Cambridge University Press
“Handbook of Medical Image Perception and Techniques”

Computer Aided Detection: Risks and benefits for radiologists' decisions

Eugenio Alberdi, Andrey A. Povyakalo, Lorenzo Strigini, Peter Ayton

1 Introduction

Computer aids for reading medical images bring obvious potential benefits, as discussed in previous chapters in this handbook. But incorporating them into medical practice presents also risks, which may be less obvious. This chapter discusses some of the things that can go wrong in the use of computer aids, as well as ways of assessing their benefits. Designing computer tools and planning their use needs to take into account how best to balance the potential for gain for certain patients and damage for others. This trade-off also adds complexity to the task of assessing such tools and their impact in medical practice.

Some chapters in this handbook have been written from the point of view of radiologists, or other medical practitioners. To complement this perspective, in this chapter we present the outcomes of interdisciplinary research work, which combined insights from a variety of disciplines: reliability engineering, computing, psychology, human factors and sociology (Alberdi et al., 2005).

We document and discuss these issues with reference to the use of computer support for the early detection of cancer in breast screening. In screening for breast cancer, expert clinicians (whom we will call “readers”) examine *mammograms* (sets of X-ray images of a woman's breasts), and decide whether the patient should be recalled for further tests because they suspect cancer.

For over ten years now, since 1998, Computer Aided Detection (CAD)[1] tools have been available to assist the interpretation of mammograms. CAD is designed to alert a reader (usually a radiologist) to areas of a mammogram. Typically the CAD tool processes a digitized 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 the readers to notice features in a mammogram that might indicate cancer but that they may otherwise miss. CAD is not meant to be a diagnostic tool, in the sense that it only marks areas of interest (possibly with additional indications about suspected type or degree of suspicion), which should be subsequently classified by the reader to reach a “recall/no recall” decision. The goal of using CAD is to increase 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).

[1] As is common in the radiological literature, we will also 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’.

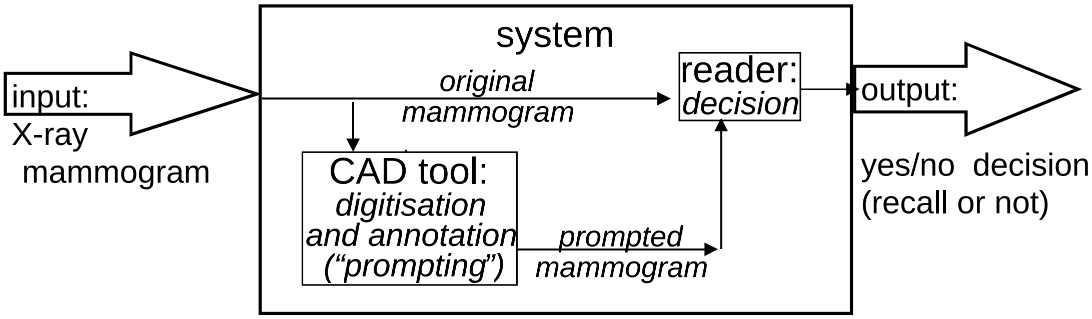

Figure 1. The decision system formed by a human "reader" with the Computer Aided Detection tool.

In a typical procedure for using CAD, the reader looks at a mammogram (on film or on a video screen) and interprets it as usual, then activates the tool and looks at a digitized image of the mammogram with the tool's prompts for regions of interest, checks whether he/she has overlooked any features with diagnostic value and then revises his/her original assessment, if appropriate. Fig. 1 schematically shows the "system" formed by the CAD tool and its human user.

Early evaluations of CAD, based on work by Warren Burhenne and co-workers (Warren Burhenne et al., 2000), showed the *potential* of this technology to help mammogram readers, and were used to support the approval of CAD by the U.S. Food and Drug Administration (FDA) (FDA, 1998). Subsequent studies have shown not only CAD's potential to help but also factual evidence of significant increases in cancer detection by radiologists when using the tool (Freer and Ulissey, 2001, Cupples et al., 2005, Gromet, 2008). However other studies have shown negligible effects of CAD on reader performance (Gur et al., 2004, Taylor et al., 2004b), even inferior to the benefits of double reading (Khoo et al., 2005), that is, separate reading of a patient's mammograms by two human readers. Furthermore, there is evidence of harmful effects of CAD use in certain situations, which manifest themselves either as a decrease in the specificity of CAD-supported readers (Fenton et al., 2007) or as readers' reduced ability to detect cancers if these are not prompted by CAD (Alberdi et al., 2004a, Taplin et al., 2006, Zheng et al., 2004).

This research naturally calls for a reassessment of the circumstances under which CAD prompts will actually improve readers' decisions. The first requirement for the effectiveness of CAD is of course accuracy of the prompts. It seems clear that, to be of any use, the prompts must be correct "often enough", although no CAD tool can identify all cancers and avoid all false prompts. However, we can expect any approved CAD system to be quite accurate. Ease of use, readability, etc. are also obvious requirements. But research suggests that less obvious aspects of CAD performance are also important. First, there is the issue of whether the CAD tools' help is focused on the real needs of the readers: we will refer to this as the issue of *diversity* between the CAD tools and the readers, because a CAD tool that only offered correct advice in cases that readers would process correctly anyway would not actually reduce reader errors. Secondly, there are changes in the way readers process information when they are assisted by CAD prompts. The simplest change to envisage is that any user of a computer support system may over time become too reliant on the computer, blindly following its advice, even when it is wrong. One may be tempted to dismiss these problems as only requiring better discipline by the users. However, in reality, there is a family of possible related changes of reader behaviour, and some may well be inevitable: rather than trying to avoid them by better discipline, it may be easier or even necessary to tweak the design of CAD to achieve the best overall results given these inevitable changes. We will refer to this set of problems as "automation bias".

This chapter describes studies and analyses that could help us understand some of the disparate results obtained by studies of CAD effectiveness, with pointers to relevant work by the computer engineering and human factors communities. In the last few decades, a growing body of research has shown potential pitfalls of using automated tools to support human decisions. We review, in section 2, some relevant literature on the decision biases associated with using automated decision support tools ("automation bias"). We will also provide evidence (from our previous work and from work by other authors) of automation bias in the use of CAD for mammography (section 3). This will be followed by explanations (conjectured by us and other authors) to account for these 'automation bias' phenomena (section 4). Section 5 discusses some implications of the work described in this paper for the evaluation and design of CAD. Section 6 contains concluding remarks.

2 A Brief Overview of the Literature on Automation Bias

Literature on human factors in engineering and computing has long recognized that even well-designed automation may fail to deliver the improvements it promised, due to initially unexpected effects on people. For instance, (Bainbridge, 1983) summarized known "ironies of automation" in industrial control: automated systems that were meant to substitute human operators and deliver more accurate, error-free performance actually relegated operators to narrower but more critical tasks (monitoring of the automated process and emergency intervention), while depriving them of the hands-on practice that these more critical tasks required. So, the task originally performed by a person alone was now performed by a "human-machine system" in which, however, the person's performance was made worse by how automation reshaped human tasks.

A partial solution to this has been seen in automation that assisted human experts rather than supplanting them; and this category of computer systems is actually widespread in "decision support" applications. However, here too there are concerns. We will refer to this set of related concerns by the label "automation bias", used by some researchers during the last decade.

As early as 1985, Sorkin & Woods (Sorkin and Woods, 1985) using signal detection theory concluded that in a monitoring task performed by a human-machine system (a person assisted by a computer), optimizing the computer's performance as though it had to perform the task by itself would not always optimize the performance of the human-computer system. A human-machine system can only be optimized or improved as a whole: improving the automation alone or human training alone may be ineffective or wasteful.

"Automation bias" refers to those situations where a human operator makes more errors when being assisted by a computerized device than when performing the same task without computer assistance (Mosier et al., 1998, Skitka et al., 1999). Similar or related concepts are automation-induced “complacency” (Azar, 1998, Singh et al., 1993, Wiener, 1981) and “overreliance” on automation (Parasuraman and Riley, 1997).

Studies of automation bias have been predominantly conducted in the domain of aviation (and only rarely on medical domains) and typically investigate the behaviour of pilots or air traffic controllers assisted by computerized decision aids. Although, at first, the activities conducted by aviation professionals may appear very different from the activities involved in breast screening, they arguably have much in common. They are both "signal detection" tasks (McNicol 1972), in which the human user is faced with the task of detecting some crucial perceptual pattern (a target) in a complex display. Both missing some targets (false negative errors) and identifying a target where none is present (false positive error) carry costs. For example, one aviation task is to identify, in a complex radar image, any airplanes that present dangers of collision. In the case of breast screening, a reader must interpret the appearances in an X-ray to determine the presence of cancer. Automation, in the form of computerized advisory tools, is introduced in both domains to prompt the human operators to help them detect their targets.

In an influential paper, Parasuraman & Riley (1997) discuss ways in which this kind of human-computer interaction can go wrong, citing both anecdotal evidence and results from various empirical studies. They talk about three aspects of ineffective human use of automation: *disuse*, i.e., underutilization of automation, where humans ignore automated warning signals; *misuse*, i.e., over-reliance on automation, where humans are more likely to rely on computer advice (even if wrong) than on their own judgment; and *abuse*, when functions are automated "without due regard to the consequences for human [...] performance and the operator’s authority " (Parasuraman and Riley, 1997).

The phrase “automation bias” was introduced by (Mosier et al., 1998) when studying the behaviour of pilots in a simulated flight. They divided inappropriate responses to automated prompts into *errors of commission* and *errors of omission*. An *error of commission* occurs when a decision-maker follows automated advice even when the automated aid is wrong. In the case of CAD for mammography, an error of commission would be to recall a patient for a mammogram appearance that does not indicate cancer but which CAD has incorrectly prompted. On the other hand, an *error of omission* occurs when a decision maker, given evidence of a problem, fails to react appropriately because the automated tool did not highlight the problem. In the case of mammography, a reader’s error of omission would be failing to recall a patient because CAD failed to prompt the area of the mammogram where a cancer is located.

The finding that automation may bias pilot decisions was subsequently replicated in studies using non-pilot samples, namely, student participants in laboratory settings simulating aviation monitoring tasks (Skitka et al., 1999). Essentially these studies showed that, when the automated tool behaved reliably, participants using it made more correct responses than those without the tool's aid. However, if the automation was imperfect (i.e. unreliable for some of the tasks), participants using it were more likely to make errors than those who performed the same tasks without automated advice. For the monitoring tasks that Skitka and colleagues used in their studies, the participants had access to other (non automated) sources of information. In the automated condition they were informed that the automated tool was not completely reliable but all other instruments were 100% reliable. Still, many chose to follow the advice of the automated tool even when it was wrong and was contradicted by the other sources of information. The authors concluded that these participants had been biased by automation and interpreted their errors (especially their errors of omission) as a result of complacency or reduction in vigilance.

Individual differences seem to play an important role in human reactions to automation (Dzindolet et al., 2003, Skitka et al., 1999, Skitka et al., 2000). One would expect that more experienced people would tend to be less susceptible to bias from automation, when performing tasks in their area of expertise. However, as highlighted above, automation bias effects have been reported for both laymen (e.g. student participants) and experts (e.g. air pilots and clinicians). In fact, Galletta and colleagues (Galletta et al., 2005) have shown (in a verbal skills domain: the use of spelling and grammar checking software) that more skilled and experienced individuals were more likely to be damaged by certain types of computer errors (i.e., false negatives) than were the less skilled individuals.

Some researchers have discussed human reliance on automation in terms of "trust" (Muir, 1987, Dzindolet et al., 2003, Bisantz and Seong, 2001, Dassonville et al., 1996, Lee and Moray, 1994, Lee and See, 2003, Muir, 1994, Singh et al., 1993, Tan and Lewandowsky, 1996, Muir and Moray, 1996). The idea is that the more human operators trust an automated aid the more likely they are to rely on the aid. So, it is important that this trust be well placed. However if a human trusts (and thus relies on) an unreliable tool, then automation bias may occur (or misuse of automation as defined above). Similarly if a person does not trust a highly reliable tool, the person may end up ‘disusing’ (as defined above) or under-using the tool, hence not obtaining the full potential benefits of automation.

Subjective measures of the trust of human operators in a computer tool have been found to be highly predictive of people’s frequency of use of the tool (de Vries et al., 2003, Dzindolet et al., 2003). Use of automation (or reliance on automation in its generic sense) is usually assessed with observations of the proportion of times during which a device is used or by assessing the probability that operators will detect automation failures (Meyer, 2001).

A concept related to “trust” in automation is that of the “credibility” or “believability” of automation (Tseng and Fogg, 1999). There are indications that people tend to perceive computers as infallible, and may put excessive trust in them (Fogg and Hsiang, 1999, Martin, 1993). Indeed, (Dzindolet et al., 2003) have shown empirically that people have an inclination to trust and rely on an automated aid regardless of its reliability. However, there is also evidence of a pattern of behaviour in which, as soon as humans become aware of the errors made by a computer tool, their trust in the tool, and their subsequent reliance on it, decrease sharply (de Vries et al., 2003, Dzindolet et al., 2003, Dzindolet et al., 2002). This reduction in trust, in turn, can be attenuated by increasing people’s understanding of the computer errors. (Dzindolet et al., 2003) found that people who were told the reasons for the aid’s errors were more likely to trust it and follow its advice than those who were not aware of these reasons.

3 Automation Bias in CAD for Mammography

The effectiveness of CAD in assisting mammogram readers in breast screening has been the subject of many studies, with variable results between different settings, and some controversy. See reviews in (Astley and Gilbert, 2004) and (Bazzochi et al., 2007). Here we present evidence of automation bias in the use of CAD, first reporting on a case study of the effects of CAD on the decisions of mammogram readers.

The case study described here was a follow-up to a previous retrospective controlled study of CAD conducted by other researchers for the UK Health and Technology Assessment (HTA) program, on a specific CAD tool: “Study 1” in (Taylor et al., 2004b), hereon “the HTA study”. This study is described in some detail in section 3.1. Our team subjected the raw data from the HTA study to various exploratory statistical analyses, and focused on the interactions between correctness of computer output, difficulty of individual decision problems and reader skills (section 3.2). We conducted some subsequent experiments to study the effects of incorrect computer output on decisions (section 3.3). We also report the results of more fine-grained analyses, which included previously unexamined details about reader decisions and CAD effects at the level of individual features in mammograms (3.4). Finally, we briefly summarize findings from other researchers that report similar automation bias in the use of CAD for mammography (3.5).

3.1 HTA Study

The goal of the HTA study was to assess the impact of a commercial CAD tool, the then market leader in mammography (R2 ImageChecker M1000) (FDA, 1998). This CAD tool had been shown to have a high sensitivity, that is, to prompt a high proportion of breast cancers appearances. An overall sensitivity of up to 90% had been reported (Castellino et al., 2000). However, high sensitivity comes at the expense of low specificity, that is, this tool generated many false prompts, with an average of 2.06 prompts per case in one study.

The HTA study was run with 50 readers, experienced in breast screening, who examined 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: 1) with CAD; and 2) without CAD. The order of conditions was randomized 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 digitized version of the mammograms printed on paper. In the “with CAD” condition, the printouts contained the prompts generated by CAD. Participants were asked to make their decisions as to whether a case should be recalled for further tests as though they were viewing the mammograms as single readers in the U.K. breast screening program. More details of the procedures can be found in (Taylor et al., 2004b).

Analysis of the results showed no statistically significant impact of CAD (no improvement and no reduction) on the readers’ sensitivity and specificity (Taylor et al., 2004b).

3.2 Interactions between correctness of computer output, difficulty of cases and human skills

The authors of the HTA study kindly granted us permission to analyse their data in our investigation of the fault-tolerant characteristics of human-machine systems (i.e., the ability of computers and their users to compensate, to some extent, for each other's errors), performed under the Interdisciplinary Research Collaboration on Dependability, "DIRC", funded by the U.K. Engineering and Physical Sciences Research Council. Although conventional statistical analyses (e.g., ANOVA) of the data from the HTA study showed no significant impact of CAD on average, our subsequent exploratory analyses indicated that CAD was actually affecting readers' decision making in systematic ways (Alberdi et al., 2005).

These supplementary exploratory analyses were informed by probabilistic modeling of the system formed by the CAD machine and the reader (Strigini et al., 2003), and motivated by previous work on computer systems with diverse redundancy (Littlewood et al., 2002): systems in which the risk of erroneous results is reduced by having two or more different programs to calculate these results, and comparing their outputs foe consistency.

Probabilistic modeling highlighted how variation and co-variation in the "difficulty" of input cases (patient mammograms) for CAD and for the reader substantially affect the accuracy of the overall system (i.e. of the final decision about a patient). For example, it shows that it may be misleading to focus on the average probabilities of the failures of either the reader or CAD: reducing either or both of these may not substantially improve the statistics of the accuracy of decisions made. What matters more is how often the CAD tool fails on the same case on which the unaided reader would also fail, so that CAD brings no improvement; and whether, and how often, erroneous behaviour by the CAD tool may negatively affect the reader's own decision.

Our analyses focused on the different effects of correct vs. incorrect CAD output, evaluating their interaction with *case difficulty*. The difficulty of a case was defined as the fraction of readers in the study who reached a wrong decision about that case. The correctness of CAD prompting and of human decisions was defined as follows. CAD output is considered to be 'incorrect':

- for a cancer, if CAD 'misses' the cancer (i.e. it gives a 'false negative', FN), which it can do in two ways: a) by failing to place any prompt on the mammogram, or b) by placing prompts in areas other than the actual location of the cancer (all prompts are 'false'). CAD is said to provide '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, even if it also prompts other areas of the mammogram.
- for a normal case, if CAD places any prompt on the mammogram ('incorrectly marked' normal case; false positive, FP). CAD is said to provide 'correct output' for a normal case if it places no prompts ('unmarked' normal; a 'true negative').

When referring to readers' decisions, 'error by a reader' (or 'incorrect decision') means that she/he recalls a normal case (reader's FP) or does not recall a cancer (reader's FN); 'correct decision' means that she/he does not recall a normal case or recalls a cancer.

In this chapter we focus on evidence concerning CAD effects on the frequency of FN failures by readers. This will be sufficient to illustrate the possibility of “automation bias” phenomena. In breast screening, a false negative is a more severe failure than a false positive, although false recalls and the unnecessary investigations (radiology, biopsy) associated with them can contribute to serious psychological and physical damage to healthy patients as well as impose high costs on a medical organization.

The supplementary analyses of the HTA results (Alberdi et al., 2005) indicated that:

- the correctness of the computer output affected the decisions of the readers. In particular, correct computer prompting was likely to help readers in reaching a correct decision while incorrect prompting hindered their decision making;
- reader and computer errors were strongly correlated; CAD tended to a) prompt correctly those cancers which readers were also likely to recall without CAD; and b) process incorrectly most of the cancer cases that readers would also miss unaided; consequently, for those cases that were more difficult for humans to interpret, the computer was less likely to give useful support.

Additionally, we used logistic non-linear regression (a form of interpolation on the raw data) to look for possible general patterns in the effect of CAD. This showed that CAD tended to make cancers that were relatively "easy" (low "difficulty") "easier", and cases which were relatively difficult even more difficult (Povyakalo et al., 2004).

An abstract representation of these effects is shown in Figure 2. The horizontal axis represents case difficulty without computer aid; the vertical axis shows the impact of CAD, as the difference between difficulty of a case when seen with CAD and its difficulty without CAD. So, a point below 0 on the y-axis indicates a cancer case for which CAD appears to reduce the rate of FN errors. In other words, readers detect these cancers more often when using CAD than when not using CAD. The curves show the regression estimates for the mean value of the impact of CAD as a function of the difficulty of cases. The dashed (red) curve is obtained from the data for the incorrectly prompted cancers, the (blue) dotted-dashed curve to the correctly prompted cancers and the solid curve to all cancers together.

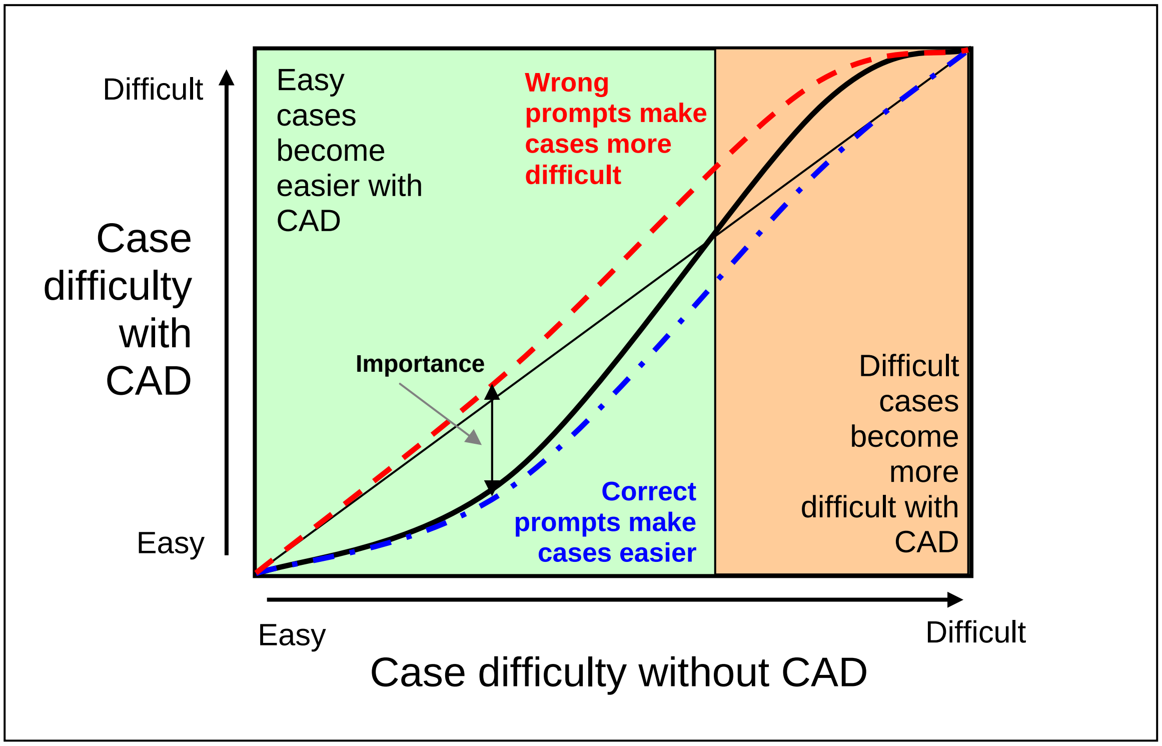
Difficult
Easy
Case difficulty with CAD
Easy cases become easier with CAD
Wrong prompts make cases more difficult
Importance
Correct prompts make cases easier
Difficult cases become more difficult with CAD
Easy
Difficult
Case difficulty without CAD

Figure 2. Impact of CAD in the HTA results, as a function of “case difficulty”

Figure 2 shows that, for “difficult” cases, incorrect decisions were more common with computer aid than without. Since less sensitive readers were very unlikely to recognize “difficult” cancers anyway, this increase in FN errors for “difficult” cases must be due to an adverse effect of the computer aid on the decisions of the more sensitive readers, plausibly caused by incorrect computer prompts (cf Fig 2, again). Similarly, for the “easy” cases, without computer aid the less sensitive readers made a larger number of incorrect decisions: they were thus more likely to benefit from the correct computer prompts on “easy” cases than on “difficult” ones, and on easy cases they were more likely to benefit than the more sensitive readers.

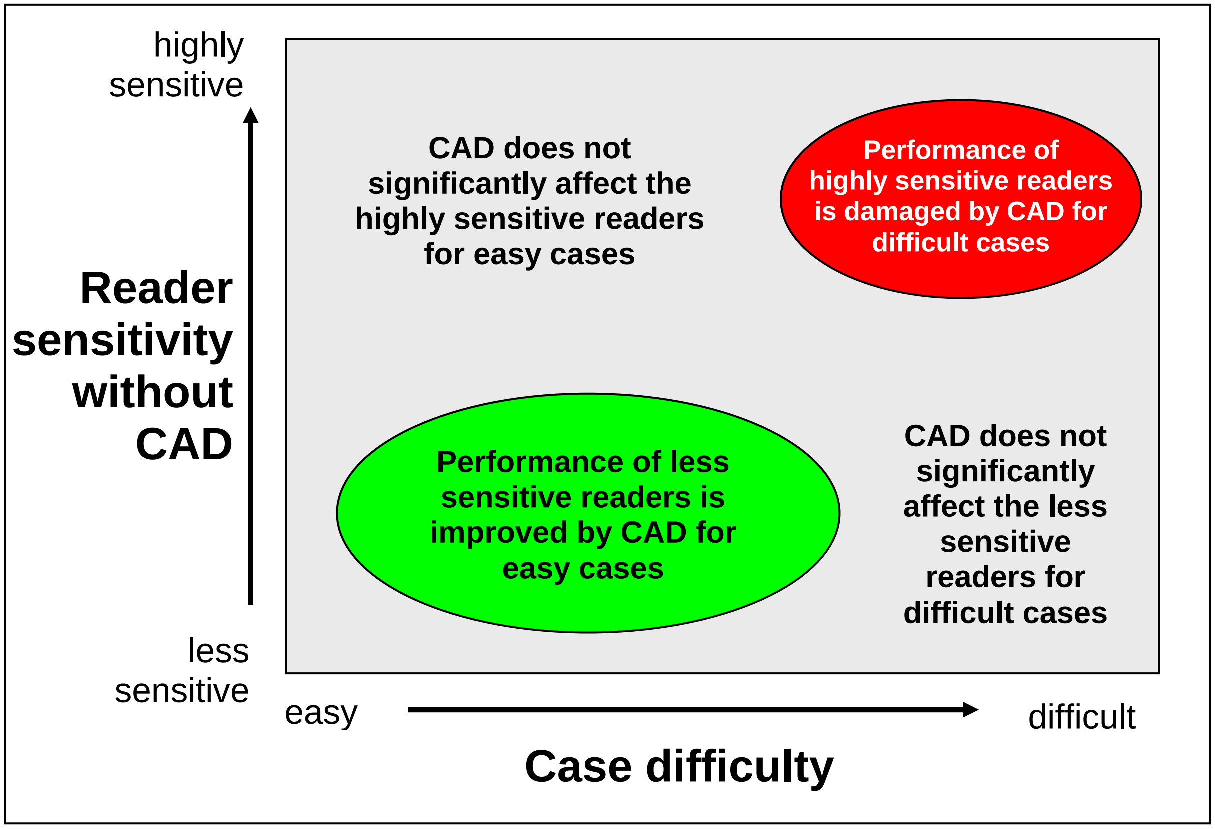
highly
sensitive
Reader sensitivity without CAD
less
sensitive
CAD does not significantly affect the highly sensitive readers for easy cases
Performance of highly sensitive readers is damaged by CAD for difficult cases
Performance of less sensitive readers is improved by CAD for easy cases
CAD does not significantly affect the less sensitive readers for difficult cases
easy
difficult
Case difficulty

Figure 3. The pattern of effects of CAD suggested by the exploratory statistical analyses of the data from the HTA study (schematic representation). Whether the average effect of CAD is positive or negative would depend on the frequencies of the different groups of cases and of readers.

Thus, the use of computer support is likely to be more beneficial for less sensitive readers and more questionable for the more sensitive ones, especially when these are dealing with the more "difficult" cases. This conjecture is supported by additional (logistic regression) analyses of these data (Povyakalo et al., 2006). Figure 3 shows an abstract representation of the effects of the interaction between case difficulty and unaided readers' sensitivity as estimated by those analyses.

3.3 Empirical study of the effects of false negative prompting

We summarize here the results of two follow up studies, investigating in more detail the effects of false negatives from CAD on human decision making (Alberdi et al., 2004a).

Study 1 used a set of mammograms containing a large proportion of cancers that CAD had missed. All other characteristics of the test set were kept as similar as possible to the sets used in the HTA study, to make sure readers perceived this study as a natural extension of the HTA study (all 20 readers in Study 1 had also participated in the HTA study) and to behave in a comparable way. Study 1 used essentially the same procedures used in the HTA study, except that the readers in Study 1 saw all the cases only once: always with the benefit of CAD. Human and CAD errors were defined as in the previous section.

Our original goal for study 1 was to collect empirical data to estimate the probability of reader error given a FN error by the CAD tool, rather than to compare the performance of readers with and without CAD. However, the results were unexpected and intriguing: the proportion of correct decisions was very low, and particularly so for cancer cases not prompted by the tool. This suggested that CAD errors may have had a significant negative impact on readers' decisions. To explore this issue further, we designed Study 2, in which a new but equivalent set of readers read the same mammograms without computer aid. Study 2 used the same procedure as in Study 1, except that readers did not see the computer prompts.

Reader sensitivity in Study 2 (without CAD) was significantly higher (73%) than in Study 1 (with CAD: 61%). The measured difference between the with-CAD and without-CAD conditions was more acute for cancer cases that CAD had processed incorrectly and, crucially, even more so for cancer cases where the tool had placed no prompts at all (54% sensitivity without CAD vs. 33% with CAD). See (Alberdi et al., 2004a) for details.

These findings strongly suggest that, at least for some categories of cases, FN errors by the CAD tool (especially absence of prompts on cancer cases) had a significant detrimental effect on human decisions. Similar damaging effects of CAD have been reported by (Zheng et al., 2004) and (Taplin et al., 2006) (*cf* section 3.6).

3.4 Analysis at the level of individual mammographic features

CAD tools are designed to help in the identification of individual features in a mammogram. However, medical studies most frequently measure CAD's effectiveness in terms of readers' recall of a case seen as a whole. This makes sense for the purpose of determining whether a particular tool helps or not, but not to understand whether its effect on the reader's work is as intended by the designers. Consistently with most studies, and partly for practical reasons, our analyses of the HTA study (and the follow-up experiments, cf section 3.3) also used data about reader decisions collected at the level of *cases* (recall or no-recall decisions), rather than at the level of mammographic *features* (which features the reader judged to be suspicious enough to warrant a recall). Reader decisions were classified simply as recall/no recall decisions, 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 example, as defined earlier, a reader's decision was classified as correct if it entailed recalling a patient who had cancer. However, this classification did not take into account whether the reader had actually detected the area in the mammogram where the cancer was located. To illustrate how an apparently correct recall may be due to the wrong process, we can imagine a mammogram where the cancer features were 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 judging as suspicious some non-cancer area. Although this was seemingly a correct decision, it was, in fact, a recall prompted by the "wrong" mammographic appearance. In other words, the decision appears to be correct at the *case level* because a patient with cancer has been recalled. However, it is a clearly incorrect decision at the *feature level*: the specific area that has triggered the recall contains no cancer.

We outline in this section the results of a study (Alberdi et al., 2008) funded by Cancer Research U.K., whose goal was precisely to conduct detailed feature-level analyses of previously unprocessed data from the HTA study.

Information about the specific features used by readers to make their decisions was available in the mammographic reports that readers filled in during the HTA study. However it had not been incorporated into the electronic databases used in the earlier analyses: these only contained information about readers' recall decisions, malignancy of the case and correctness of CAD prompting. In this new study (Alberdi et al., 2008), the previously unexamined, detailed information from several thousands mammography reports from the HTA study (5,839 mammography reports produced by 50 readers for 59 cancer cases in the previous study) was entered into a new electronic database, including the specific mammographic features that had been identified by the readers. This information was used to elucidate how judgements about features were affected by CAD.

The new electronic database contained information about all the different mammographic features (areas of interest) which had one or more of these characteristics: a) identified by an expert radiologist as cancer regions, *a posteriori*, based on biopsy reports; b) marked by CAD ("prompted") with a prompt; and c) marked by a reader on the printout of the mammogram in the mammography report during the HTA study. Henceforth we will refer to these areas (features) as Marked Areas of Interest (MAOIs).

Additionally, the new database included other information that readers entered in each mammography report, namely, the degree of suspicion and type of abnormality (mass, calcification, asymmetry) the reader assigned to each MAOI s/he marked, plus other possible notes, and a recall/no recall decision about the whole case.

Using this newly extracted information, the recalls *of cancer cases* were carefully classified into "true-target", if the reader identified at least one cancer feature as suspicious (or at least as the most suspicious in the mammogram), and "false-target" otherwise. About 13.5% of recall decisions were found to be "false-target recalls", that is, they were due to non cancerous features. This pattern was independent of whether the readers were using CAD or not.

In other analyses, we investigated how readers reacted to features prompted by CAD and how the tool affected readers' identification of cancer features. These analyses showed 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 (Taylor et al., 2004b), is also present, *on average*, at the level of individual cancer features. This is the result of significant effects (in opposite directions) on the reader by individual prompts. More specifically, in comparing the decisions made with CAD against those made without, we found that:

- for those lesions that CAD prompted, readers were more likely to examine them when using CAD (i.e., seeing the prompts) than when not using CAD (a desirable effect of CAD);
- for those lesions that CAD prompted, readers were more likely to consider them malignant when using CAD than when not using CAD; in other words, many lesions that readers *did* consider in their decisions even when looking at a case without CAD were treated as *more suspicious* when the case was seen with CAD; this suggests that CAD prompts affect readers' *interpretation* of features and were not used merely as attention cues but also as *diagnostic* cues, going beyond the scope of computer aided *detection*;

- for those cancer features that CAD did not prompt, readers were less likely to detect them, or to consider them malignant, when using CAD than when not using CAD (an undesirable effect of CAD);
- false prompts made non-cancer features more likely to be classified as cancer by readers when using CAD than when not using CAD; this effect outweighed that of true prompts increasing the probability of readers correctly classifying cancer features.

Additionally, similarly to what we described in 3.2, exploratory (logistic non-linear) regression analyses were conducted using as factors:

- readers' "*effectiveness*", that is, readers' ability to discriminate between cancers and normal cases, calculated as the difference between a reader's measured sensitivity and 1 minus the reader's measured specificity
- the "*difficulty*" of the features, that is, the fraction of all readers that failed to detect the feature.

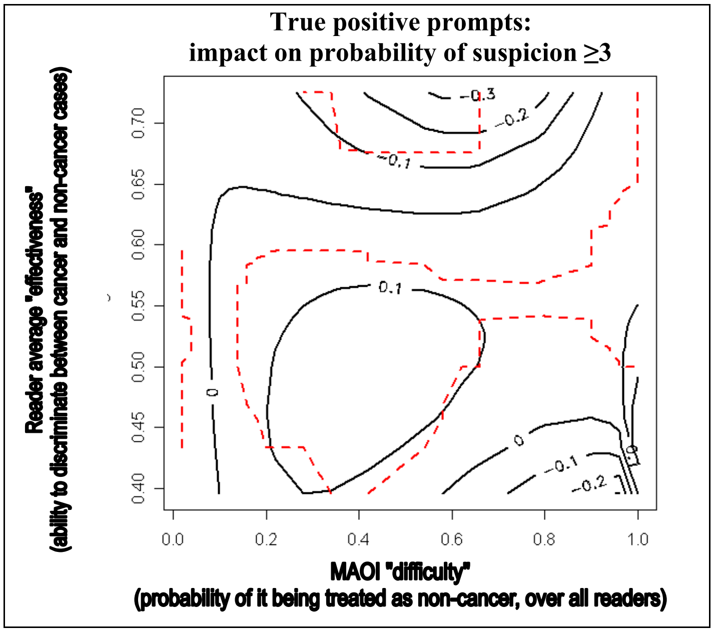

Figure 4. "Feature-level" study of the HTA results: exploratory regression analysis of readers' reactions to prompts

Figure 4 shows the results of the regression analysis. The areas marked in the figure represent different degrees of estimated CAD impact. The dashed lines bound areas where the effect is statistically significant. For example, for readers of "effectiveness" 0.50, the estimated CAD impact is between 0.1 and 0.2 when judging correctly prompted cancer MAOIs of 40% difficulty, that is, they fall within the big closed area in the centre. Since this closed area is mostly within a dashed outline, this estimated impact is statistically significant. The estimates suggest that a CAD prompt significantly reduced (by up to 10%), compared to what it was without CAD, 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. Again, this supports, at the level of features, the earlier finding that prompting may damage the decisions of more skilled readers for some categories of patients. See (Alberdi et al., 2008) for more details.

3.5 Further evidence of automation bias in the use of CAD

In this section we present evidence collected by other researchers that supports some of the findings reported so far.

Zheng and colleagues (Zheng et al., 2001, Zheng et al., 2004) conducted a retrospective laboratory study of 7 radiologists interpreting digitized mammograms in 5 different conditions. In one condition, the radiologists saw no prompts, whereas the other four conditions were created by manipulating the images, by adding CAD prompts to produce different false negative (FN) and false positive (FP) rates of CAD prompting. These authors found that a reduction of prompting sensitivity and specificity significantly increased the radiologists' false negative rates in unprompted areas. Their conclusion was that CAD tools have the potential to significantly improve diagnostic performance in mammography, but poor combinations of CAD's FN and FP rates could adversely affect radiologists' performance, for both prompted and unprompted areas.

Similar results were found by (Taplin et al., 2006) in a retrospective study comparing the sensitivity of 19 radiologists who read a set of 341 mammograms twice, with and without CAD. They found that visible mass cancer lesions that had not been prompted by CAD were less likely to be considered as cancer by radiologists than they were when CAD was not used. Use of CAD was associated with: a) improved radiologist sensitivity for visible cancers prompted by CAD; and b) decreased reader sensitivity for unprompted visible cancers.

A different type of bias has been reported by (Fenton et al., 2007), in one of the largest studies to date, which reviewed decisions, with and without CAD, by a large population of community radiologists (in 43 screening centres) for over 200,000 patients in the course of 4 years. The authors report negligible improvement in cancer detection for readers using CAD and significantly reduced overall accuracy, assessed by use of receiver-operating-characteristic (ROC) curves. CAD-supported readers exhibited reduced specificity; recall and biopsy rates were significantly increased without an associated rise in sensitivity.

4 Possible Explanations for Automation Bias in CAD use

Here we look at cognitive processes or strategies that may be underlying the examples of apparent automation bias described so far. We present these mechanisms in the context of CAD for mammography, but we believe they are relevant for other computer assisted decision environments.

4.1 Effects of CAD's False Negatives

The results of our analyses and empirical studies summarized earlier in this chapter suggest that incorrect outputs of the CAD tools may have biased the decision making of at least some of the readers for some categories of cases. In particular, the results indicate that, for difficult cancers, FN errors in prompting often led to incorrect reader decisions.

Using Skitka et al.’s (Skitka et al., 1999, Skitka et al., 2000) terminology introduced earlier, these incorrect decisions could be classified as “errors of omission”: readers interpreted absence of prompts on (an area of) a mammogram as an indication of “no cancer” and therefore failed to take appropriate action (i.e., they failed to recall the patient). There are, however, important differences between Skitka et al.’s tasks and mammogram reading as investigated in the HTA study and our follow-up studies. In the former, the participants seemed to use the computer output to replace calculations they could perform otherwise by using alternative, highly reliable, non-computer mediated indicators. In contrast, uncertainty in mammogram reading is greater. The readers did not have access to any sources of information other than the X-ray films (mammograms) and the output of the CAD tool on digitized versions of the films. It is important to remember that CAD prompts are designed merely as attention cues – and not to replace other sources of information.

With CAD, over-reliance on automation could manifest itself in readers using prompts instead of thoroughly examining the mammograms, that is, readers becoming complacent, or less vigilant, when using automation – as hypothesized by Skitka’s team and others to explain errors of omission (Skitka et al., 1999, Skitka et al., 2000, Parasuraman and Riley, 1997, Parasuraman and Miller, 2004, Meyer, 2001, Meyer and Bitan, 2002, Meyer et al., 2003, Meyer, 2004).

One could argue that, based on their past experience with the tool, readers tended to assume that the absence of prompts was a strong indication that a case was normal. CAD generates many false positives, which readers claim to find distracting. As a result, the absence of prompts in a mammogram may be seen as more informative than their presence, so readers may have become complacent, paying less attention than necessary to those mammograms that had no computer prompts.

A problem with the term “complacency” (and similar accounts of automation bias) is that they imply value judgments on the human experts. As Moray (Moray, 2003) has recently pointed out, the claim that automation fosters complacency implies that operators are at fault, when the problem often lies in the characteristics of the automated tools.

An alternative (perhaps complementary) way of accounting for the association between incorrect prompting and reader errors contemplates how readers deal with “indeterminate” cases - that is, cases where they have detected anomalies with unclear diagnosis, so that they are uncertain as to whether the cases should be recalled. For example, a reader perhaps noticed an ambiguous abnormality without using CAD; then, when looking at the CAD output, the reader revised her/his decisions for those features that he/she had already detected. In other words, they may have used the absence of prompts as reassurance for a "no recall" decision when dealing with features they found difficult to interpret. Conceivably readers were using any available evidence to resolve uncertainty. The implication is that the CAD tool was being used not only as a detection aid but also as a classification or diagnostic aid – specifically *not* what the tool was designed for. Readers’ use of CAD in such a way has been reported before for the CAD tool investigated here (Hartswood et al., 2003) and for other similar CAD tools (Hartswood et al., 2000) The following transcript is an example of a reader’s use of prompts to inform her/his decisions: “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.” (Hartswood et al., 2003) In other instances, readers were observed using the absence of a prompt as evidence for ‘no recall’ (Hartswood et al., 2003).

Arguably, this use of CAD violates an explicit warning in the device labeling (1998): "...a user should not be dissuaded from working up a finding if the device fails to mark that site". However, such warnings could only be useful if readers were aware they exploited prompts in this fashion.

An alternative conjecture about "indeterminate cases" posits that CAD may alter readers’ decision thresholds. The conjecture is as follows: when seeing certain "indeterminate cases" without computer support, readers are likely to recall it for further investigation. However, with computer support, they know that supplementary information will be available in the form of CAD prompts; therefore, their preliminary decisions (before checking the prompts) are more likely to be "no recall" than they would be without CAD. A similar point was previously raised about other studies (Astley and Gilbert, 2004, Taylor et al., 2004a). In practice, this makes the presence of computer outputs a determinant for the "recall/no recall" decision, despite readers *never* changing a preliminary decision from "recall" to "no recall" based on absence of prompts, or being in violation of the prescribed procedures. For those indeterminate cases that are "difficult" cancers, CAD is likely to produce incorrect output, substantially reducing the chances of the case being recalled. Readers may not change any individual decisions because of absence of prompts, but they may change their decision thresholds for some cases due to the very presence of computer support. It may be very difficult (or even impossible) to prevent this by simple prescription.

4.2 Effects of false prompts

In this section we focus on observed and hypothesized effects of false prompts on reader decisions. As noted earlier, CAD tools show high sensitivity at the expense of specificity. A large proportion of the prompts generated by CAD are false. An expected outcome of the low specificity of CAD is that it will affect the specificity of its readers, predictably making readers less specific, as reported by (Fenton et al., 2007, Astley, 2005). However the opposite effects have been reported as well: some studies show that readers are more specific when using CAD than when not using it (Alberdi et al., 2004a, Taplin et al., 2006).

Less predictable, perhaps, is the finding that a high FP rate from CAD also reduces readers’ sensitivity. We have already noted readers’ tendency to over-rely on the absence of prompts as an indication of the absence of cancer. This phenomenon could be partly explained by an excess of false prompts. When a large proportion of the prompts are false, the absence of prompts becomes more informative than their presence.

Readers also report frequently that they find many of the CAD prompts distracting and confusing, possibly because the false prompts cause attentional overload. This distracting effect can manifest itself in different ways. For example, false prompts may distract readers’ attention to unimportant areas of the mammogram, making them spend more time than due on normal appearances and leaving them with less time to exhaust their search for true cancer areas, especially if these have not been prompted by CAD.

Another related but different effect is that the presence of ‘obvious’ false prompts may affect how readers interpret true prompts (Astley, 2005). Let us imagine that a reader is assessing a cancer area which is difficult to interpret but which has been correctly prompted by CAD. If the area and the true prompt are surrounded many false prompts that are easily recognizable as spurious, the value of the true prompt may get diminished in the eyes of the reader. As a result, an area that a reader may have considered as cancer without CAD may be dismissed by the same reader when seeing it with CAD, even if the tool has prompted the area correctly. Analyses of readers’ decisions at the level of features have provided evidence of many instances in which a reader performs worse with CAD than without for cancer appearances that CAD has prompted correctly. Similar effects have also been reported by (Cupples et al., 2005).

5 Some implications for the Evaluation and Design of CAD

5.1 Methodological Implications for Assessment of CAD

Some of the findings reported earlier in this chapter have obvious methodological implications for the evaluation and study of CAD use. We focus here on two limitations that arguably affect many evaluation studies.

5.1.1 Limitations of studying the impact of CAD at the level of case (vs. feature)

We have noted how, when looking at fine-grained analyses of CAD evaluation data Alberdi, 2008) “false-target recalls” (i.e., recalls of cancer cases without correct identification of the cancerous features) appeared to be rather frequent (about 13.5%). 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 – but the evidence that these situations can occur so frequently has important methodological implications for the evaluation of CAD, for two reasons. Although CAD is designed to help in the identification of individual features, a CAD tool’s effectiveness is most frequently measured in terms of readers’ recalls of cases. This might seem appropriate for deciding how effective a CAD system will be in clinical practice (one of the important goals of studies about CAD). But if “false-target recalls” are frequent, the “case level” sensitivity measures might not represent the probability of the cancer being detected by biopsy, and this estimation error could differ between the with-CAD and without-CAD conditions, giving misleading estimates of CAD’s effects.

Imagine, for example, a patient who has cancer on her *right* breast, and this area goes unnoticed by the reader, who is misled by CAD to interpret an area in the *left* breast as the site of cancer. It is difficult to judge during a trial whether such recall decisions would be a detrimental outcome of screening. In the best case, the patient would be properly diagnosed. In the worst case, the further examinations of the patient may not only fail to detect the true cancer but also result in damage to healthy areas of the patient’s body.

The need for examining effects at the feature level is even more apparent with respect to the other purpose of research, i.e., to improve the design and use of CAD. It is important to know *how* a reader with CAD achieves a certain sensitivity and specificity at the case level. If this happens through unintended effects of CAD at the feature level, designers may receive misleading signals. For instance, in the above scenario of false-target recall, CAD might be contributing to a recall of a *patient* with cancer *despite not achieving its primary goal* of helping readers to notice cancerous *features*.

In conclusion, although the ultimate goal of CAD is to improve case recall decisions in breast screening, we argue that a fine-grained, feature-level analysis of data is necessary to understand how the tool actually affects the readers' decisions. This is necessary, for example, in order to discriminate among the different conjectures we have outlined in the previous section to explain the detrimental effects observed in specific categories of cases, and allow improvements in the design and use of CAD.

5.1.2 Limitations of studying the average effects of CAD on the readers population

Evaluation studies often focus on the effects of automation on readers' *average* performance. But averages may hide substantial variations between sub-populations. Our statistical analyses have proven very useful here. These analyses were motivated by the 'diversity modelling' approach (Littlewood et al., 2002, Strigini et al., 2003), which focuses on how performance varies across classes of cases.

Whether analyzing data at the level of cases (considering readers' recall decisions) or at the level of features (considering readers' evaluation of specific mammographic areas), we found that CAD had both beneficial and detrimental, *systematic* effects, depending on the reader-case (or reader-feature) pairs. In the HTA study, these effects just happened to cancel each other out, due to the composition of the sample of cases used, leading to the original study's result of no significant *average* impact.

If the systematic effects that we detected also occurred in practical clinical use, the net overall effect of CAD on the number of FN errors might still be close to zero, if the sample used in the study were representative of the populations of actual patients and readers. If not, the net overall effect of CAD could be positive, negative or null, in addition to some possible transfer of risk between categories of patients. Samples used in studies are not necessarily representative of the population for every clinical use, both because these populations may differ, and because studies may use intentionally skewed samples for practical reasons. Then, the observed average effect from the use of CAD in a study is *no guidance* for its effect in clinical practice, nor does it provide precise enough feedback for improving CAD. We therefore argue that the effects of CAD should be analyzed by stratifying the sample of decisions appropriately (as proposed e.g. in (Alberdi et al., 2005)), attempting to identify "strata" of readers and cases (or features) within which CAD effects vary less dramatically. Our classifications of cases or features by "difficulty" and of readers by sensitivity or "effectiveness" are just a first step. Ideally, one would classify them by variables that can be estimated both in studies and in clinical use before introducing CAD and that are sufficiently predictive of the effect of CAD. Further studies are necessary to find any such variables. This would give specific feedback to designers about strengths and weaknesses of tools; and stratified measurements would support preliminary estimation of the net effect of a CAD system in any future clinical use from the results in an evaluation study ("preliminary" as there will be other differences between the study and clinical environment; importantly, the magnitude of "automation bias" effects is probably affected by the prevalence of targets, and many study samples have artificially high cancer prevalence). The overall effect on the whole population of patients and readers could be estimated as a weighted sum of the different effects expected for each class of case-reader pairs (or feature-reader pairs), weighted with the relative frequencies of these classes (Strigini et al 2003).

5.2 Implications for Design of the Human-CAD System (tool, procedures, training)

As noted earlier, CAD errors in the HTA study were heavily correlated with the errors of readers without CAD (Alberdi, 2005). In other words, there was limited *diversity* between readers and the CAD tool. This may be due to two factors: the same mammograms that are hard to interpret correctly by readers are also hard for the tool to prompt correctly; or, incorrect prompting of a case by the CAD tool makes readers more likely to err in their turn. The data from the HTA study suggest that both factors were present.

Multiple approaches are open to those seeking to improve the effectiveness of CAD. We have pointed out that for some of the “automation bias” effects observed, there are alternative explanations, and research so far has not determined which ones are the real or predominant causes. More refined results would help to finesse approaches to improvement. Nonetheless, we can discuss here some approaches that are feasible, some of which are adopted in practice.

5.2.1 Diversity between errors by unaided readers and by CAD tools

An obvious path towards improvement is to improve either the CAD tool or its user, the reader. The tool may be improved by improvements in algorithms. The quality of the reader behavior may be improved by training. However, such improvements (besides being possibly expensive, and subject to a law of diminishing returns) may be ineffective because of lack of “diversity” between the error patterns of the CAD tool and the readers (cf the mathematical models in (Littlewood et al., 2002, Strigini et al., 2003). For example, if we improve a CAD tool’s ability to prompt types of cancers that readers very seldom miss anyway, the final FN rate may not improve.

It may be better to target the interaction and correlation between machine and reader error. Informally, the goal would be to focus the design of CAD tools on correctly prompting those cases where unaided readers would tend to fail. So, improvements in CAD effectiveness could be sought by increasing this diversity, even without improving the average sensitivity or specificity of a CAD tool. The tool could be tuned to be more sensitive for classes of cases on which readers tend to be less effective: these cases are natural candidates for CAD to make a difference. This strategy is feasible, since the designers of an alerting tool have a degree of freedom in choosing the trade-off between its FP and FN rates. Especially if the tool uses a combination of algorithms for identifying situations that require an alert, tuning these multiple algorithms may allow some "targeting" of the peaks and troughs in FN and FP rates to different classes of cases. The overall improvement is determined by the changes achieved for each class, weighted with their relative frequencies (cf Strigini et al, 2003). Thus tuning a CAD tool for “diversity” from its readers may improve the latter’s performance more than tuning it just to be as good as possible (in terms of sensitivity, specificity or any weighted combination of the two). An additional desirable effect of this strategy may be the scope for designers to reduce the overall FP rate of the tool (which tends to be high if the main design goal is to ensure low FN rate for all cases): as discussed earlier, readers find FP prompts annoying, and a high FP rate is plausibly one of the factors contributing to the observed “automation bias” effects. In view of how effects vary among readers, it may even be desirable to tune CAD differently for each individual reader, automatically or manually.

The importance of “diversity” between reader and CAD may have more general implications for designers of these tools (and alerting tools in general). A design philosophy aiming at reproducing in the CAD tool the outward behaviour of human experts may limit the effectiveness of these tools: they may tend to help a reader most reliably on those cases where the readers needs less help. They will still be immune to fatigue and random lapses of attention, and thus serve the goal of making reader performance more uniform, if these were (as they are thought to be) important causes of human failure. But, even from this viewpoint, a tool design focused on helping readers with cases where they are least effective could still be the more effective solution (Strigini et al., 2003).

5.2.2 The impact of absence of prompts on reader error

The evidence presented for possible error-inducing effects of CAD points to apparent “automation bias”, with readers behaving in “inappropriate” ways, e.g. tending to miss, when using CAD, cancers that they would notice when not using CAD. Defenses against such “inappropriate” behaviours may be devised with procedures and training as well as with changes to CAD tools.

Instructions for CAD use may say that readers must not allow machine prompts (or lack thereof) to change their opinion about a case from “recall” to “do not recall”. If readers comply, then a FN error from a CAD tool would not lead to a reader’s FN, except on cases on which the reader failed to notice the symptoms to start with. Computer support "could do no harm". But can readers actually comply with such guidelines? Much of an expert’s skill is formed by non-explicit pattern recognition rules and heuristics. If experts slowly adapt to relying on the tool's prompts for advice, at least for some types of cases, they may not realize that they are doing so. At least two forms of protection can be pursued.

One protection would be to make it easier for the reader to obey the prescribed procedure, or more difficult to violate it. For instance, the user interface can request readers to mark all features they are going to consider in their decision, before showing the tool's prompts. However, such measures may be cumbersome to enforce, especially with high load on the readers, causing them to take advantage of all available help.

Another possible protection is to make the reader less prone to being influenced in the wrong way by the tool, for instance by practices that remind readers of the possibility of erroneous prompting. Which forms of reminders would be appropriate depends on the specific working practices in a given environment. Examples of erroneous prompting could be given during training, just as, for example, pilot simulator training includes an unrealistically high rate of mechanical failure. A probably more effective, and more controversial, process would be to plant fictitious cases with incorrect CAD output in readers’ normal workload, rarely but yet frequently enough to refresh their memory of types of possible prompting errors.

One should notice that simply asking readers not to allow the presence or absence of a CAD prompt to affect their judgment about a feature could be “normatively incorrect” – that is, a prescription for non-optimal decision making. Prompts have some informative value: with high quality tools, absence of a prompt may be indeed a good indication of absence of cancer. It may be desirable to specify a heuristic procedure that readers could be trained to follow, that would give absence of prompts approximately the right weight in decisions. Simply training readers to ignore absence of prompts altogether is a simple, although suboptimal such heuristic. The question is whether readers could ever succeed in applying it, since their reactions to absence of prompts may be involuntary.

A factor that will certainly affect readers' performance is their level of experience with CAD. Several authors assume, both for cancer screening CAD and for advisory systems in general, that the phenomenon of incorrect prompting causing reader error is limited to users who lack experience with the prompting tool, and disappears with use, as users learn the "strengths and weaknesses" of their tools. This seems a one-sided, over-optimistic statement, since it is not yet supported by decisive empirical evidence:

- experience will help readers to recognize that the tool is fallible, but it will also teach them that in many situations it is normally reliable. CAD tools are tuned to have low FN rate at the cost of a high FP rate, so they are indeed quite reliable for the kinds of cancer signs they target. A natural learned behaviour may be to rely on absence of prompts to exclude cancer, even when cancer is present and could have been detected by the unaided reader
- experience will teach readers to recognize some situations in which the tool tends to fail, but this will be more likely the more the situation is one in which the reader is reliable and so needs the tool's support less.

5.2.3 Cognitive overload

A plausible cause for many errors with CAD is simple overload. Readers with CAD have additional work to do compared to readers without CAD: examining prompts, which may be many if the CAD tool has high FP rate. Reducing these FP rates may thus improve the sensitivity of readers with CAD (in addition to the probable effect of improving their specificity). As we suggested earlier, tuning CAD for lower overall sensitivity but more diversity (from readers) may allow designers to further reduce its FP rates. Other improvements can be sought by changing the information the CAD tool gives readers, for example, by not repeating prompts on features that readers have already noticed and marked.

A likely factor in the readers' cognitive load is their concern to explain the presence or absence of prompts in terms of regular patterns of CAD behaviour (Hartswood and Procter, 2000, Hartswood et al., 2003). To reduce this load, the CAD tool could be made to explain its behaviour when asked, for example incorporating decision support in the form of knowledge based diagnostic advice (Alberdi et al., 2004b). The challenge is to reliably 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 (Hartswood et al., 2003).

6 Final Remarks

The main findings we have shown in this chapter are indications that:

- CAD prompts do help readers detect some cancers, which is the main purpose of CAD use
- the presence of a prompt on a cancer feature makes readers not only more likely to detect the feature but also more likely to *interpret* the feature as cancerous, which goes beyond the scope of computer aided *detection*
- the absence of prompts on a cancer feature can reduce a reader's ability to detect the abnormality

- an excessive number of false prompts might damage not only reader specificity but sensitivity as well; it is one of the possible reasons for readers being less likely to notice true cancer features if these have not been prompted by CAD
- CAD's effects on reader decisions vary in complex ways, being beneficial for some groups of readers and cancer appearances and detrimental for others
- this combination of systematic positive and negative effects of CAD on reader decisions is likely to explain the lack of average effects of computer prompts on reader decisions reported in some evaluations of CAD
- any detrimental effects from CAD are not limited to the less accurate readers but, in fact, can be more acute for very skilled readers, at least for some categories of cancer features.

The general picture is thus mixed. In an ideal scenario, CAD would simply help readers to notice cancers that they would have otherwise missed, and thus improve their sensitivity, without excessive reductions to their specificity. However, this benefit is limited by various, somewhat subtle factors including: 1) possible "lack of diversity" between CAD tools and readers, meaning that the rare errors by CAD occur in precisely those cases in which a reader would need correct prompting; 2) "automation bias", meaning that sometimes CAD prompting (if incorrect, and sometimes even if correct) causes mistakes that a reader would not have made without CAD, because CAD use changes the readers' mental processes. To design CAD tools and activities, or to assess the benefit to be expected from CAD, requires considerations of all these factors.

We have also suggested some methodological recommendations:

- to assess the benefits and risks of using a CAD tool, statistical analyses need to go beyond average sensitivity and specificity, but take into account the role of readers' skills, the difficulty of features and cases, and the effect of correct or wrong computer prompting on reader decisions
- to understand better how CAD affects reader performance, it is important to study the decisions of readers at the level of features (vs. case level)

CAD tools are continually evolving and improving. It is tempting to assess a tool by its average false positive and false negative rates. But we have shown how effectiveness for improving readers' decisions depends on more complex statistical characteristics of a tool's behavior, and on subtle effects on readers' work. Improving a tool's sensitivity and specificity may not be the only (nor, in fact, the best) way for making it more effective. Striving to improve CAD tools' discrimination by improving it for "easy" cases might actually *amplify* risks for difficult cases (when read by more effective readers) without improving performance of the more effective readers. Dedicated studies would be necessary to refute this hypothesis.

More research is certainly desirable towards more detailed understanding of how the technology affects human decisions. Its results would help to resolve the controversies about the effectiveness of CAD tools as well as to improve their design and assessment.

Acknowledgements

The work described in this paper was supported in part by: Cancer Research UK (Population & Behavioural Science Committee) via grant C22515/A7339; the U.K. Engineering and Physical Sciences Research Council via the Interdisciplinary Collaboration on the Dependability of Computer Based Systems, "DIRC", and via project

INDEED, “Interdisciplinary Design and Evaluation of Dependability (EP/E000517/1); and the European Union Framework Programme 6 via the ReSIST (Resilience for the Information Society) Network of Excellence, contract IST-4-026764-NOE.

We would like to thank: Rosalind Given-Wilson for her collaborations and expert radiological advice; 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 in the Cancer Research UK project; our colleagues in the DIRC project for early discussions of these topics.

References

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

Alberdi, E., Taylor, P. & Lee, R. (2004b) Elicitation and Representation of Expert Knowledge for Computer Aided Diagnosis in Mammography. *Methods of Information in Medicine 239-46.,* 43**,** 239-246.

Alberdi, E., Povyakalo, A. A., 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.

Alberdi, E., Povyakalo, A. A., Strigini, L., Ayton, P. & Given-Wilson, R. (2008) CAD in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions. *Journal of Computer Assisted Radiology and Surgery*.

Astley, S. M. & Gilbert, F. J. (2004) Computer-aided detection in mammography. *Clin Radiol,* 59**,** 390-399.

Astley, S. M. (2005) Evaluation of computer-aided detection (CAD) prompting techniques for mammography. *Br J Radiol,* 78**,** S20-25.

Azar, B. (1998) Danger of automation: It makes us complacent. *APA monitor,* 29**,** 3.

Bainbridge, L. (1983) "Ironies of Automation". *Automatica*, 19, 775-779.

Bisantz, A. M. & Seong, Y. (2001) Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics,* 28**,** 85-97.

Castellino, R., Roehrig, J. & Zhang, W. (2000) Improved computer aided detection (CAD) algorithms for screening mammography. *Radiology***,** 400.

Cupples, T. E., Cunningham, J. E. & Reynolds, J. C. (2005) Impact of Computer-Aided Detection in a Regional Screening Mammography Program. *Am. J. Roentgenol.,* 185**,** 944-950.

Dassonville, I., Jolly, D. & Desodt, A. M. (1996) Trust between man and machine in a teleoperation system. *Reliability Engineering & System Safety (Safety of Robotic Systems),* 53**,** 319-325.

de Vries, P., Midden, C. & Bouwhuis, D. (2003) The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies,* 58**,** 719-735.

Dzindolet, M. T., Pierce, L. G., Beck, H. P. & Dawe, L. A. (2002) The perceived utility of human and automated aids in a visual detection task. *Human Factors,* 44**,** 79-94.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G. & Beck, H. P. (2003) The role of trust in automation reliance. *International Journal of Human-Computer Studies,* 58**,** 697-718.

FDA (1998) Pre-market approval decision. Application P970058 - http://www.fda.gov/cdrh/pdf/p970058.pdf. US Food and Drug Administration.

Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E. & Elmore, J. G. (2007) Influence of Computer-Aided Detection on Performance of Screening Mammography. *N Engl J Med,* 356**,** 1399-1409.

Fogg, B. J. & Hsiang, T. (1999) The elements of computer credibility. *CHI99.* Pittsburgh, PA.

Freer, T. W. & Ulissey, M. J. (2001) Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology,* 220**,** 781-786.

Galletta, D. F., Durcikova, A., Everard, A. & Jones, B. M. (2005) Does spell-checking software need a warning label? *Commun. ACM,* 48**,** 82-86.

Gromet, M. (2008) Comparison of Computer-Aided Detection to Double Reading of Screening Mammograms: Review of 231,221 Mammograms. *Am. J. Roentgenol.*, 190, 854-859.

Gur, D., Sumkin, J. H., Rockette, H. E., Ganott, M., Hakim, C., Hardesty, L., Poller, W. R., 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.

Hartswood, M. & Procter, R. (2000) Computer-Aided Mammography: A Case Study of Error Management in a Skilled Decision-making Task. *Journal of Topics in Health Information Management,* 20**,** 38-54.

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 Kuutti, K., Karsten, E. H., Fitzpatrick, G., Dourish, P. & Schmidt, K. (Eds.) *Eighth European Conference on Computer Supported Cooperative Work (ECSCW 2003).* Helsinki, Finland.

Khoo, L. A. L., Taylor, P. & Given-Wilson, R. M. (2005) Computer-aided Detection in the United Kingdom National Breast Screening Programme: Prospective Study. *Radiology,* 237**,** 444-449.

Lee, J. D. & Moray, N. (1994) Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies,* 40**,** 153-184.

Lee, J. D. & See, K. A. (2003) Trust in computer technology. Designing for appropriate reliance. *Human Factors*.

Littlewood, B., Popov, P. & Strigini, L. (2002) Modelling software design diversity - a review. *ACM Computing Surveys,* 33**,** 177-208.

Martin, C. D. (1993) The myth of the awesome thinking machine. *Commun ACM,* 36**,** 39-44.

McNicol, D. (1972) *A Primer of Signal Detection Theory,* London, George Allen & Unwin.

Meyer, J. (2001) Effects of warning validity and proximity on responses to warnings. *Hum Factors,* 43**,** 563-572.

Meyer, J. & Bitan, Y. (2002) Why better operators receive worse warnings. *Human Factors,* 44**,** 343-353.

Meyer, J., Feinshreiber, L. & Parmet, Y. (2003) Levels of automation in a simulated failure detection task. *IEEE International Conference on Systems, Man and Cybernetics, 2003.* Washington DC, USA.

Meyer, J. (2004) Conceptual issues in the study of dynamic hazard warnings. *Human Factors,* 46**,** 196-204.

Moray, N. (2003) Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics,* 31**,** 175-178.

Mosier, K. L., Skitka, L. J., Heers, S. & Burdick, M. (1998) Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology,* 8**,** 47-63.

Muir, B. M. (1987) Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies,* 27**,** 527-539.

Muir, B. M. (1994) Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics,* 37**,** 1905-1922.

Muir, B. M. & Moray, N. (1996) Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics,* 39**,** 429-460.

Parasuraman, R. & Riley, V. (1997) Humans and automation: Use, misuse, disuse, abuse. *Hum Factors,* 39**,** 230-253.

Parasuraman, R. & Miller, C. A. (2004) Trust and etiquette in high-criticality automated systems. *Communications of the ACM,* 47**,** 51-55.

Povyakalo, A. A., Alberdi, E., Strigini, L. & Ayton, P. (2004) Evaluating 'Human + Advisory computer' systems: A case study. IN Watts, A. D. a. L. (Ed.) *HCI2004,18th British HCI Group Annual Conference.* Leeds, UK, British HCI Group.

Povyakalo, A. A., Alberdi, E., Strigini, L. & Ayton, P. (2006) Divergent effects of computer prompting on the sensitivity of mammogram readers. London, UK, Technical Report, Centre for Software Reliability, City University.

Singh, I. L., Molloy, R. & Parasuraman, R. (1993) Automation-induced "complacency": development of the complacency-potential rating scale. *International Journal of Aviation Psychology,* 3**,** 111-122.

Skitka, L. J., Mosier, K. & Burdick, M. D. (1999) Does automation bias decision making? *International Journal of Human-Computer Studies,* 51**,** 991-1006.

Skitka, L. J., Mosier, K. & Burdick, M. D. (2000) Accountability and automation bias. *International Journal of Human-Computer Studies,* 52**,** 701-717.

Sorkin, R. D. & Woods, D. D. (1985) Systems with human monitors: A signal detection analysis. *Human-Computer Interaction,* 1**,** 49-75.

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

Tan, G. & Lewandowsky, S. (1996) A comparison of operator trust in humans versus machines. *Presentation of First International Cyberspace Conference on Ergonomics.*

Taplin, S. H., Rutter, C. M. & Lehman, C. D. (2006) Testing the Effect of Computer-Assisted Detection on Interpretive Performance in Screening Mammography. *Am. J. Roentgenol.,* 187**,** 1475-1482.

Taylor, P., Given-Wilson, R., Champness, J., Potts, H. W. & Johnston, K. (2004a) Assessing the impact of CAD on the sensitivity and specificity of film readers. *Clin Radiol*., 59, 1099-1105.

Taylor, P. M., Champness, J., Given-Wilson, R. M., Potts, H. W. E. & Johnston, K. (2004b) An Evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. *Brit J Radiol,* 77**,** 21-27.

Tseng, S. & Fogg, B. J. (1999) Credibility and computing technology. *Communications of the ACM,* 42**,** 39-44.

Warren Burhenne, L. J., Wood, S. A., D'Orsi, C. J., Feig, S. A., Kopans, D. B., O'Shaughnessy, K. F., Sickles, E. A., Tabar, L., Vyborny, C. J. & Castellino, R. A. (2000) Potential Contribution of Computer-aided Detection to the Sensitivity of Screening Mammography. *Radiology,* 215**,** 554-562.

Wiener, E. L. (1981) Complacency: is the term useful for air safety? *26th Corporate Aviation Safety Seminar.* Denver, CO, Flight Safety Foundation, Inc.

Zheng, B., Ganott, M. A., Britton, C. A., Hakim, C. M., Hardesty, L. A., Chang, T. S., Rockette, H. E. & Gur, D. (2001) Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings. *Radiology,* 221**,** 633-640.

Zheng, B., Richard, G. S., Sara, G., Christiane, M. H., Ratan, S., Luisa, W. & David, G. (2004) Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments1. *Academic radiology,* 11**,** 398-406.