City University London

CSR
Centre for Software Reliability

Computer assistance, diversity and dependability

Lorenzo Strigini
report of work with Andrey Povyakalo, Eugenio Alberdi, Peter Ayton
City University, London
Rob Procter, Mark Hartswood (U. of Edinburgh), Mark Rouncefeld (U. of Lancaster)

DIRC the Interdisciplinary Research Collaboration
on the Dependability of computer-based systems

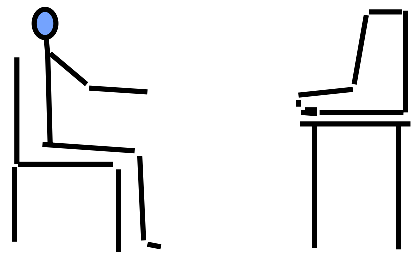
Outline of presentation

warning:
not really about interfaces,
and not just about evaluation

- human-computer systems as fault-tolerant systems, relevance of some known models
- outline of a case study
  - insight from modelling
  - analysis of empirical data
- implications for this system, and category of systems

Redundancy/diversity in human-machine systems

- example: simple two-component system

User  computerized advisory system

- each detects/corrects <u>some</u> errors of the other
- details vary between systems

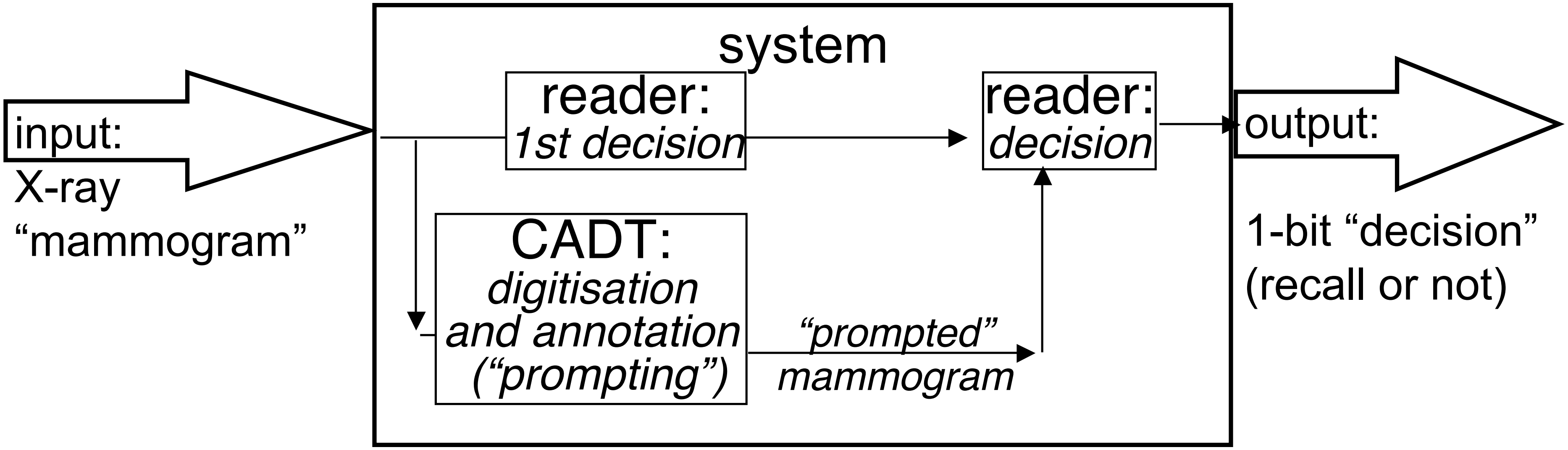
Redundancy/diversity potential not always noticed, exploited in design and assessment

- *even highly reliable* human *-only performance normally relies on fault tolerance*
- *common mode, un-recovered failures matter*
- *we in this community know a few tricks for reasoning about them - how useful are they?*

A case study: Computer Aided Detection in reading X-rays

- system: a clinician assisted by a computer
- “mammography”: X-rays “read” by [usually two] “readers” (scarce, highly skilled specialists)
- “computer-assisted detection” machines exist
  - to help readers not to miss image “features” that may indicate cancer
- some studies indicate this is an effective aid
- being evaluated for use in U.K. screening programme for breast cancer detection
  - controlled trial led by University College, London
- additional DIRC study by City University, U. of Edinburgh, U. of Lancaster

System: human *reader* with **C**omputer **A**ided **D**etection **T**ool

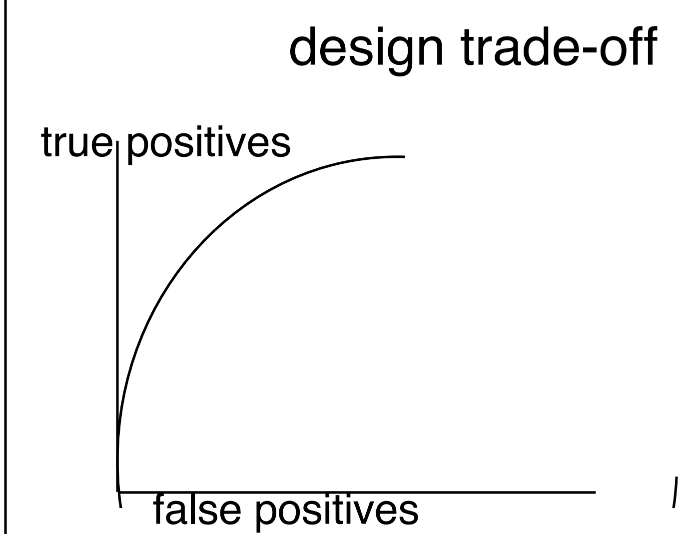


procedure for readers:

- check mammogram, select visual features to diagnose
- *then* see whether CADT "prompted" more features
- examine all and decide (autonomously)

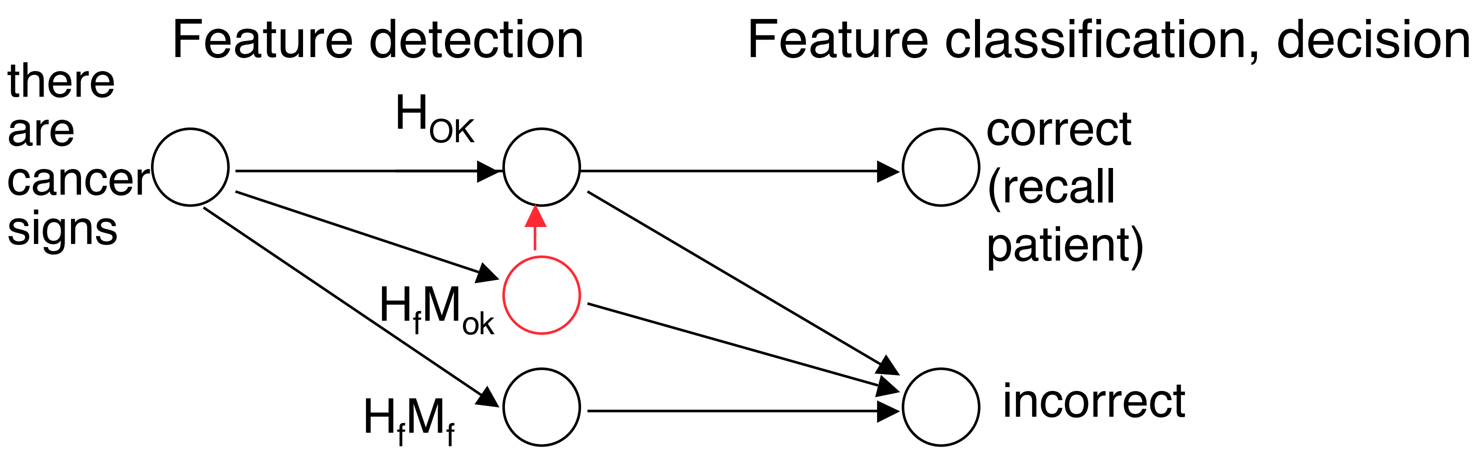
Two kinds of system failures

- false negative (dangerous):
  not recalling patient with cancer
- false positive (painful, expensive):
  recalling healthy patient
- CAD meant to reduce false negatives (FNs)
  - without excessive increase in false positive
- similar definitions for failures of subsystems, except:
  - human has special role: final decision
  - CADT is tuned to produce few FNs at the cost of many FPs (too many prompts)



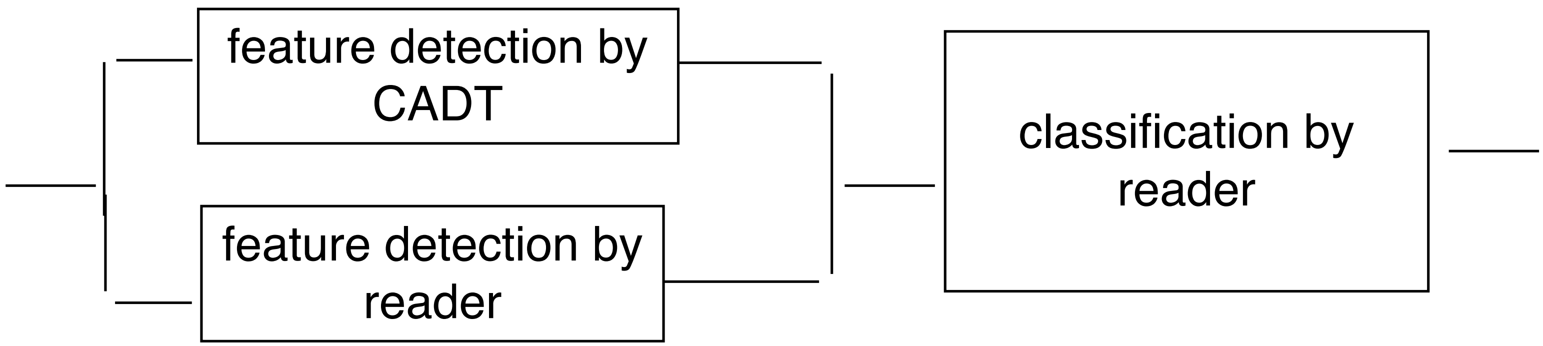
Design rationale

- The CADT provides fault tolerance for users' *FN errors of detection*:
  - it only matters if the reader missed some "interesting" feature of the image
  - what matters is the coverage of its error detection function, and of recovery function by human



Ideally...

Reliability block diagram (for FN failures) is:



notice that CADT "can do no harm"

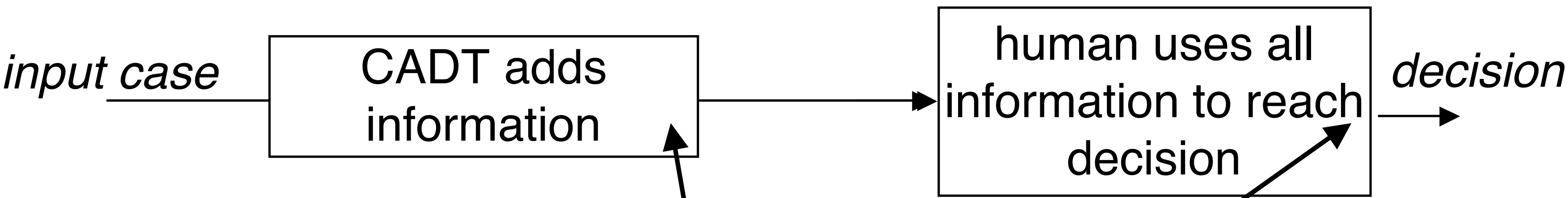
common assumption about *warning* systems

However,

- can we use this model and “can do no harm” assumption?
  - readers’ task not neatly divided into 2 stages
    - + intermediate results not observable
  - presence of CADT might change how reader behaves before using CADT output
    - + e.g. readers might grow dependent on CADT
    - + and use of CADT perhaps not as intended
  - plus some extra modelling difficulties in describing failure correlation via conditional independence

- clinical trial analysis rightly avoided this assumption
  - simple black box comparison of human performance in “prompted” vs “unprompted” condition
  - but did it therefore ignore useful information?
  - we used a clear[er] box analysis

A valid model, and its parameters



- probability of failure here, $\boldsymbol{P_{Mf}}$
- probability of failure here
  - –given correct advice by CADT: $\boldsymbol{P_{Hf\,|Ms}}$
  - –given failure of CADT: $\boldsymbol{P_{Hf\,|Mf}}$

all these depend on the input case (how “difficult”)?

("diversity modelling" approach)

Probability of system failure

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

then for the next randomly chosen input case:

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

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

"Importance index" (of CADT for system)

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

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

2 aspects of CADT: intrinsic reliability, diversity from human

Using the model for dependability prediction

- estimating dependability ***in the field*** using statistics ***from a trial*** is problematic
- many factors may change
  - distribution of patient types, readers
  - adaptation to CADT, developing over time

    mistrust and/or overreliance, workload, other unexpected effects?
  - upgrades of CADT
- with model, conjectured changes can be represented explicitly to study ranges of effects

  a motherboard for plugging in diverse knowledge

  + failure statistics, observation/questioning of readers
  + "human" and "engineering" disciplines
  + specific and general knowledge

  into a common, formal picture

Some results from observation of readers

- readers need mental model of CADT operation
  - "why did it mark this?", "why didn't it mark that?"
  - answer needed for decision
  - they'll form a model from scant instruction plus observed behaviour
    - this model may be incorrect
  - their reaction to prompts or lack thereof depends on trust in their reliability, produced in part by models
  - readers did not follow instructions to ignore CADT in classification of features
  - readers resent the high frequency of false prompts

- other observations of previous practice, suspected effects of introducing CAD
  - "double reading" is part of continuous recalibration of readers
  - readers [believe they] use all complex evidence about a patient, about their own abilities

UCL CHIME experiment

50 readers looked at 180 cases:

- 120 normal cases (normals)
- 60 cancers

in two conditions:

- without CADT support (unprompted session)/
- with CADT support (prompted session)

details

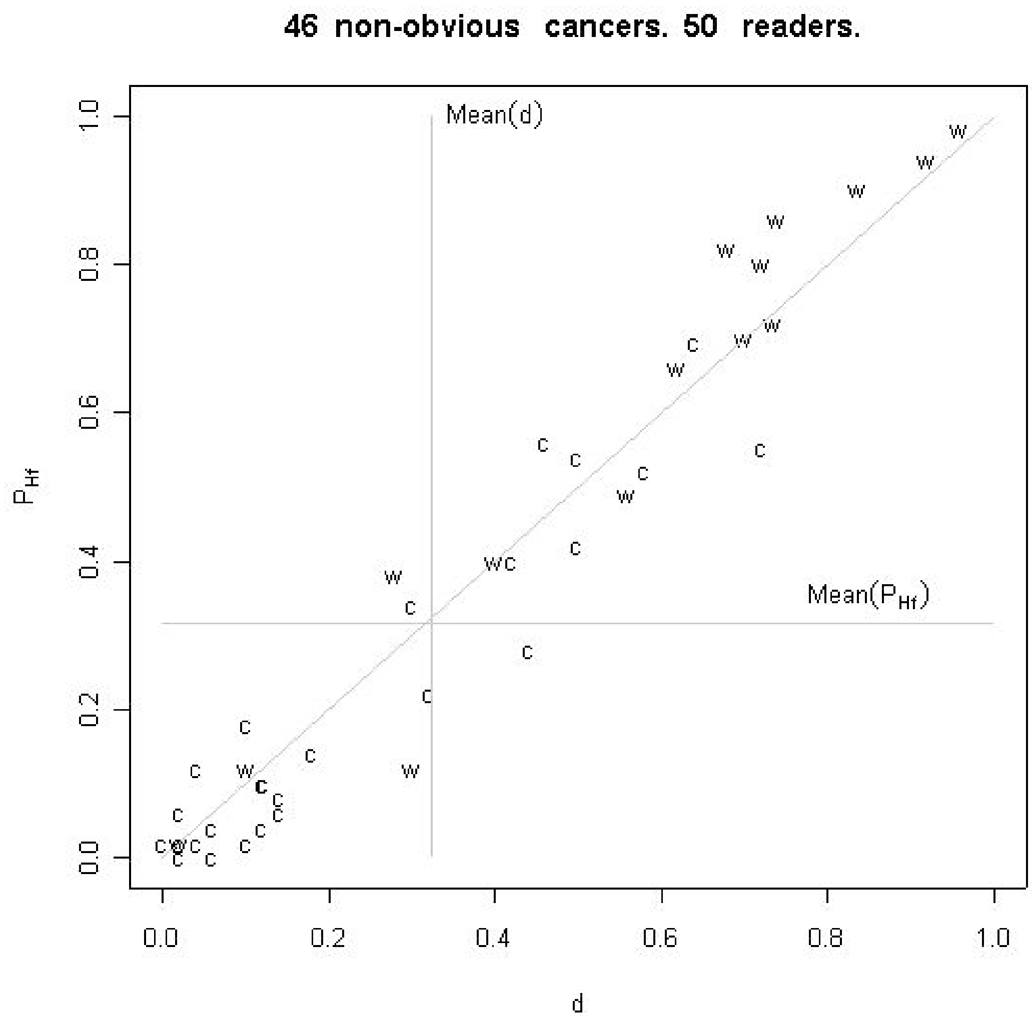
- order of seeing cases in prompted and unprompted conditions partially randomised
- **Rate of cancers** much higher than in real working conditions
- **CADT printout** used instead of using real system
- **Each reader** marks a case with mark:
  - “1” recall (definitely cancer)
  - “2” discuss and possibly recall (possibly cancer)
  - “3” discuss and possibly not recall (possibly normal)
  - “4” not recall (definitely normal)

Standard statistical analyses of the UCL trial results

… revealed no significant difference between reader performance with and without CADT support in the UCL trial,

- artefact of experimental setup?
  - very good readers?
  - lack of fatigue/boredom?
  - “Hawthorne effect”?
- lack of diversity between readers and CADT?
- indeed, [unaided] readers’ and machine’s failures are highly correlated
  - cause: “similarly trained” subsystems?
  - potential for improvement: increase diversity?
    - + (appears popular with readers)

e.g., UCL data for cancers



However, further analyses indicate

- a lot is going on that averages hide
  - e.g. readers change their minds a lot between the two conditions
  - need to separate “random” error from systematic effect of case (“difficulty”) and of CADT
- results: correctness of CADT output is a significant factor affecting decision reliability for non-obvious cancers
  - people do not just ignore prompts.
  - correct computer output helps
  - wrong computer output makes things worse
  - total effect in trial not significant

- e.g.
  - false positives reduced by correct CADT output (reasonably! ...?)

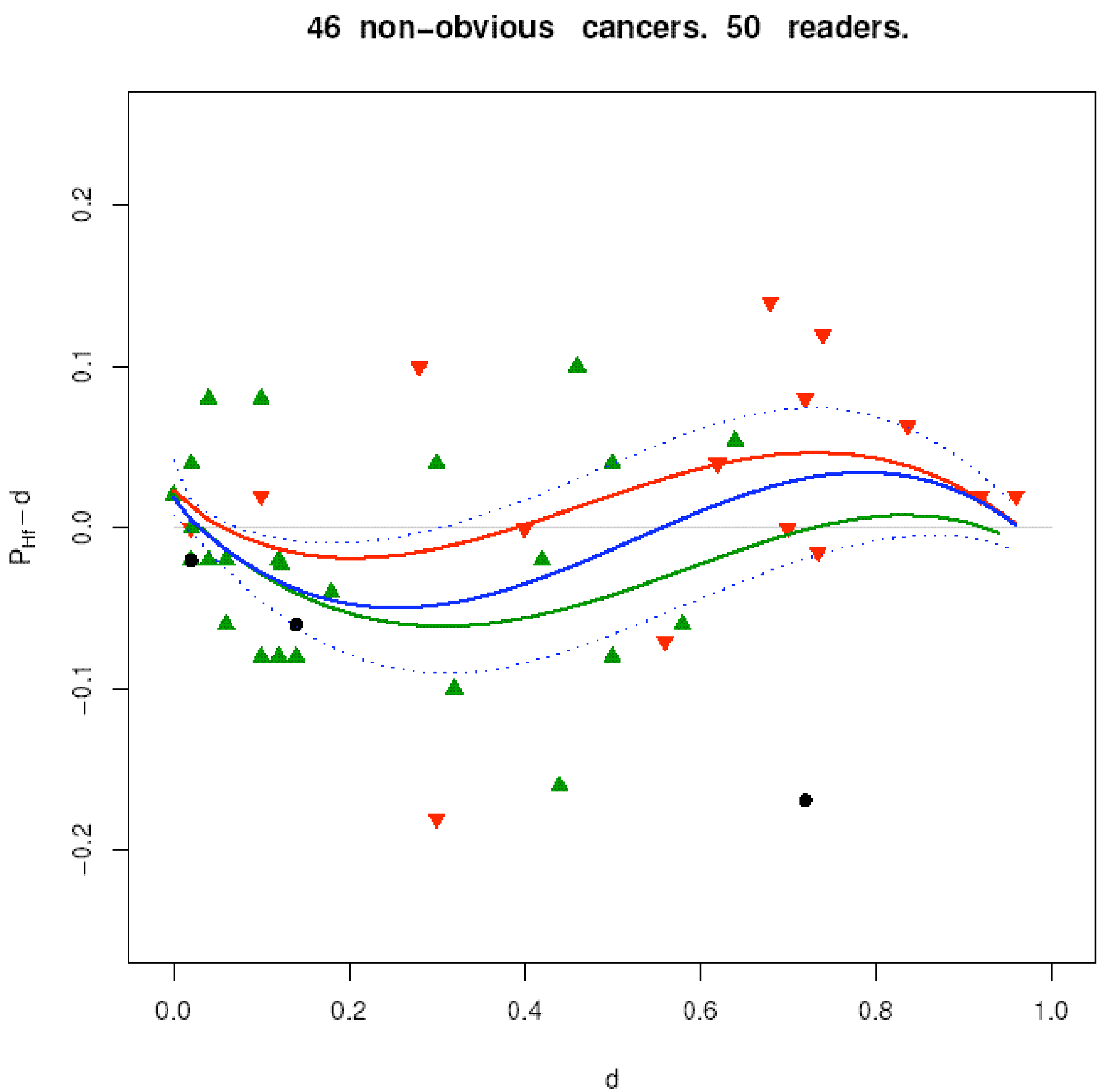
    (CADT meant to reduce false negatives)
  - reader with CAD is more likely to miss unprompted cancer
  - but these effects compensate on average
    - + difference might be amplified in transition to field
  - causal mechanisms?
  - consider apparent effect from absence of prompts (see later)

Example of exploratory analysis

Regression to filter out noise:

difference between unprompted, prompted readers as function of unprompted "difficulty" alone,

or of “difficulty” and of correct or wrong prompting



Supplementary studies

•Supplementary experiment (Study 1):

- goal: estimate probability of reader failure given incorrect CADT output
- data set with an artificially large proportion of cancers not prompted by computer
- readers saw cases with computer support
- we observed poor reader sensitivity - in particular, for cancers not prompted by computer

•Control study (Study 2):

- goal: to test whether indeed wrong CADT output has large effect
- same data set but with a control group of readers without computer support
- significantly higher sensitivity - especially for cancers that computer would not prompt

Findings from Study 1 vs Study 2

Average Readers Sensitivity:

- Study1(CADT): 52%* (min 27% - max 70%)
- Study2 (No CADT): 67%* (min 50% - max 87%)

Average Readers Specificity:

- Study1(CAD): 90% (min 72%- max 100%)
- Study2 (No CAD): 84% (min 71% - max 100%)

% "Correct" Human Decisions

	Cancers		Normals	
CADT Output	CADT	No CADT	CADT	No CADT
Correct Prompts:	81%	88%	n/a	n/a
Incorrect Prompts:	53%	67%	92%	86%
No Prompts:	21%*	46%*	94%	86%

*Statistically significant difference

Conclusions about CAD and this CADT

even if no bias from inevitably limited realism of *work setting* in trial,

- "no effect" conclusion may be wrong
- effect with realistic population may be good, bad, negligible depending on population
- observed effects suggest "automation bias”
  - absence of prompts appears informative (it is!) and over-influences readers
  - time-effort issues?
- more study needed
  - scrutinise conjectured mechanisms of systematic effects (submit conjectures from exploratory analysis to experimental test)
  - estimate effects in actual practice
  - feedback to *system* designers (machine, procedures)
    - + short term fault tolerance re decisions
    - + long-term fault tolerance re calibration

about advisory/warning systems, evaluation/design methods

– "no harm" assumption methodologically harmful
– importance of diversity:
  + computer reliability may be useless, if reliable useless advice
  + whole-system evidence necessary: effect on reliability of decision
  + perhaps not a good idea to support better humans with replica of an average human
– "diversity" approach - focus on variations between cases - helps
  + systematic effects that "on average" analysis may hide
  + the same support system that is good for an easy situation may damage decision quality in a difficult situation and vice versa (risk transfer isues?)

about advisory/warning systems, ctd

– ….
– "HCI" observations
  + [unintended] use : user may focus on informative signals rather than intended signals
  + dependability of HC interaction determined by design factors way beyond HC interface: user’s “mental model” includes [perceived] reliability of machine in various circumstances
– worth considering engineering tricks like
  + forcing more diversity to improve dependability?
    * even tunable to individual user?
  + procedures to make human appeal for computer support unlikely in situations where it could be detrimental?
  + artificial error background to avoid automation bias?

Some salient conclusions

- HC interaction concerns go far beyond issues of visible interface (with common focus on low-level errors)
- modelling gives *helpful novel* insight
  - can treat humans as components
  - quantify their reliability (in letters not numbers)

  and learn something
- many adverse effects are plausible by common sense and “human” lab research...
  - ... here observed in actual behaviour of experts
  - data may ring alarm bells, and yet standard analysis procedures hide them

Bibliography

L. Strigini, A. Povyakalo, and E. Alberdi, “Human-machine diversity in the use of computerised advisory systems: a case study,” DSN 2003, San Francisco, U.S.A., 2003.

E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton, “Does incorrect computer prompting affect human decision making? A case study in mammography,” presented at CARS 2003: Computer Assisted Radiology and Surgery, 2003.

E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton, “Decision support or automation bias? A study of computer aided decision making in breast screening,” presented at SPUDM 2003 (Subjective Probability, Utility and Decision Making), 2003.

E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton, “Effects of Incorrect CAD Output on Human Decision Making in Mammography,” Academic Radiology, vol. 11, 2004, in print.

writeups of more recent analyses: in preparation