychology 3:1-23