The Work of Reading Mammograms and the Implications for Computer-Aided Detection Systems

Mark Hartwooda, Rob Procter, Mark Rouncefieldb, Roger Slacka and James Southera

aSchool of Informatics, University of Edinburgh, Edinburgh, EH8 9LW
bComputing Department, Lancaster University, Lancaster, LA1 4YR

1. What is the key intellectual contribution of the work presented?

The paper presents findings from an ethnographic study of a computer-based tool for mammography in the context of a clinical trial. Our key contribution is to show the ‘lived work’ of reading mammograms and so contribute towards a better understanding of such technologies in use. In so doing, we show how reading mammograms is a thoroughgoingly social enterprise and is achieved in, and through, the making available of features that are relevant to the community of readers. We show how current presumptions of how readers might use and interact with such computer-based detection tools are inadequate and how they need to be guided by a better understanding of the work and the routine problems of reading mammograms. In turn, this has implications for the design of such tools and digital imaging systems in general. Our point is that as digital imaging systems evolve from performing basic image rendering to incorporate increasingly sophisticated image processing, then users’ interactional requirements become more demanding. In particular, their interfaces need to furnish richer ‘accounts’ of system behaviour if users are to be able to ‘make sense’ of and ‘trust’ in them.

2. Have you published related work? If so, does this work differ? and how?

This paper is the latest in a series of publications that have come from eight years of research into work practices in breast screening and the use of ethnographic studies for informing the design of IT systems for medical work. Here, we evaluate a computer-aide detection tool for mammography and relate findings on work practice to the design of interfaces for these tools.

3. What is the most closely related work by others? How does this work differ?

This paper draws on the work of Goodwin and the notion of ‘professional vision’ as a way of understanding the social, intersubjective characteristics of work practice. In so doing, we demonstrate it’s relevance to understanding the work of reading mammograms.

4. Has your work been tested or developed in association with clinical practice? If so, how?

The work is entirely based on studies of clinical practice.
The Work of Reading Mammograms and the Implications for Computer-Aided Detection Systems

Mark Hartswood\textsuperscript{a}, Rob Procter\textsuperscript{a}, Mark Rouncefield\textsuperscript{b}, Roger Slack\textsuperscript{b} and James Soutter\textsuperscript{a}

\textsuperscript{a}School of Informatics, University of Edinburgh, Edinburgh, EH8 9LW
\textsuperscript{b}Computing Department, Lancaster University, Lancaster, LA1 4YR

Abstract. We examine the ways that readers make sense of mammograms in context, showing how a consideration of the social aspects of this work might illuminate practice and suggest ways for the building of computer-based tools to support such work. We show how sense-making is a situated activity and raise some concerns as to the ways that technologies have been developed to support reading may impact negatively upon the very practice they were intended to support. We show how it is important to consider technologies in use and discuss how they might be developed to support real world use, as opposed to some idealised formulation of it. We conclude with some outline suggestions towards better user interfaces for computer-aided detection systems, in particular, and for digital imaging systems in general.

1 Introduction

The practice of breast screening calls for readers to exercise a combination of perceptual skills to find what may be faint and small features in a complex visual environment, and interpretative skills to classify them appropriately – i.e., as benign or suspicious. Current UK NHS breast screening practice is for each mammogram to be ‘double read’, i.e., assessed independently by two readers \cite{1}. Because of the growing shortage of trained readers, there is interest in using computer-aided detection (CADe) systems to replace double reading with a single reader using a CADe system. We report here ethnographic studies of readers using a CADe system which we conducted during clinical trials.

2 Method

As a complement to the quantitative emphasis of the conventional clinical trial, we used ethnographic investigative and evaluative techniques \cite{7}. Ethnography argues for understanding the situatedness of individual activities and of the wider work setting, highlighting the interdependencies between activities, and stressing the ‘practical participation’ of individuals in the collaborative achievement of work. For the purposes of designing and developing computer-based tools, the advantage of applying ethnographic methods lies in the ‘sensitising’ they promote to the real-world character of activities in context and, consequently, in the opportunity to help ensure that systems resonate with the circumstances of use. This is, we argue, particularly important to medical work, where the lack of attention to work practice has been responsible for many failures of IT systems.

3 The Trial

The CADe system being trialed was the R2 Imagechecker. Following conventional clinical trial design, three sets of 60 prompted and unprompted cases were prepared using historical mammograms. During the trial, readers were shown the appropriate mammogram – CCs and Obliques, but not previous mammograms (or any notes) – and asked to indicate areas of concern and to make a decision as to whether the case should be recalled for further investigation using a four point decision scale: 1. Recall; 2. Discuss but probably recall; 3. Discuss but probably no recall; 4. No Recall. Before the trial was run, each reader was given a brief explanation of how the CADe system worked, emphasising that it was intended to be used for detection rather than for diagnosis. Readers were told that the system ‘spotted’ masses and calcifications and about the appropriate prompts. They were also advised that the threshold of sensitivity of the system had been set such that there would inevitably be a lot of false prompts; and warned that since this was a trial set there would be more cancers than in a ‘normal’ reading session.

4 Observations

As part of the trial, readers were observed doing the various test sets and then asked about their experiences of using the prompts. Readers were also taken back to cases identified in the test set where they had appeared to have had difficulty or spent a long time making their decision and asked to talk through any problems or issues to do with the prompts and their decisions. Although there were variations in how readers approached the trial, the fieldwork extract below gives some idea of the process observed:
Case 10: Looking at film – using a blank film to mask area outside that of immediate interest. Magnifying glass. Looking at booklet prompts - looking back at film. "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 ..." Marks decision.

The main strengths of the CADE system in supporting this kind of work seemed to lay in picking up subtle signs – signs that some readers felt they might have missed – and stimulating interaction between reader and the available technology by motivating them to re-examine the mammogram. As one reader said:

"Those micros that the computer picked up ... I might have missed it if I was reading in a hurry ... I'd certainly missed them on the oblique ... This one here the computer certainly made me look again at the area. I thought they were very useful, they make me look more closely at the films ... I make my own judgement ... but if the prompt is pointing things out I will go and look at it again."

There was also a perception that the CADE system was more consistent than readers might be:

"... it's just the fact that it's more consistent than you are ... because it's a machine."

Readers also frequently express the opinion that they are better at ‘spotting’ some cancers – as having skills or deficiencies in noticing particular types of object within films. This was another area where the CADe prompts were seen as useful, as both compensating in some (consistent) way for any individual weaknesses of the reader and as a reminder of ‘good practice’:

"My approach tends to be to look for things that I know I’m not so good at ... there are certain things that you do have to prompt yourself to look at, one of them being the danger areas."

Amongst the weaknesses identified by readers was the distracting appearance of too many prompts:

"This is quite distracting ... there’s an obvious cancer there (pointing) but the computer’s picked up a lot of other things ... there’s so many prompts ... especially benign calcifications ... you’ve already looked and seen there are lots of benign calcs."

The CADE system was also seen to prompt the ‘wrong’ things – benign features or artefacts of the mammogram generation process: “... what the computer has picked up is benign ... it may even be talcum powder ... I’m having trouble seeing the calc its picked up there ... (pointing). I can only think its an artefact on the film."

At the same time, the CADe system was seen to be missing obvious prompts that raised wider issues to do with trusting and ‘understanding’ the system:

"I’m surprised the computer didn’t pick that up ... my eye went to it straight away."

Our wider studies of breast screening show how reading mammograms is a thoroughlygoingly social enterprise and is achieved in, and through, the making available of features that are relevant to the community of readers as opposed to some idealised individual cogniser [6]. It is for this reason that we turn to the work of Goodwin and, in particular, his notion of ‘professional vision’ [4], to explicate the social, intersubjectively available nature of doing reading. In mammography, a reader has to learn how to interpret the features on the mammogram and what they mean, as well as how to find them. We have described how readers ‘repertoires of manipulation’ make features visible [6]. Methods for doing this include using the magnifying glass and adopting particular search patterns:

"Start at top at armpit ... come down ... look at strip of tissue in front of armpit ... then look at bottom ... then behind each nipple ... the middle of the breast."

Readers also attempt to ‘get at’ a lesion by measuring with rulers, pens or hands from the nipple in order to find a feature in the arc; comparing in the opposite view; aligning scans; looking ‘behind’ the scans; ‘undressing lesions’ by tracing strands of fibrous tissues into and out of the lesion area and so on. A magnifying glass may be used to assess the shape, texture and arrangement of calcifications or, where the breast is dense, the mammogram may be removed and taken to a separate light box. These repertoires of manipulations are an integral part of the embodied practice of reading mammograms. Such features are not work arounds, but an integral part of the ecology of practice built up in and as a part of doing reading mammograms.
The positioning of an object in a particular area of the breast renders it more suspicious than if it had been elsewhere. At the same time, certain areas within the mammogram are regarded as more difficult than others to interpret and readers particularly orient to them in their examinations. As one reader noted:

“I do … I have areas where I know I’m weak at seeing … you know ones that you’ve missed … one is over the muscle there … its just because the muscle is there … if you don’t make a conscious effort to look there you tend not to see that bit of breast and the other area is right down in the chest wall – breast and chest wall area … because in older women the cancers tend to be in the upper outer quadrant so I look in that area very carefully … it depends on the type of breast really.”

We would also stress the self-reflective nature of readers’ behaviour. Readers know about their own strengths and weaknesses (in one centre, a reader is referred to as ‘the calcium king’ because of his ability to detect calcifications; a member of another centre is referred to as ‘Mrs Blobby’ because of her ability to detect lesions in dense areas). Readers are sensitive towards the set of criteria for correctness and what is required for the satisfaction of the maxims that constitute it.

5 Discussion

It is important to note that the CADe system should not be taken to make reading mammograms less uncertain – decisions still have to be made and these fall to the readers. Prompts are ‘docile’ in that their character is simply to draw the reader’s attention to candidate features as opposed to say what should be done with them. That a prompt occurs is a meaningful thing, but what to do about it is still a readers’ matter. In other words, the system still requires the professional vision of the reader to remedy prompts as what they accountably are. A reader makes what is seen or prompted accountable in, and through, the embodied practices of professional vision. That a mammogram feature or a prompt is there is not of itself constitutive of a lesion or other accountable thing, it must be worked up through these embodied practices and ratified in the professional domain of scrutiny. The CADe system knows – and can know – nothing of what it is to be a competent reader and what it is to look for features in a mammogram beyond its algorithms, and the reader must ‘repair’ what the system shows, making it accountable in, and through, their professional vision. This is, as we have argued, a thoroughlygoingly social procedure and, as such, something that the CADe system cannot be a part of. Beyond its algorithms, the CADe system cannot account for what it has and has not prompted, and it cannot be queried as a colleague can.

Readers used prompts to develop some understanding of the CADe system’s scope and function. However, they occasionally held incorrect notions about, e.g., the system prompting for asymmetry and were often baffled by the high level of false positive prompts. In part, this ability to make sense of how the system behaves also impacts on issues of dependability and trust in the system. We have argued elsewhere that how readers use prompts to inform their decision-making, and how they make sense of a CADe system’s behaviour, may be important for maximising effectiveness [5]. We find that readers rationalise false prompts by devising explanations or accounts of its behaviour that were grounded in the properties of the mammogram image. This points to general issues concerning trust – users’ perception of the reliability of the evidence generated by such tools – and how trust is influenced by users’ capacity for making sense of how the system behaves. The need to account for a prompt – even if it is dismissed – distracts the reader. In other words, its docile prompts often call attention to features that the readers have decided are not important enough to merit attention.

The CADe system prompts features that are not cancers, as well as missing features that may be obviously cancers to the reader. For example, normal features in the breast such as calcified arteries or crossing linear tissues can be prompted as micro-calcifications, while other normal features such as ducts and tissue radiating from the nipple or inadvertent crossing of parenchymal tissue can produce a prompt for a cancerous mass. That the system prompts features other than the cancer is regarded as problematic but still in need of account. It might be said that the system works too well, providing not just too many prompts, but prompting features that a skilled reader would not accept as promptable. In part this is a feature of the technology that the readers (at least in this trial) effectively ‘forget’, but which might be incorporated into readers’ ‘biography’ of the system in time.

How do readers construct, achieve or make sense of the system? Following Schutz, we might argue that readers render mammograms intelligible using a mosaic of ‘recipe knowledge”: “a kind of organisation by habits, rules and principles which we regularly apply with success.” [8]. While the common experiences and rules embodied in the ‘mosaic’ are always open to potential revision they are, nevertheless, generally relied upon for all practical purposes as furnishing criterion by which adequate sense may be assembled and practical activities – reading the mammogram – realised. Of course, in everyday interaction with colleagues any breakdown in sense is rapidly repaired and ‘what is going on’ readily understood. But, when the other participant in the interaction is a computer, difficulties can arise as readers (in this case) characteristically rush to premature and often mistaken...
conclusions about what has happened, what is happening, what the system ‘meant’, what the system is ‘thinking’, and so on. The problem is, of course, that the CADe system is not capable of reciprocating the perspective of the skilled practitioner.

It would therefore seem desirable to increase the scope for a CADe system to be ‘self-accounting’ [2] through the provision of richer and more sophisticated user interfaces. It is certainly possible to conceive of richer representations of a CADe system’s behaviour, but it is an open question as to whether such representations could be sufficiently contextualised in a manner that would enable readers to use them easily and in any meaningful sense. It seems to us that such representations are not accounts in themselves, but resources for the realisation of accounts in context. We argue that even a series of representations from which readers could choose may not provide sufficient detail to answer all conceivable ‘why that now’ types of questions.

6 Conclusions

The current generation of CADe systems are designed with user interfaces that presume that all readers need to see is the bare and unadorned prompt. Our ethnographic investigations of the CADe system on trial show, however, that this presumption is false. Indeed, we would argue that as digital imaging systems, in this and other medical work domains, evolve from performing basic image rendering to incorporate increasingly sophisticated image processing, then users’ interactional requirements become more demanding.

It is clear from our study that readers need an understanding of what the CADe system has prompted and why. The key problem we observe is that the system does not provide accounts of its behaviour. The docile nature of the prompts generated requires the reader to formulate ad hoc an explanation for their presence. Thus, there is a need for the reader to engage in some kind of retrospective search for what it is that the CADe system might have ‘meant’ or ‘intended’. Without the possibility of being able to assemble an account from the source, so to speak, the reader has to develop some notion of the potentialities of the system – which, as we have seen, may or may not be consistent with what the CADe system actually does.

If readers are to ‘trust’ CADe systems, they need accounts of why prompts come – or come not – to be there. It is also important to consider how far accounts of prompts might be intrusive and thereby impact negatively on the work of readers. We therefore need to consider what these accounts would look like and for whom they would be intended. In other words, these accounts must be designed to relevant to readers’ concerns. We suggest that one way of moving towards assembling readers’ accounts is for CADe system developers to work with closely with readers as the latter become acquainted with the system’s performance characteristics over time.

Acknowledgements

We would like to thank the readers who participated in this study for their time and patience. Also, we are grateful to Paul Taylor and his colleagues at CHIME for giving us the opportunity to carry out this study. This work was supported by the EPSRC under grant number GR/ R24517/01.

References