



## Combining qualitative and quantitative methods in a study of computer-assisted medical decisions

**Mark Hartswood**

School of informatics  
University of Edinburgh

**Lorenzo Strigini**

Centre for Software Reliability  
City University, London



[www.dirc.org.uk](http://www.dirc.org.uk)

# Overview of talk

Introduce the context

- Breast screening
- Alerting systems to assist detection of cancers

Discuss findings taken from

- ethnography
- experiments and probabilistic modelling

Draw some methodological conclusions

# Screening mammography

- More than 40,000 cases of breast cancer are diagnosed in the UK every year
- Women between the ages of 50 and 70 are invited for periodic screening
- Mammography is the principal screening test for breast cancer
  - One or more breast X-Rays, or mammograms, are taken of each breast
  - These are examined by one or more film readers for signs of cancer
- Roughly 5% of women screened are invited back for further tests: ~10% of whom will actually be found to have cancer

# Decision-aids in mammography

## Computer-aided Detection (CAD)

- Use image analysis software to detect potential abnormalities
- Draw these to the reader's attention using a 'prompt'
- Designed to prevent readers from overlooking a possible abnormality
- Has a number of potential roles:
  - Making screening more sensitive
  - Supporting single reading
  - Supporting less experienced reader

CAD tends to be reasonably sensitive, but not very specific

- Abnormalities are detected with a high probability (e.g. 90% of targeted cancers identified)
- At the expense of a large number of false alarms (e.g. 2 FP per case)

# Evaluating CAD

## Reasons for evaluating CAD systems

- For policy makers: is CAD effective – to what degree?
  - In-vitro studies show that CAD can detect cancers that readers miss
  - In-vivo the evidence of effectiveness is mixed
- How does CAD impact on readers' decision-making?
- How can CAD be best exploited (training, procedures)?
- Feedback to CAD tool design

## Problems of evaluating CAD systems:

- The small 'base rate' of cancers
- Variability between the performance of individual readers
- Providing a realistic environment for testing
- Changing character of machine and screening practice

## Ethnographic studies of screening practice and CAD use

Have been doing this in Edinburgh for about 10 years

- initially with system called PROMAM – developed in Edinburgh
- most recently been involved in the evaluation of a US system – R2 “ImageChecker” M1000

Ethnographic studies of

- screening practice
- use of CAD in experimental settings (‘think aloud’ protocol)
- Detailed observation: video, note-taking, collecting examples of documents, interviews

# Biographical familiarity

Readers' understandings of:

- own and others' abilities
- equipment and procedures

Formal and informal feedback

- Double reading
- Assessment clinics
- Multidisciplinary meetings
- Formal assessments (e.g. PERFORMS)

## Biographical familiarity with CAD

Making sense of what the system does:

- Prompts are accountable
- Sometimes prompts are found to be confusing
- Gaining a sense of what CAD is good at and where its deficiencies lie
- Sometimes making incorrect inferences about CAD's capabilities
- In the experimental settings we have studied there has been no feedback
- No contact with developers

## How CAD is used

- The idea is for prompting systems to act as attention cues
- Look at the images and reach own conclusion before looking at the prompts
- However, we saw evidence of prompts being used as decision-aids:

“I’m not really that worried about it. [At all?].  
But as CAD’s marked it now, it’s a case of – do I really take more notice of it? ... I’ll mark it. I’m going to mark it down - as possibly being something.”  
(transcript from video)

# The case study in the DIRC project

## Inputs:

- a large study conducted by UCL (CHIME) in collaboration with various hospitals for the Health and Technology Assessment (HTA) programme, on a specific CAD tool
- DIRC interdisciplinary team

## The HTA experiment (UCL CHIME)

- Using mammograms of past cases with known diagnosis
- 50 readers examined 180 cases
  - 120 normal cases, 60 cancers
  - in two conditions:
    - + without CAD (*unprompted session*)
    - + with CAD (*prompted session*)
  - for "recall / do not recall" decisions
- standard statistical analyses of the decisions revealed *no significant difference* between reader performance with and without CAD
  - did readers ignore the tool's prompts?
  - artefact of experimental setup?
    - + very good readers?
    - + different consequence of errors?
    - + lack of fatigue/boredom?
    - + Hawthorne effect?

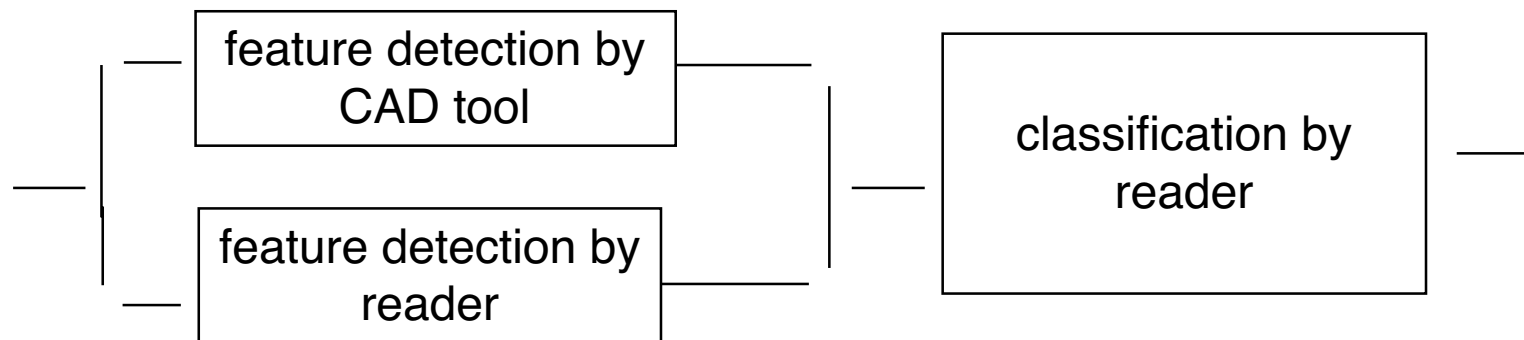
## **Additional work by the DIRC team**

- direct observation
  - probabilistic modelling
  - additional experiments
  - exploratory statistical analyses
- 
- largely in parallel
  - cues from each feeding into the others

# Probabilistic modelling of the system (CAD tool + reader)

In the style of reliability engineering:

- how do the (probabilistic) failure characteristics of the components affect those of the whole system?
- a method for “what if” reasoning, bounds on uncertainty
- makes assumptions explicit
  - CAD tool + reader meant as a *fault-tolerant system*
  - the tool ought to compensate for human FN errors in *detection*, *without ever harming* human decisions.



Some studies take this as an *assumption*.... which we discarded

- psychologically doubtful
- direct evidence from ethnography

## Probabilistic modelling for insight ...

If  $p(x)$  is probability of input case  $x$

then for the next randomly chosen input case:

$$P_{Hf} = \sum_x p(x) [P_{Hf|Ms}(x)P_{Ms}(x) + P_{Hf|Mf}(x)P_{Mf}(x)] =$$

$$\sum_x p(x) (P_{Hf|Ms}(x) + P_{Mf}(x) t(x)) =$$

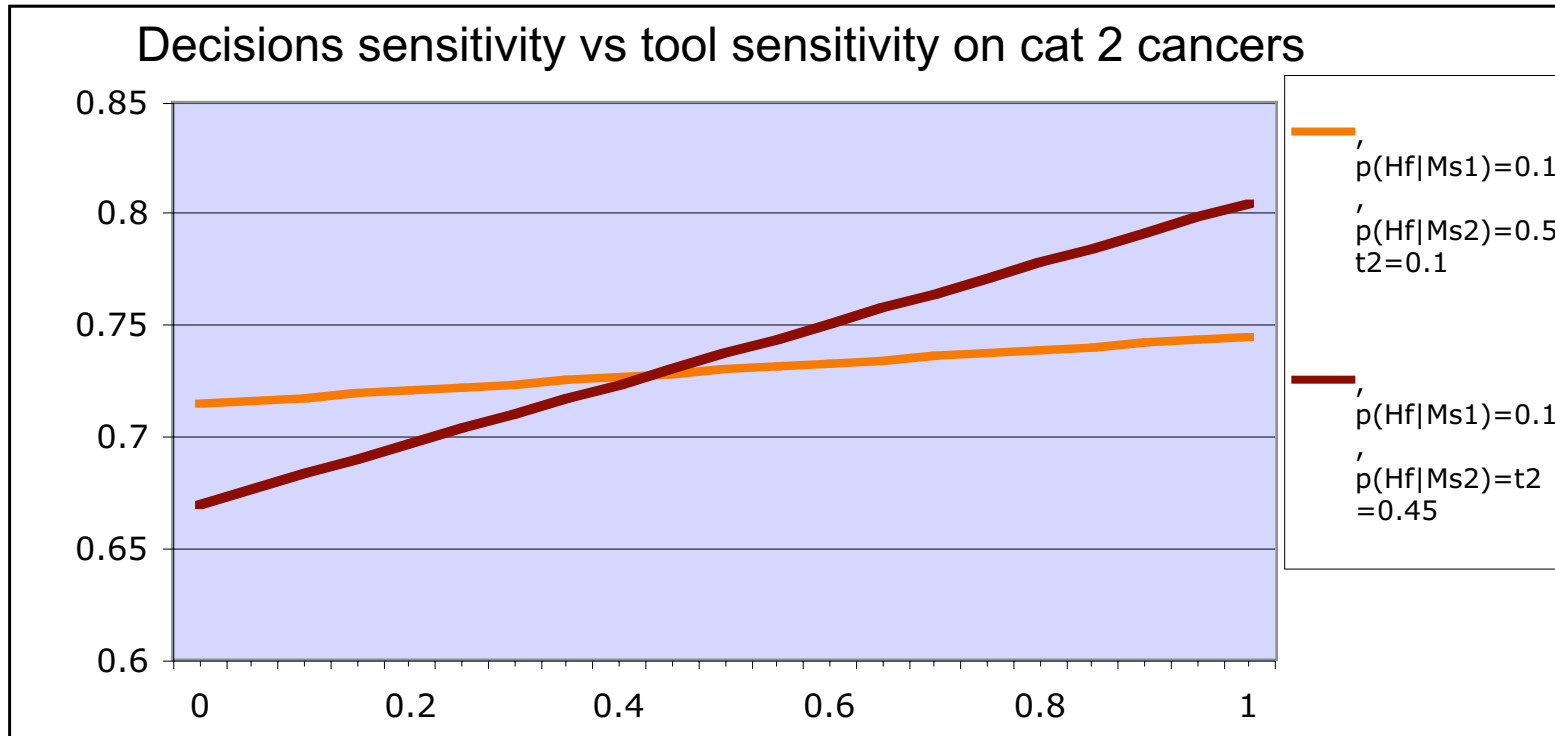
"Importance index" (of CAD tool for system)

$$E[P_{Hf/Ms}(x)] + E[P_{Mf}(x)] E[t(x)] + \text{cov}_x(P_{Mf}(x), t(x))$$

note role of *covariance*:  
e.g. negative covariance means machine more reliable in cases when its support affects human more

*2 important aspects of CAD tool: intrinsic reliability, diversity from human*

# "What if" reasoning, e.g. ...



## Supplementary experiments

A parameter in the models is the probability of reader FN error, *given* wrong “advice” from the tool:

- FN errors by tool too rare in original sample
- new study: 20 readers, with CAD, high rate of "wrong advice"
- so many FN decisions that we recruited a “control” group to test without CAD
- results are intriguing
  - FN errors *increased* with CAD, especially with wrong prompts
  - FP errors *decreased* with CAD

**..... did the tool's FN errors *cause* human errors?**

## Data from the supplementary experiments

CAD Output	% “Correct” Decisions			
	Cancers		Normals	
	CAD/No CAD	CAD/No CAD	CAD/No CAD	CAD/No CAD
Correct Prompts:	81%	90%	n/a	n/a
Incorrect Prompts:	<u>53%</u>	66%	92%	87%
No Prompts:	<u>21%</u>	<u>46%</u>	<u>94%</u>	<u>88%</u>

## Exploratory analysis on HTA trial data

looking for *average* effect can mask variations of CAD effect between *readers* and between *cases*

- since average effect was effectively 0
  - either “CAD can do no harm”: then it was no good either
  - or it *can* harm some decisions, improve others: then
    - + which ones? How?
    - + what will be the balance in real use?

- new statistical analyses to tease this out

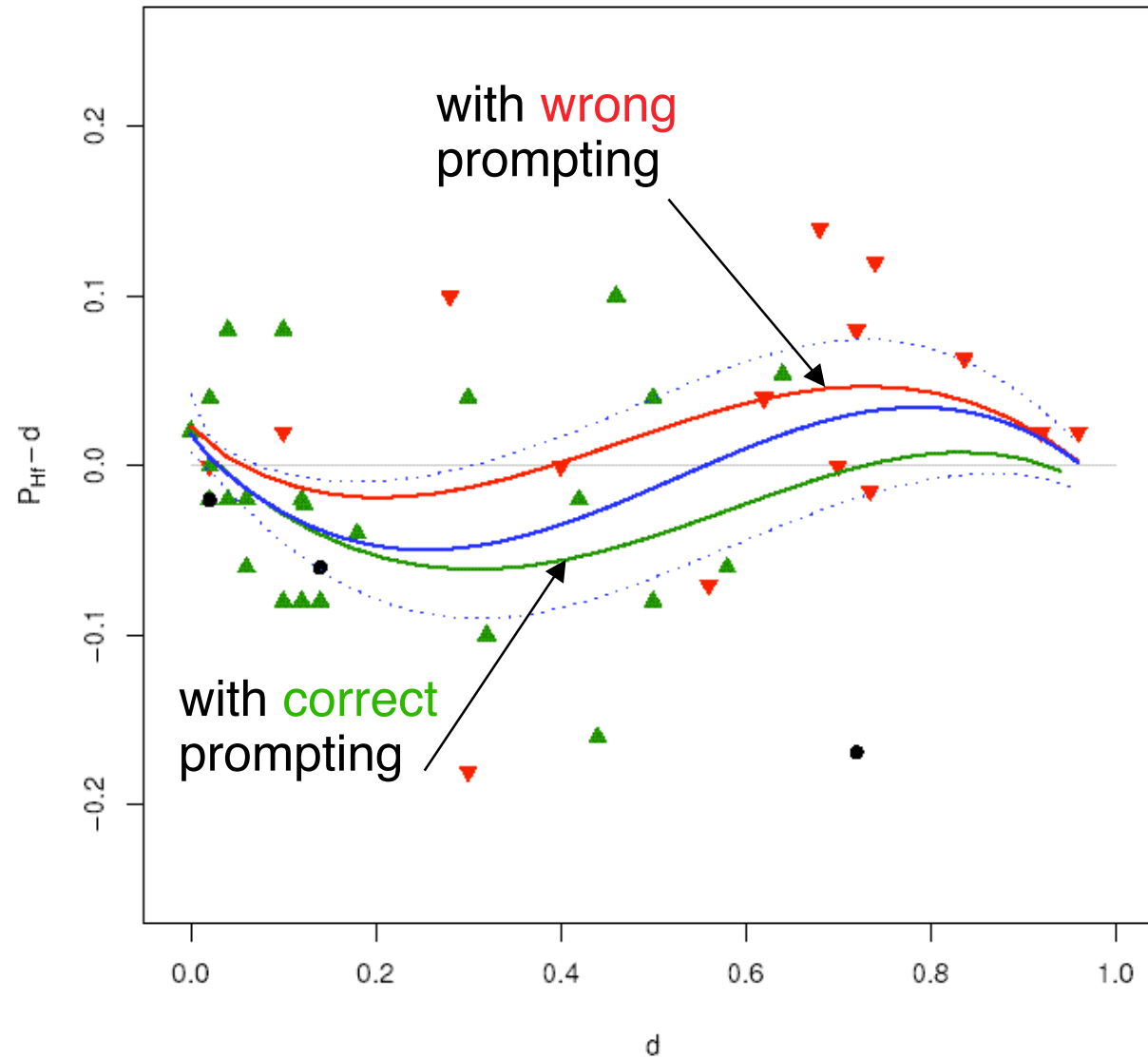
*notes:*

- *the data are “noisy”: readers appear often to “change their minds”*
- *we are using data collected for a different purpose*  
*e.g., no separate indication of unaided reader's “natural noise”*

# Example of exploratory analysis

46 non-obvious cancers. 50 readers.

Regression to filter out noise:  
difference between error rates of prompted & unprompted readers as function of unprompted "difficulty" and **correct** or **wrong** prompting



## Some results of this study

For the evaluation of this CAD tool:

- evidence that it did affect how readers decide, and **variety of these effects**, e.g.
  - + worse FN rates in the better (more sensitive) readers on difficult cancers
  - + better FP rates in the worse (less specific) readers
- support for reasoning about
  - effectiveness in real use
  - improving design
    - + e.g. target “diversity” rather than absolute error rates of tool

About similar decision support systems

- contribution to “automation bias” literature
- conjectures about processes underlying effect of CAD
  - + e.g. specific changes of strategy on “uncertain” cases?

*More work needed*

## Integrating methods *worked*

- some insight about questions for which pure controlled trials would be infeasible
- methods for *peering through the black box* of standard controlled studies
  - direct observation
    - + point out things that happen (how tools are used), though not frequency, effect on overall performance
  - models of the causal structure of the system
    - + explicit, formal representation of factors that may affect outcome
    - + means for integrating evidence, highlighting critical issues
    - + support with the problems of estimating effects *in the field* using statistics *from a trial*: representing effects of conjectured changes
  - focused, partially unrealistic experiments
  - exploratory statistical analysis of data *as available*
    - + challenge conjectures on what happened in a specific experiment, produce conjectures of general patterns

familiar approach in engineering science

## For more information

<http://www.dirc.org.uk/research/DIRC-Results/MammographyCity.html>

<http://www.csr.city.ac.uk/people/eugenio/dirc-mammo.html>

DIRC mammography case study team:

Eugenio Alberdi <[e.alberdi@csr.city.ac.uk](mailto:e.alberdi@csr.city.ac.uk)>

Peter Ayton <[p.ayton@city.ac.uk](mailto:p.ayton@city.ac.uk)>

Mark Hartswood <[mjh@inf.ed.ac.uk](mailto:mjh@inf.ed.ac.uk)>

Mark Rouncefield <[m.rouncefield@lancaster.ac.uk](mailto:m.rouncefield@lancaster.ac.uk)>

Roger Slack <[rslack@inf.ed.ac.uk](mailto:rslack@inf.ed.ac.uk)>

Andrey Povyakalo <[andrey@csr.city.ac.uk](mailto:andrey@csr.city.ac.uk)>

Rob Procter <[rnp@inf.ed.ac.uk](mailto:rnp@inf.ed.ac.uk)>

Lorenzo Strigini <[strigini@csr.city.ac.uk](mailto:strigini@csr.city.ac.uk)>