
Article:
Povyakalo, AA, Alberdi, E, Strigini, L et al. (2013) How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography. Medical Decision Making, 33 (1). pp. 98-107. ISSN: 0272-989X

https://doi.org/10.1177/0272989x12465490

TITLE:

How to discriminate between computer-aided and computer-*hindered* decisions: a case study in mammography.

AUTHORS:

Andrey A. Povyakalo, PhD[1]

Eugenio Alberdi, PhD[1]

Lorenzo Strigini, M.Eng[1]

Peter Ayton, PhD[2]

ADDRESSES:

[1] Centre for Software Reliability, City University London, Northampton Square, London, EC1V 0HB, UK

[2] Psychology Department, City University London, Northampton Square, London, EC1V 0HB, UK

SHORTENED VERSION OF THE TITLE:

How to discriminate between computer-aided and computer-hindered decisions.

FUNDING:

The work described in this paper has been partly funded by: the UK Engineering and Physical Sciences Research Council (EPSRC) through projects DIRC (the Dependability Interdisciplinary Research Collaboration) and INDEED ("Interdisciplinary Design and Evaluation of Dependability", EP/E000517/1) as well as by Cancer Research UK (grant GR/C22515/A7339).

Structured abstract - 252

Main text - 3980

Figures - 5

Tables - 3

References - 32

ABSTRACT

Background: Computer aids can affect decisions in complex ways, potentially even making them worse; common assessment methods may miss these effects.
We developed a method for estimating the quality of decisions and how computer aids affect it, and applied it to computer-aided detection (CAD) of cancer, re-analysing data from a published study where 50 professionals ("readers") interpreted 180 mammograms, both with and without computer support.
Method: We used stepwise regression to estimate how CAD affected the probability of a reader making a correct screening decision on a patient with cancer (sensitivity), thereby taking into account the effects of the difficulty of the cancer (proportion of readers who missed it) and the reader's discriminating ability (Youden's Determinant.) Using regression estimates we obtained thresholds for classifying *a posteriori* the cases (by difficulty) and the readers (by discriminating ability).
Results: Use of CAD was associated with a 0.016 increase in sensitivity (95% CI: 0.003, 0.028) for the 44 least discriminating radiologists for 45 relatively easy, mostly CAD-detected, cancers. However, for the 6 most discriminating radiologists, with CAD sensitivity *decreased* by 0.145 (95% CI: 0.034, 0.257)
for the 15 relatively difficult cancers.
Conclusions: Our exploratory analysis method reveals unexpected effects. It indicates that, despite the original study detecting no significant *average* effect, CAD *helped* the less discriminating readers but *hindered* the more discriminating readers. Such differential effects, although subtle, may be clinically significant and important for improving both computer algorithms and protocols for their use. They should be assessed when evaluating CAD and similar warning systems.

Keywords: Automation bias, breast cancer screening, computer-aided detection, medical decisions, computer advice

INTRODUCTION

Computer aids are increasingly important in medical decision making (1). Software algorithms may provide highly reliable hints and advice, and this reliability can be estimated with methods in common use in software engineering. However, what really matters is whether using computer aids make users' decisions better - or worse. Negative effects of decision support have been reported in other fields (2-4). In aviation, for example, research on "automation bias" has documented situations in which a human operator makes more errors when being assisted by a computerized device than when performing the same task without computer assistance (5).

Here we introduce an analysis method, applied to a case study in computer aided detection (hereafter CAD) in breast cancer screening. This case study provides evidence of both positive and negative effects of automation and justifies our proposed method.

There is a debate about whether current breast cancer screening regimes report too many abnormalities that should not be treated as cancers. However, our topic is not the assessment of screening regimes and whether they target the right class of abnormalities; but rather, given a target class of abnormalities, the extent to which a computerised aid helps or hinders their accurate detection by clinicians. Therefore, we use the terms "cancer" and "cancer case" as they were used in the clinical study (6) which we re-analyse here.

Cancer Screening and previous studies of CAD

In breast cancer screening, expert clinicians (readers) examine mammograms and decide whether the patient should be recalled for further tests because they suspect

cancer. CAD tools have been designed to help readers by identifying and marking ("prompting") regions of interest (ROI) on a digitised mammogram, to prevent clinicians from overlooking them. However, evidence about the benefits of CAD is inconclusive (5-14). Extensive literature surveys can be found in (7), (8) and (9).

Three leading approaches to the evaluation of CAD effectiveness are:

(a) Estimate the potential of CAD impact by comparing CAD performance with human performance.

E.g. Warren-Burhenne et al (10) estimated the potential benefit of CAD by checking whether a cancer was prompted by the CAD tool and how many radiologists had judged it actionable without using the CAD tool.

(b) Compare the performance of readers using CAD with their performance without CAD. E.g. Taylor et al (6, 8, 11) assessed the impact of CAD prompts on the sensitivity and specificity of mammogram readers with and without CAD.

(c) Infer the effectiveness of CAD by establishing statistical equivalence between readers' use of CAD and another well established procedure (e.g. double reading). E.g. Gilbert et al (12) compared double reading without CAD and single reading with CAD.

We argue that these methodological approaches can be useful, but are ultimately insufficient and even potentially misleading for assessing the impact of using CAD and similar tools. A concern is that the observed effects may be an artefact of the composition of the samples used, and that implicit assumptions about the tool's effects (e.g. that the tool, when used as prescribed, will not undermine human performance) may bias the analysis of results. We propose an alternative method to estimate the

impact of decision support that reveals otherwise latent systematic differences in effects of CAD on different users and different sets of cases. The remainder of this paper describes an application of this approach.

An approach based on analysis of Human-computer diversity in computer-based systems.

We studied CAD in mammography as part of the U.K. DIRC project (the Interdisciplinary Research Collaboration on Dependability of computer-based systems - www.dirc.org.uk). The CAD tool we considered in all our analyses is the R2 ImageChecker M1000, as evaluated by Taylor et al (6, 8, 11).

In the terminology of engineering, CAD is an example of protective redundancy with diversity: the computer is meant to help users to correct some of their omission errors. The diversity between the user and the software should make them unlikely to exactly duplicate each other's errors (13) . We know from previous studies of redundancy, especially of computer systems (14, 15), that its effectiveness is greatly affected by variations in "difficulty" between cases. Accordingly we looked for similar effects in CAD. We defined "difficulty" of a cancer case for human readers as the probability of a randomly selected reader, without CAD, not recalling that case. Difficulty of a cancer case for the CAD tool was defined as the probability of the CAD tool prompting that case incorrectly (i.e. not prompting cancer areas, whether or not it placed prompts on other areas). We developed probabilistic models showing the probability of incorrect decisions as a function of how these two "difficulty" measures vary across the population of cases (13). To estimate model parameters we analysed the data from (6), the best data set, to our knowledge, for this purpose, since each mammogram was

examined by each reader *both* with and without CAD. We also conducted two supplementary studies (16, 17) with a set of cases selected to have large proportions of cancers incorrectly prompted by CAD. In these studies, radiologists proved significantly less sensitive with CAD than without, suggesting that incorrect CAD prompts may hinder readers under some circumstances, which is consistent with the findings of Zheng et al (18). Our further analyses of data from (6) suggested that CAD reduced radiologists' sensitivity in decisions about the more difficult cases – most of which were incorrectly prompted by the CAD tool – and increased sensitivity for comparatively easy cases (16). Other indications were that the CAD tool's errors correlated positively with those of readers, and that its errors affected the readers' decisions (17).

These results appear to contradict the statement from the "ImageChecker M1000 - Device labelling" (19) that "... the potential for missed lesions is not increased over routine screening mammography when the ImageChecker is used as labelled...". Possible explanations can be conjectured (2) with the help of direct observations of readers (17).

Some studies have highlighted the importance of reader variability when interpreting the results of clinical trials in mammography. For example, Wagner et al (20) conjecture that "... the ambiguity surrounding the effectiveness of mammography is due in part to the observed range of reader skills...".

Here we present a more complete method of statistical analysis than in (17), which takes into account not only the effects of case difficulty but also of readers' performance (i.e. the difference between their true positive rate and false positive rate, when operating

without CAD). Other factors we consider are whether CAD prompts were correct or wrong and whether the cancer was screen detected or a false negative at screening (interval cancer). We also discuss the potential role of these effects in explaining discrepancies between results from different controlled studies.

METHOD

Data source

We conducted supplementary statistical analyses of the data from the study published by Taylor and colleagues (6) . This study is, to our knowledge, the only study where each reader examined all the cases in both conditions, with and without CAD.

Taylor et al (6, 8, 11) evaluated the R2 ImageChecker M1000 in a retrospective study which assessed the impact of CAD prompts on the sensitivity and specificity of film readers (30 radiologists, 5 breast clinicians and 15 radiographers), who read 180 films, all of which had a proven outcome. The set included 60 cancers (20 false negative interval cancers – i.e., cancers that had not been detected during screening using these mammograms but which were classified as false negatives when the mammograms were later reviewed by a panel of experts – and 40 screen detected cancers, i.e., detected during screening). The case set was selected in such a way that it contained a mixture of different types of cancerous signs (masses, micro-calcifications, spiculated lesions and asymmetries) and a combination of easy and difficult cases. The selection criteria (6, 8, 11) ensure that the case set covers a variety of categories of cases, although the set is not necessarily a representative sample of the population addressed

by screening. Each reader read each case twice, once with CAD ("prompted condition") and once without CAD ("unprompted condition"). The order of the reading sessions was randomised separately for each reader (8, 11). The conclusions of the original study were that "No significant difference was found for readers' sensitivity or specificity between the prompted and unprompted conditions" and “There was no evidence that the use of R2 affects able and less able readers’ sensitivities or specificities differently”

We applied each of the statistical procedures described below to various subsets of the cancer cases in the set (Table 1).

Classification of cases by “difficulty” and Readers by “Discriminating Ability”

We examined how the apparent effect of CAD on readers' sensitivity varied with two parameters: 1) the "difficulty" of the cancer: we define the difficulty d_i of cancer i as the fraction of readers who, without CAD ("unprompted condition"), missed the cancer; 2) the discriminating ability (DA) s_j of the reader j in the unprompted condition, used as a measure of a reader’s skill.

We measure a reader's DA with Youden’s Determinant (21), i.e. as the difference between the proportion of cancers each reader (correctly) recalled (sensitivity, true positive rate) and the proportion of normal cases the reader (wrongly) recalled (false positive rate):

$$DA = \text{sensitivity} - (1 - \text{specificity}) = \text{sensitivity} + \text{specificity} -- 1,$$

(each reader’s specificity here is calculated for the 120 non-cancer films, which do not otherwise enter into the present analysis)

If *RR* is the reader recall rate and *BR* is the base rate of cancers, then

$$RR = (1 - \text{specificity}) + BR \times DA.$$

If the reader does not discriminate between cancers and normal cases, she recalls any case with the same probability and her DA = 0. If the reader recalls all the cancers and no normal cases (perfect discrimination), her DA = 1. In the extreme (and unrealistic) case, in which the reader recalls no cancers but all the normal cases, her DA = -1. For real readers, we expect $0 \leq DA \leq 1$.

"Impact" of CAD: Logistic Regression Estimates

Our method is based on the assumption that reading is affected both by systematic differences between readers (more or less effective at the task) and between cases (more or less difficult to decide), and by random variations in the performance of a specific reader, even when examining a specific case. Thus, we assume that for each reader and each case, there is a probability (between 0 and 1) of that reader recalling that specific cancer. The study data provide a binary indicator describing whether the reader recalled that case on two specific occasions (one unprompted and one prompted, in randomised order). From these data, we sought to estimate the probability of a specific reader recalling a specific cancer in the unprompted (*un*) and prompted (*pr*) condition, choosing as estimates two functions $p_{un}(d_i, s_j)$ and $p_{pr}(d_i, s_j)$ of the readers' *DA* (averaged over all the 180 cancer and non-cancer films), s_j , and the cancer's difficulty (averaged over all readers), d_i. We obtained these functions as logistic regression models:

$$\text{logit}(p_{un}(d_i, s_j)) = a_0 + a_1 \cdot d'_i + a_2 \cdot (d'_i)^2 + a_3 \cdot (d'_i)^3 + a_4 \cdot s'_j + a_5 \cdot (s'_j)^2 + a_6 \cdot (s'_j)^3 + a_7 \cdot d'_i \cdot s'_j + a_8 \cdot (d'_i)^2 \cdot s'_j + a_9 \cdot d'_i \cdot (s'_j)^2 \quad [1]$$

$$\text{logit}(p_{pr}(d_i, s_j)) = b_0 + b_1 \cdot d'_i + b_2 \cdot (d'_i)^2 + b_3 \cdot (d'_i)^3 + b_4 \cdot s'_j + b_5 \cdot (s'_j)^2 + b_6 \cdot (s'_j)^3 + b_7 \cdot d'_i \cdot s'_j + b_8 \cdot (d'_i)^2 \cdot s'_j + b_9 \cdot d'_i \cdot (s'_j)^2 \quad [2]$$

where n is the number of cases; m is the number of readers,

$$d'_i = \text{mlogit}(d_i, n);$$
$$s'_j = \text{mlogit}(s_j, m);$$

and

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right);$$
$$\text{mlogit}(x, k) = \log\left(\frac{x + 0.5/k}{1 - x + 0.5/k}\right);$$

are the well known logit and modified logit transformations (22).

To determine the form of the polynomials [1] and [2] and estimate the unknown coefficients a_l, b_l for $l = 0..9$, we applied stepwise regression with the Akaike information criterion (AIC) (23), which makes the coefficients a_l, b_l for the non-significant terms of logistic models [1] and [2] equal to 0 while estimating the others with the standard procedure for generalised linear models (24, 25) .

Our analyses were exploratory. Therefore, we were not fitting a pre-selected statistical model to the data, but extracting a model from the data by stepwise regression using AIC (see e.g. (24)).

The procedure starts with the 'empty' model, in which the right-hand sides of the equations for $\mathrm{logit}(p_{\mathrm{un}}(d_{\mathrm{i}}, s_{\mathrm{j}}))$ or $\mathrm{logit}(p_{\mathrm{pr}}(d_{\mathrm{i}}, s_{\mathrm{j}}))$ contain a single, constant term a_0 or b_0. Then, given a certain collection of possible model terms as in equations [1] and [2], at each step it adds the most informative term to the model and removes a term when it becomes noninformative, until no term can be added or removed without increasing AIC:

$$\mathrm{AIC} = \mathrm{Deviance} + 2 \cdot \text{number of parameters} + \mathrm{const}$$

(Deviance is a measure of discrepancy between the observed responses and those estimated by the model (24))

Thus, AIC resolves the trade-off between model accuracy and model complexity (which may lead to over-fitting) by penalizing the addition of every new term to the model. A new term may be added to the model only when the resulting reduction of deviance is greater than the penalty for adding the term. In the final model all the terms are statistically significant.

Because the data were sparse (26) the distributions of model deviances are not Chi-squared and the standard analysis of deviance is not applicable. Therefore, we applied the le Cessie-van Houwelingen global test (27) to the null hypothesis of no difference between the observed responses and those estimated by the regression models [1] and [2]. For all sub-populations of cancers from rows and columns of Table 1, the le Cessie-van Houwelingen global test did not reject the null hypothesis for either of these two regression models, corroborating the validity of estimates obtained by these analyses. For instance, for the sub-population including all cancers the test showed: p-

value = 0.52 (Z=-0.64) for the unprompted condition and p-value = 0.25 (Z= -1.14) for the prompted condition.

Then, we estimated the systematic impact of computer prompts on readers' sensitivity as

$$imp(d_i, s_j) = p_{pr}(d_i, s_j) - p_{un}(d_i, s_j) .$$

A positive value of this "impact" estimate for a certain pair of values *s* and *d* indicates that computer support is helpful: the probability of a reader with discriminating ability *s* recalling a case of difficulty *d* would be greater with CAD than without it. Similarly, a negative value indicates a detrimental effect.

We classified the impact of CAD as significant if the 95% pointwise confidence intervals (28) for $p_{pr}(d_i,s_j)$ and for $p_{un}(d_i,s_j)$ did not overlap.

The above method produces best-fit estimates for probabilities of correct (case i, reader j) decisions in the two conditions, prompted and unprompted, as a function of the two independent variables chosen, to account for the differences between the reader-case pairs. With the same independent variables, alternative regression methods could also be used; for instance focusing only on the differences observed for each case-reader pair between decisions in the two conditions, one can apply the same method to the (mutually exclusive) paired outcomes:

1. *Decision is aided by CAD*: Reader *i* recalls case *j* in the prompted condition *and* does not recall the same case in the unprompted condition with probability $p_{aid}(d_i, s_j)$.

2. *Decision is hindered by CAD*: Reader i recalls case j in the unprompted condition *and* does not recall the case in the prompted condition with probability $p_{hin}(d_i, s_j)$.

One can see that,

P(correct prompted decision) – P(correct unprompted decision) =

P(decision aided) + P(both decisions correct) –

(P(decision hindered) + P(both decisions correct)) =

P(decision aided) – P(decision hindered).

Therefore, if the errors of the models are negligible, then

$$imp(\, d_i, s_j \,) = p_{pr}(d_i, s_j) - p_{un}(d_i, s_j) \approx \; p_{aid}(d_i, s_j) - p_{hin}(d_i, s_j)$$

RESULTS

Case Difficulty, Reader DA without CAD and CAD tool sensitivity

Many cancer cases in the study (Figure 1) were "easy": in the unprompted condition, 30% (18/60) of all cases had d = 0, that is, they were recalled by all 50 readers. The average case difficulty was 0.24 (CI: 0.17, 0.32).

The readers' DA without CAD (Figure 2) varied between 0.37 and 0.76 with average value 0.57.

The correlation between readers' DA and their sensitivity in the unprompted condition was not statistically significant (Figure 3; the linear regression coefficient is 0.25 with t-value = -1.79 and p-value = 0.08). Therefore, we do not believe that our results were significantly influenced by regression towards the mean (29) .

The CAD tool's sensitivity for different sub-populations of cases (Table 1) is summarised in Table 2.

Logistic models

Our method resulted in the following two logistic models for isolated outcomes (correct/incorrect decisions in the prompted and unprompted conditions)

$$\text{logit}(p_{un}(d_i, s_j)) = a_0 + a_1 \cdot d'_i + a_3 \cdot (d'_i)^3 + a_5 \cdot (s'_j)^2 + a_7 \cdot d'_i \cdot s'_j \quad [5]$$

$$\text{logit}(p_{pr}(d_i, s_j)) = b_0 + b_1 \cdot d'_i + b_7 \cdot d'_i \cdot s'_j + b_9 \cdot d'_i \cdot (s'_j)^2 \quad [6]$$

We note that both models contain mixed terms (term coefficients a_7, b_7, b_9), indicating that the reader-case effects are significant.

CAD's impact

Figure 4 shows contour plots of the estimated impact of CAD, *imp(d, s)* for all cancer cases. Areas with statistically significant values are outlined by dashed curves. Table 2 shows that the estimated impact *imp(d, s)* varies markedly among different categories of cases and readers.

Figure 5 shows the impact, for the same data as Figure 4, but estimated with the models fitted for paired outcomes: "decision is aided by CAD" and "decision is

hindered by CAD". The pattern is the same as in Figure 4, i.e.

$$imp(d_i, s_j) \approx p_{aid}(d_i, s_j) - p_{hin}(d_i, s_j)$$

As an extra check on these results, we fitted the 2-dimentional complete 3rd order multinomial logistic model (30) to paired outcomes. The patterns of estimated CAD impact showed no substantial differences from the results obtained by our method.

To validate the regression estimates against the raw data, we used the plot in Figure 4 to suggest an *a posteriori* classification for both cases and readers, based on the magnitude of estimated CAD impact. For instance, In Figure 4 the horizontal line at s = 0.65 seems a good "ad hoc" separator of readers hindered by CAD from readers who benefited from CAD. Thus, we classified the readers into 6 "highly discriminating" readers ($s \geq 0.65$) and the remaining 44 "less discriminating" readers and the cancers into 45 "easy" ($d < 0.5$) and 15 "difficult" ($d \geq 0.5$).

Combining groups of readers and cases defined by these two classifications, we compared observed reader sensitivities with and without CAD.

The sensitivity of the 44 less discriminating readers for the 45 easy cases improved with CAD by 0.016 (95%CI: 0.003, 0.028). However, the sensitivity of the highly discriminating readers for the difficult cases decreased with CAD by 0.145 (95%CI: 0.034, 0.257).

Repeating the check for different subsets of the cancer cases, we observed significant negative impact on the sensitivity of the highly discriminating readers for interval

cancers (average imp= -0.12; 95%CI: -0.22, -0.01) and no statistically significant impact for sub-populations of correctly or incorrectly prompted cancers.

Difficult cases for readers are difficult cases for CAD

Table 3 shows that, when processing easy cases ($d \leq 0.5$), there were no statistically significant differences between the average sensitivities of: (i) the more discriminating readers ($s > 0.65$) without CAD; (ii) the less discriminating readers ($s \leq 0.65$) without CAD; (iii) the CAD tool alone. The same is true for difficult cases ($d > 0.5$). All this is consistent with our earlier observation that the correlation between readers' discriminating ability and sensitivity is not statistically significant.

However, the three differences between the columns "Easy cases" and "Difficult cases" are all significant. That is, the cases that are difficult for an average reader are also difficult for more discriminating readers and for the CAD tool alone, which can be considered a weakness in a decision support tool. (13)

DISCUSSION AND CONCLUSIONS

Our method of analysis showed systematic, divergent effects of CAD prompts on readers' sensitivity: beneficial for some categories of readers and cases but detrimental for others. Apparently the same CAD set-up could produce a positive or a negative overall effect if used with different populations of cases and/or readers; accordingly we conclude that the finding of no significant overall effect in the original study (6) is due

to systematic positive and negative effects canceling out in the sample used, rather than the absence of any CAD influence.

We found a positive association between computer prompts and improved sensitivity of the less discriminating readers for comparatively easy cases, mostly screen-detected cancers. This is the intended effect of correct computer prompts. An unexpected finding is the association between use of CAD and degraded sensitivity of readers for comparatively difficult cases; particularly striking because this affected the highly discriminating readers. This finding is, however, consistent with our aforementioned empirical study (16, 17) and the study by Zheng et al (18), which strongly suggested that readers using CAD were biased by incorrect computer outputs.

Plainly CAD can affect readers with different skills in very different ways. Other studies have suggested limited effectiveness of CAD in mammography (31, 32), however the finding that CAD can systematically improve or worsen readers' performance depending on their skill and the difficulty of cases is a novel finding. These findings can parsimoniously be attributed to general mechanisms that affect how people respond to advice - specifically automated advice (2, 4) - which support the expectation that CAD effects may vary markedly between different contexts of use. Effects may even be different for new releases of the specific tool used in the study (6) and the commonly used summary measures of the quality of a CAD tool – its sensitivity and specificity – may not be good predictors of its impact on the actual decisions of clinicians.

Accordingly we propose that the degree of diversity between readers' and CAD's false negative errors, as well as false positive errors, and possibly other subtler

parameters, are worth including in evaluations of the impact of CAD. In particular, in assessing CAD effectiveness in a specific context, it is useful to check for patterns of systematic improvement and/or degradation of decisions for classes of readers and of cases, as observed in this case study.

Although we demonstrated our method by analysing effects on readers' sensitivity, it can also be applied to explore effects on readers' specificity.

Our study offers some methodological contributions that would be of general value in assessing computer-supported decision making:

- analyses should consider that a computer aid may sometimes do harm (cause wrong decisions); ignoring this possibility may produce substantial misinterpretation of experimental results (e.g., in this case, "there was no overall effect from CAD, therefore one can conclude that readers ignored CAD prompts");

- computer aids may have different effects for different users as well as for different decision problems (e.g. different mammograms), complicating extrapolation from small artificial samples to real populations of patients and clinicians;

- assessment methods should be chosen to detect such systematic differences, to avoid decisions based on spurious extrapolation and to inform the design of decision aids and of the protocols for their use;

- our method of exploratory analysis identifies effects that are hidden when data are analysed in the aggregate only. Our regression method is akin to clustering, or a posteriori stratification of the data, so as to reveal groups of {user, case} pairs for which CAD had markedly different effects, and thus avoids inappropriate extrapolations from average results over a diverse population, and suggests ways of improving both assessment and design of computer-assisted activities;

- using an indicator of the intrinsic difficulty of a decision and an indicator of the general ability of a decision-maker as the two independent variables for the analysis was an effective basis for detecting previously overlooked effects.

ACKNOWLEDGEMENTS

This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through projects DIRC (the Dependability Interdisciplinary Research Collaboration) and INDEED ("Interdisciplinary Design and Evaluation of Dependability", EP/E000517/1) and by Cancer Research UK (grant GR/C22515/A7339). Special thanks go to Paul Taylor, from University College London (UCL) for granting access to the data from the clinical trial (funded by the UK Health Technology Assessment programme) and for providing valuable feedback. We would also like to acknowledge the collaboration of Jo Champness (from UCL) and the readers who participated in the trial and our follow-up experiment. Many thanks to Jonathan Baron (from University of Pennsylvania), to the editor and anonymous reviewers, and to our colleagues at City University London, especially Bev Littlewood,

Martin Newby, Kizito Salako and last but not least David Wright for their useful comments.

REFERENCES

1. Lyman JA CW, Bloomrosen M, Detmer DE. Clinical decision support: progress and opportunities. Journal of the American Medical Informatics Association : JAMIA. 2010 Sep-Oct;17(5):487-92.

2. Alberdi ES, L. Povyakalo, A. and Ayton, P. Why Are People's Decisions Sometimes Worse with Computer Support. In: Bettina Buth GR, Till Seyfarth, editor. SAFECOMP 2009, The 28th International Conference on Computer Safety, Reliability and Security; Hamburg, Germany: Springer. p. 18-31.

3. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association : JAMIA. 2012;19(1):121-7.

4. Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. Hum Factors. 1997;39:230-53.

5. Skitka LJ. Accountability and automation bias. Int J Hum-Comp Stud. 2000;52:701-17.

6. Taylor PM, Champness J, Given-Wilson RM, Potts HWE, Johnston K. An Evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. Brit J Radiol. 2004;77:21-7.

7. Astley SM, Gilbert FJ. Computer-aided detection in mammography. Clin Radiol. 2004;59(5):390-9.

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

9. Noble M BW, Uhl S, Schoelles K. Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. Archives of gynecology and obstetrics. 2009;279881-90(6):881-90.

10. Warren-Burhenne LJ, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology. 2000;215(2):554-62.

11. Taylor P, Given-Wilson R, Champness J, Potts HW, Johnston K. Assessing the impact of CAD on the sensitivity and specificity of film readers. Clin Radiol 2004;59(12):1099-105.

12. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J, et al. Single Reading with Computer-Aided Detection for Screening Mammography. New England Journal of Medicine. 2008;359(16):1675-84.

13. Strigini L, Povyakalo AA, Alberdi E. Human-machine diversity in the use of computerised advisory systems: a case study. Int Conf on Dependable Systems and Networks (DSN'03) June 22 - 25 2003; San Francisco: IEEE.

14. Littlewood B, Miller DR. Conceptual modelling of coincident failures in multi-version software. IEEE Trans Software Engineering. 1989;15(12):1596-614.

15. Littlewood B, Popov P, Strigini L. Modelling software design diversity - a review. ACM Computing Surveys. 2002;33(2):177-208.

16. Alberdi E, Povyakalo AA, Strigini L, Ayton P. Effects of incorrect CAD output on human decision making in mammography. Acad Radiol. 2004;11(8):909-18.

17. Alberdi E, Povyakalo AA, Strigini L, Ayton P, Hartswood M, Procter R, et al. Use of Computer Aided Detection tools in screening mammography: A multidisciplinary investigation. Brit J Radiol. 2005;78:31-40.

18. Zheng B, Ganott MA, Britton CA, Hakim CM, Hardesty LA, Chang TS, et al. Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings. Radiology. 2001 December 1;221(3):633-40.

19. Pre-market approval decision. Application P970058: US Food and Drug Administration; 1998 June 26.

20. Wagner RF, Beam CA, Beiden SV. Reader Variability in Mammography and Its Implications for Expected Utility over the Population of Readers and Cases. Med Decis Making. 2004 November 1;24(6):561-72.

21. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5.

22. Anscombe FJ. On estimating binomial response relations. Biometrika. 1956;43:461 - 4.

23. Sakamoto Y, Ishiguro M, G K. Akaike Information Criterion Statistics. Dordrecht/Tokyo: D. Reidel Publishing Company; 1986.

24. Dobson AJ. An introduction to generalized linear models. London: Chapman and Hall; 1990.

25. McCullagh P, Nelder JA. Generalised Linear Models. London: Chapman and Hall; 1989.

26. Hosmer DW, Hosmer T, le Cessie S, S L. A comparison of goodness-of-fit tests for the logistic regression model. Stat in Med. 1997;16(9):965-80.

27. Cessie Sl, Houwelingen JCv. A goodness of fit test for binary regression models, based on smoothing methods. Biometrics. 1991;47:1267-82.

28. Smyth G. Pointwise confidence intervals for logit predictions. 24 December 2002 [cited 25 July 2012]; Available from: http://www.statsci.org/s/predlogi.html

29. Kirkwood BR, Sterne JAC. Essentials of medical statistics. Second ed. Oxford:

Blackwell Science; 2003.

30. Hosmer DW, Lemeshow S. Applied Logistic Regression: Wiley; 2000.

31. Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim C, Hardesty L, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst. 2004;96(3):185-90.

32. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer. 2008 Mar 18;44(6):798-807.

FIGURE LEGENDS

Figure 1 Distribution of difficulty (*d*) of 60 cancers.

Figure 2. Distribution of 50 readers' DA (*s*, Youden's determinant) without CAD.

Figure 3

Scatter-plot of readers' DA (Youden's determinant) without CAD vs. readers' sensitivity without CAD.

Figure 4.

Contour plot of estimated impact of CAD $imp(d, s)$ for all cancers. Here and in the following figures, we consider the impact significant if the 95% pointwise confidence intervals for $p_{un}(d_i,s_j)$ and $p_{pr}(d_i,s_j)$ do not overlap; statistically significant values of impact are outlined with the dashed curve. Values for the points with extreme impact are given in Table 2.

Figure 5.

Contour plot of impact of CAD: $imp(d, s)$ for all cancers, estimated with logistic models for paired outcomes: "decision is aided by CAD" and "decision is hindered by CAD". See Figure 4 for an explanation of significance values.

FIGURES AND TABLES

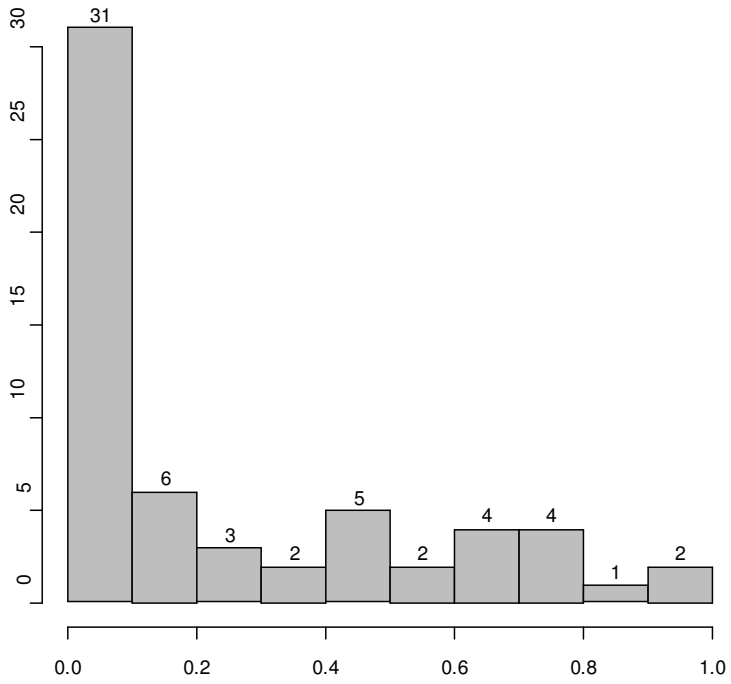

Figure 1 Distribution of difficulty (*d*) of 60 cancers.

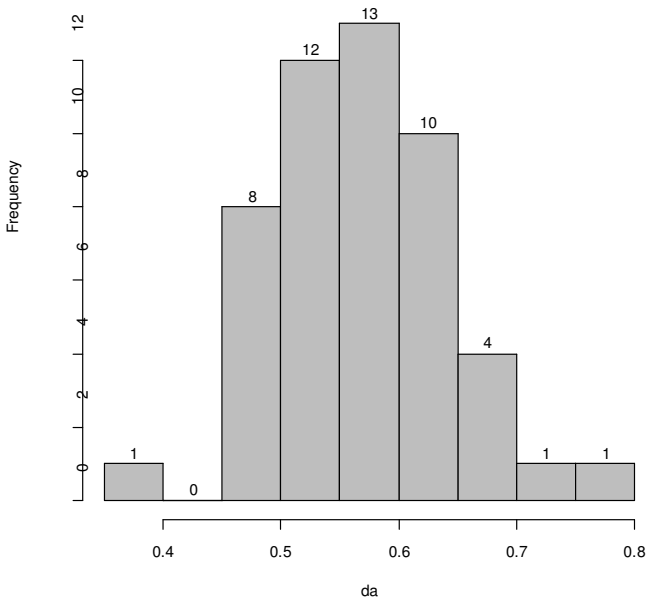

Figure 2. Distribution of 50 readers' DA (*s*, Youden's determinant) without CAD.

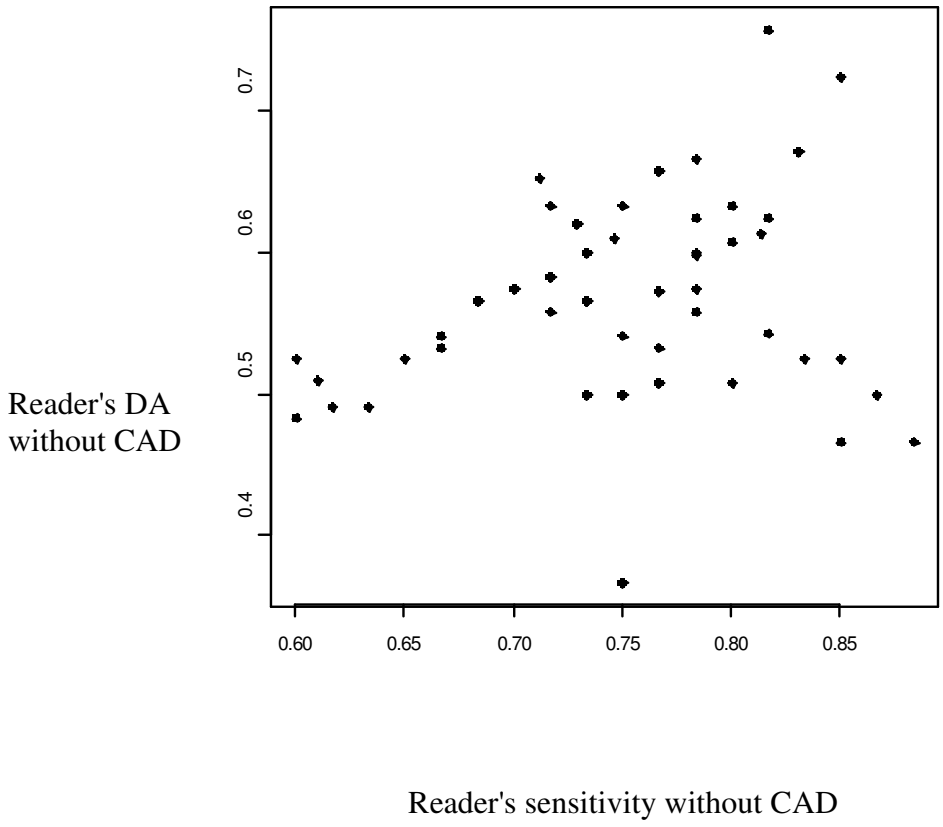

Figure 3

Scatter-plot of readers' DA (Youden's determinant) without CAD vs. readers' sensitivity without CAD.

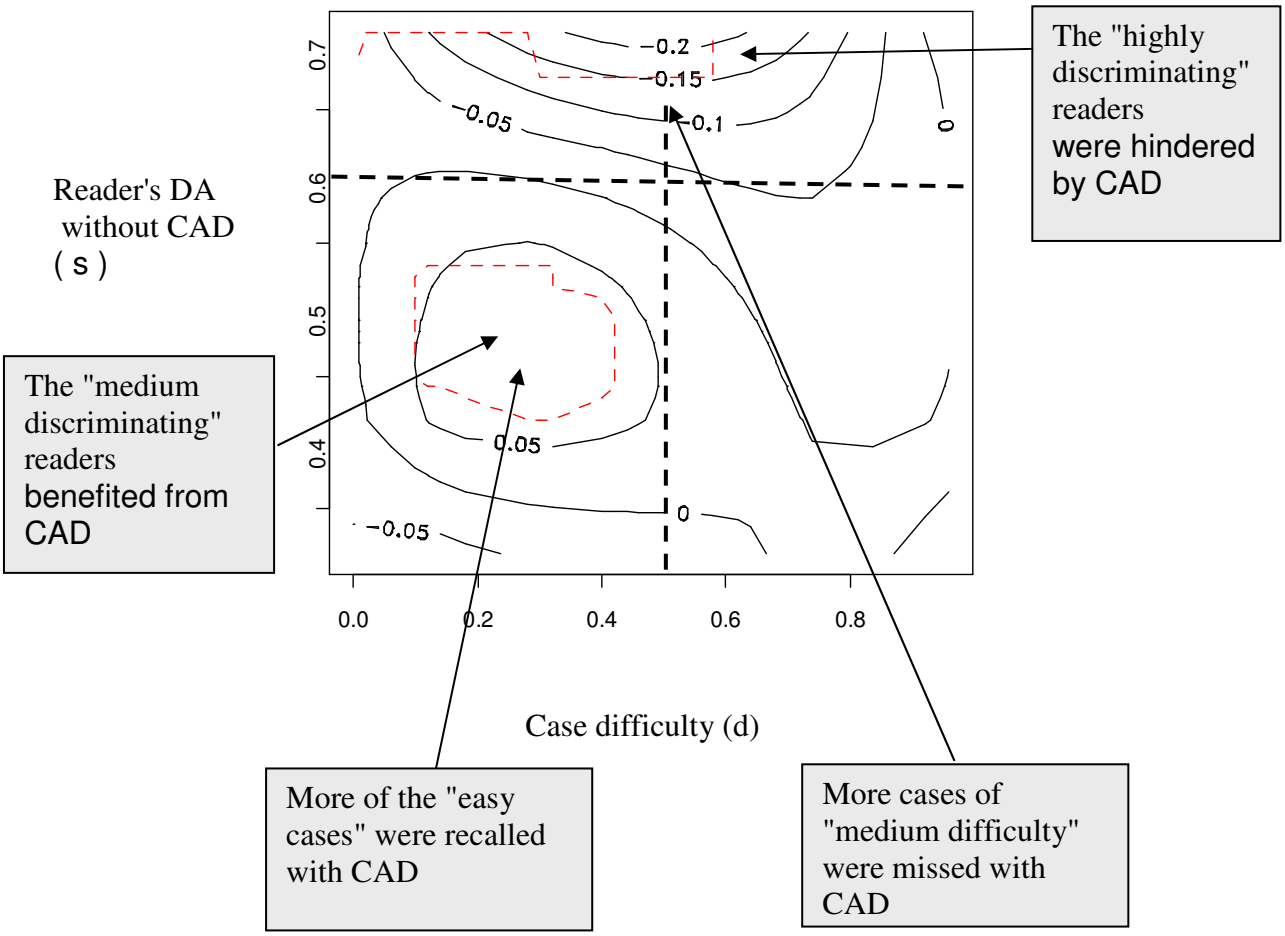

Figure 4.

Contour plot of estimated impact of CAD $imp(d, s)$ for all cancers. Here and in the following figures, we consider the impact significant if the 95% pointwise confidence intervals for $p_{un}(d_i,s_j)$ and $p_{pr}(d_i,s_j)$ do not overlap; statistically significant values of impact are outlined with the dashed curve. Values for the points with extreme impact are given in Table 2.

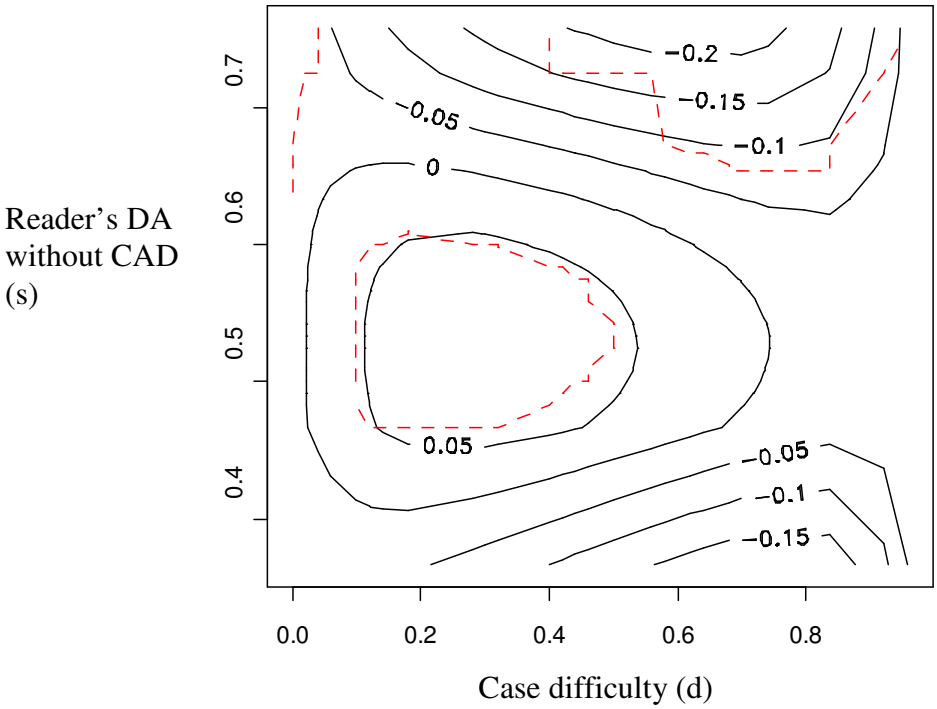

Figure 5.

Contour plot of impact of CAD: *imp(d, s)* for all cancers, estimated with logistic models for paired outcomes: "decision is aided by CAD" and "decision is hindered by CAD". See Figure 4 for an explanation of significance values.

Table 1 Different categories of cancer cases.

	Cancers correctly prompted by CAD	cancers not correctly prompted by CAD	Total
Screen detected cancers, i.e. cancers detected through routine screening	36	4	40
Interval cancers, i.e. cancers missed during routine screening	9	11	20
Total	45	15	60

Table 2. Extremes of estimated impact *Imp(d,s)* of CAD by different categories of cancers.

Category of cases	Sensitivity of CAD	Type of extreme of impact	Imp(d,s)	s	d	p_{un}(d, s)	p_{pr}(d, s)
All cases (n=60)	0.750 (0.621,0.853) [45 correctly prompted cases]	Min	-0.225	0.758	0.500	0.735(0.603, 0.835)	0.510 (0.480, 0.540)
		Max	0.087	0.510	0.280	0.663(0.622, 0.701)	0.750 (0.722,0.775)
Correctly prompted by CAD (n=45)	1.000	Min	-0.428	0.758	0.460	0.705 (0.499, 0.852)	0.277 (0.145, 0.464)
		Max	0.107	0.533	0.300	0.654 (0.607, 0.698)	0.761 (0.725, 0.794)
Incorrectly prompted by CAD (n=15)	0.000	Min	-0.160	0.672	0.620	0.467 (0.388, 0.547)	0.306 (0.238, 0.384)
		Max	not signif	0.633	0.920	0.068 (0.041, 0.109)	0.029 (0.015, 0.053)
Interval cancers (n=20)	0.450 (0.231, 0.685) [9 correctly prompted cases]	Min	-0.316	0.367	0.920	0.372 (0.146, 0.674)	0.057 (0.040, 0.080)
		Max	0.268	0.367	0.100	0.657 (0.378, 0.858)	0.925 (0.893, 0.958)
Screen detected cancers (n=40)	0.900 (0.763, 0.972) [36 correctly prompted cases]	Min	-0.303	0.367	0.122	0.910 (0.871, 0.938)	0.607 (0.408, 0.776)
		Max	0.293	0.367	0.700	0.329 (0.244, 0.427)	0.621 (0.450, 0.767)

Comments: $p_{un}(d, s)$, $p_{pr}(d, s)$ are the estimated probabilities of a reader with DA *s* recalling a case of difficulty d in the unprompted condition and prompted condition respectively; $Imp(d, s) = p_{pr}(d, s) - p_{un}(d, s)$ is the estimated impact of CAD; 95% Confidence intervals for the probabilities are given in brackets;

Table 3 : Sensitivity of readers (without CAD) and CAD (alone) for *a posteriori* classification of readers and cases

	Easy cases (d < 0.5, 45 cases)	Difficult cases (d ≥ 0.5, 15 cases)
Sensitivity of less discriminating readers without CAD (s < 0.65, 44 readers)	0.898 (binomial 95%CI: 0.884, 0.911) 1776 correct decisions out of 1977 known	0.293 (binomial 95%CI: 0.258, 0.328) 193 correct decisions out of 659 known
Sensitivity of highly discriminating readers without CAD (s ≥ 0.65, 6 readers)	0.922 (binomial 95%CI: 0.883, 0.951) 248 correct decisions out of 269 known	0.404 (binomial 95%CI: 0.301, 0.513) 36 correct decisions out of 89 known
Sensitivity of CAD tool	0.822 (binomial 95%CI: 0.679, 0.920) 37 of 45 cases were correctly prompted	0.267 (binomial 95%CI: 0.078, 0.551) 4 of 15 cases were correctly prompted.

CONFLICT OF INTERESTS STATEMENT

The authors declare that they have no competing financial interests. The corresponding author (A.P) had full access to all the data in the study and had final responsibility for the decision to submit for publication.

ROLE OF THE FUNDING SOURCE

This work was supported in part by: the UK Engineering and Physical Sciences Research Council (EPSRC) through projects DIRC (the Dependability Interdisciplinary Research Collaboration) and INDEED ("Interdisciplinary Design and Evaluation of Dependability", EP/E000517/1) as well as by Cancer Research UK (grant GR/C22515/A7339).

Correspondence and requests for materials should be addressed to A.P (andrey@csr.city.ac.uk)