Does incorrect computer prompting affect human decision making? A case study in mammography

E. Alberdi[a], A. Povyakalo[a], L. Strigini[a] and P. Ayton[b]

[a]Centre for Software Reliability and [b]Department of Psychology, City University, Northampton Square, London, EC1V 0HB, United Kingdom

Corresponding author: Eugenio Alberdi - Email: e.alberdi@csr.city.ac.uk,
Tel: +44 20 7040 8424 - Fax: +44 20 7040 8585

Abstract

The goal of the data collection and analyses described in this paper was to investigate the effects of incorrect output from a CAD tool on the reliability of the decisions of its human users. Our work follows on a clinical trial that evaluated the impact of introducing a computerised prompting tool (R2 ImageChecker) as part of the breast screening programme in the UK. Our goal was to use data obtained in this trial to feed into probabilistic models (similar to those used in "reliability engineering") which would allow us to find and assess possible ways of improving the interaction between the automated tool and its human user. A crucial requirement for this modelling approach is estimating the probability that a human will fail in his/her task when the output of the automated tool is incorrect. The data obtained from the clinical trial was not sufficient to inform this aspect of our probabilistic model. Therefore, we conducted a follow-up study to elucidate further the effects of computer failure on human performance. Preliminary analyses of the data resulting from the follow-up study are reported and discussed.

Keywords: CAD, mammography, evaluation

1. INTRODUCTION

"Reliability modelling" in engineering is concerned with predicting probabilities of future failures[1] (or failure-free behaviour) of man-made systems. It most typically involves building a "model" which gives a mathematical specification of how failures of its components may cause the failure of the system (specifically, which combinations of component failures would lead to system failure) [1]. These models bring insight about which components are most critical for the dependability of the whole system, about which component is it best to spend effort to improve it, and sometimes may allow one to estimate the dependability measures for

[1] We will use "failure" to mean any case in which something (a machine or a human or a set of co-operating humans and machines) produces an incorrect output.

the whole system, like the probability of the system behaving improperly, from the known reliability levels of the individual components.

In our current work, we are applying this approach to the situation in which a human uses the output of a computer aid to issue a decision (e.g., a diagnosis, a suggestion for treatment, etc.) [2]. We see the human and the automated aid (which we will call "the machine" for brevity) as parts of a single system. The use of computer aids in breast screening is a good example of this sort of "human-machine system".

We use data from a clinical trial [3] which was run by researchers in University College London, UCL (in conjunction with various hospitals and breast screening centres in SE England), to assess the potential benefit of introducing a particular CAD tool, R2 ImageChecker [4], as part of the breast screening programme in the UK. The trial used a sizeable set of mammography cases and was conducted with a relatively large number of practitioners; it generated extensive empirical evidence of potential use to inform our probabilistic model. However there are features of our modelling approach for which the data provided by this trial are not sufficient. As a result, we felt the need for an additional experiment to complement those data.

2. METHODS

We outline first the data collection methods used by the researchers from UCL in their clinical trial, since we used essentially the same methodology in our follow-up study. The UCL trial was run with 50 practitioners (radiologists, radiographers and breast clinicians) and used 180 cases distributed in 3 sets of 60.

All the film readers saw all the cases in two different experimental conditions: a) "unprompted condition": without computer support (similarly to how they do it in everyday practice); b) "prompted condition": with the aid of the CAD tool. This was randomised across the participants.

In both the prompted and unprompted conditions, the readers saw two versions of each case: 1) the actual films, positioned on a standard viewing roller; and 2) a digitised version of the mammograms printed out on paper; in the prompted condition, the printouts contained the prompts generated by the CAD tool.

The readers were instructed to mark relevant mammographic features on the printouts and grade the level of risk associated with the features. Readers were also asked to make their decision on recall as if they were viewing the films as single readers in the screening programme. See more details in [3].

Researchers at UCL carried out statistical analyses comparing the sensitivity and specificity of the readers in the unprompted condition with their sensitivity and specificity in the prompted condition. The analyses showed that the prompts had no significant impact on (neither improved nor diminished) readers' sensitivity and specificity [3].

We were granted access to the trial data and conducted supplementary analyses focusing on the instances in which the readers made different decisions for the same case depending on whether they saw it with the aid of CAD ("prompting condition") or without CAD support ("unprompted condition").

Table 1. Composition of the data set used in the follow-up study

	Correctly marked (N=10)	'Mismarked' (N=23)	Unmarked (N=27)
Cancer (N=30)	**10** (11-13)	**11** (2-5)	**9** (1-2)
Normal (N=30)	**n/a**	**12** (29-34)	**18** (6-11)

Note. The numbers in bold type correspond to the actual distributions in the new data set. The intervals shown within brackets correspond to the ranges in the data sets used in the clinical trial.

In our follow-up study, we were interested in the humans' responses to a particular type of computer failure, namely, the cancers that were "missed" by the prompting tool. There are two ways in which CAD can "miss" a cancer: a) by failing to place any prompt on the mammogram ("unmarked mammogram"); b) by placing a prompt in an area of the mammogram *away* from the area where the actual cancer is located ("mis-marked mammogram").

The study was conducted with a subset (20) of the 50 readers who participated in the original trial. We used a new data set with similar characteristics to the data sets used in the original trial except that it contained a higher proportion of "missed" cancers. Sixty sets of mammograms were used in the follow-up study. They were provided by the developers of R2 ImageChecker [4], who picked them from a repository of cases used to train the image processing algorithms used by the computer tool. The diagnosis for all the cases was known, proven either by biopsy or by a clear result in a subsequent round of screening. The output of the CAD tool was also known and recorded for every case. After the start of the clinical trial at least two new versions of the image processing algorithms were developed by the manufacturers of R2 ImageChecker; we ensured that the cases in our follow-up study were processed by the same version of the image processing algorithms that was used to process the cases in the clinical trial.

The cases were selected to meet two criteria: 1) to contain a large proportion of cancers "missed" by CAD, as defined above; but 2) to resemble as much as possible the test sets used in the original trial. We wanted the readers to perceive our study as an extension to the original trial and to behave in a comparable way. So we tried and mask the fact that the CAD tool was being much less sensitive than in the original trial as this might cause them to change their own threshold (and their ways of using the machine prompts) to compensate. Table 1 summarises the composition of the data set we used in the follow-up study.

We tried to keep as many case characteristics as possible similar to those in the original data sets. We gave priority to the following parameters:

- The *number of cases* (N=60): the goal was to present a data set of the same size as each of the sets in the original trial, so that the participants would see it as a "natural" follow-up to the trial.
- The *number of cancers*: in the original data sets this number was already artificially high: it ranged between 15 and 25 out of the 60 cases in each set; we decided to go for the higher end of the range of cancers in the original data sets (25) and decided we

would add just a few more (5) hoping that this difference would not be obvious to the participants.

- The *specificity* of CAD for our data set is 27%, which is within the range of the original data sets (23%-41%).
- The *number of cancers correctly marked* in our data set it is 10, just below the lower end of the range of the original data sets (11-13) (although, obviously, the "proportion" of correctly marked cancers out of the total number of cancers, i.e., the sensitivity, is lower).

In the original trial data sets, the average of cancers missed by CAD is around 5 (between 27% and 35% of the cancers). For the new data set, we decided to increase this number as much as possible within the constraints noted above. The figure we ended up with is 20 (66% of the cancers in our set).

The reading procedures were essentially the same as in the previous sessions in the clinical trial. One important difference, however was that all the readers went through one reading session only, viewing all the cases with computer support (i.e., they saw the cases in the "prompted condition" only). Otherwise, the instructions given to the participants were the same as in the clinical trial.

In addition, in both studies, the readers filled in a series of questionnaires which enquired about various issues related to CAD, breast screening practice and the characteristics of the cases used in the studies.

3. RESULTS

Our supplementary analyses of the original trial data showed that variations in recall decision (between the "unprompted condition" and the "prompted condition" for the same case) occurred for about 28% of the readers' decisions. The most striking pattern in these variations was the following: if a reader was not confident about his/her decision for a "normal" case (non malignant) in the "unprompted condition", then, when viewing the same case in the "prompted condition", he/she was likely to provide a more correct decision than when viewing it in the "unprompted condition". In other words, for "non obvious" normal cases, the readers tended to perform better during the prompted condition than during the unprompted condition. This data pattern was statistically significant: this systematic decision variation was not observed for the cases with cancer.

As regards our follow-up study, we looked at whether the recall decisions generated by the readers were correct (i.e. they recommended recall for cancers and no recall for normal cases) or not. We found that 73% of all the decisions were correct. However, the proportion of correct human decisions varied greatly depending on whether a case was normal or cancer and whether it had been correctly prompted by the CAD tool or missed. Very few (21%) of the decisions generated for "unmarked" cancers and only over half (53%) of the decisions for the "mis-marked" cancers were correct; in contrast with the very high proportions of correct

answers for the normal cases (92%-94%) and, to a lesser extent, the correctly prompted cancers (81%).

Table 2. Summary of "majority" recall decisions

	Correctly marked (N=10)	'Mis-marked' (N=23)	Unmarked (N=27)
Cancer (N=30)	7/10	5/11	1/9
Normal (N=30)	**n/a**	12/12	16/18

Note: The numerator in each fraction indicates the number of cases in each category for which a "majority" of readers provided the "correct" recall decision out of the total number of cases in each category (the denominator). "Majority" is defined as 80% (16 out of 20) or more of the readers.

We were also interested to see to what extent the different readers agreed in their decisions for each case. Not surprisingly, we found few cases, only 12 (20%), for which there was unanimity amongst the readers. Table 2 shows the number of cases (grouped according to computer output) for which a "majority" of the readers provided a "correct" recall decision. Again, there was a very low proportion of cancers with no machine prompts (only one) that the majority of the readers chose to recall. More strikingly, the remaining seven unmarked cancers elicited an "incorrect" recall decision by 18 or more of the 20 participants; further, for two of those cases, the "incorrect" decision was unanimous: all readers chose not to recall them.

Additionally, the average sensitivity of the human readers for the cases in the follow-up study was 52% (minimum: 27%; maximum:70%), in contrast with the average 85% sensitivity in the "prompted condition" of the original trial.

These data patterns are supported by some of the responses to the questionnaires. Some of the readers did not seem to be aware of the large proportion of cancers that were missed by the CAD tool in the follow-up study. For example, at least three of the participants stated that, in the test set used in the follow-up study, there was a smaller number of cancers than in the original trial. In fact, the data set in the follow-up study contained more cancers than any of the sets in the clinical trial. The difference was that our data set contained many more "unmarked" cancers (9 vs. 1or 2 in each of the original sets).

4. DISCUSSION

Our supplementary analyses of the data from the clinical trial suggest that prompting did have an effect on the readers' decision making even if there is no statistically significant evidence that it affected their performance in terms of sensitivity and specificity. We cannot exclude the possibility of random error (e.g., it is not rare that experts change their decisions in successive presentations of the same cases). However our analyses strongly suggest that prompting might have been used by the readers as a sort of reassurance for their "no recall" decisions for normal cases [5]. We believe this possible side effect of CAD use was not anticipated by the manufacturers of the tool, who essentially designed it to assist with cancers.

To a great extent, this is consistent with the results from our follow-up study. One plausible explanation for these results is that the readers were misled by the prompting system, that is, the readers tended to assume, based on past experience with the tool, that the absence of prompting was a strong indication that a case is normal, thus they paid less attention than necessary to those cases with no prompts on. An alternative explanation is that these cases had characteristics that made them particularly difficult (perhaps undetectable mammographically) for both the human readers and the computer tool. Unfortunately, our study was not designed to study this phenomenon; because the readers saw each case only once (with CAD support), we cannot know how they would have behaved had they seen the cases without computer support. We cannot elucidate whether the readers failed to notice those cancers because of the misleading influence of CAD or because of the intrinsic difficulty of the cases or a combination of the two. To clarify this, we are designing an additional data collection session, which we hope will serve as a "control" condition to our study.

We believe the unexpected effects of prompting conjectured above (if corroborated by future data) would have serious implications for the design and use of CAD in general. Additionally, we are currently using the data collected so far to inform the parameters in our probabilistic models [2], which we hope will provide further insights about the reliability of computer assisted decision making in mammography and other domains.

ACKNOWLEDGEMENTS

The work described in this paper has been partly funded by UK's Engineering and Physical Sciences Research Council (EPSRC). We would like to thank R2 Technologies (and very especially Gek Lim) for their support in obtaining the data samples for the follow-up study; Paul Taylor and Jo Champness, from UCL, for facilitating the follow-up study and helping run it; and DIRC collaborators Mark Hartswood, Rob Procter and Mark Rouncefield for their collaboration and advice.

REFERENCES

1. Littlewood B, Popov P, Strigini L. Modelling software design diversity - a review. ACM Computing Surveys 2001; 33(2): 177-208.
2. Strigini L, Povyakalo A, Alberdi E. Human-machine diversity in the use of computerised advisory systems: a case study. To appear in: Proceedings of DSN (Dependable Systems and Networks), San Francisco, CA, 2003.
3. Champness J, Taylor P, Given-Wilson R. Impact of computer-placed prompts on sensitivity and specificity with different groups of mammographic film readers. In: Peitgen H-O, editor. Proceedings of the 6th International Workshop on Digital Mammography. Springer-Verlag, 2002.
4. http://www.r2tech.com
5. Hartswood M, Procter R, Williams J, Prescott R, Dixon P. (1996). Subjective responses to prompting in screening mammography. MIUA-96.