'Repairing' the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening

Mark Hartswood[1], Rob Procter[1], Mark Rouncefield[2], Roger Slack[1], James Soutter[1] and Alex Voss[1]

[1]Social Informatics Cluster, School of Informatics, University of Edinburgh
[2]Department of Computing, Lancaster University
sic@inf.ed.ac.uk

Abstract. In this paper, we consider the problems of introducing computer-based tools into collaborative processes, arguing that such an introduction must attend to the sociality of work if it is not to impact negatively upon the work that they are intended to support. To ground our arguments, we present findings from an ethnomethologically-informed ethnographic study carried out in the context of the clinical trial of a computer-based aid in medical work. Our findings highlight the problematic nature of traditional clinical trials for evaluating healthcare technologies, precisely because such trials fail to grasp the situated, social and collaborative dimensions of medical work.

Introduction

Our research is focused on investigating and understanding the relationships between work practices and technologies. One of the work settings in which we have a longstanding interest is healthcare. In this paper we present some findings from an ethnographic study carried out in the context of clinical trial of a computer-aided detection (CADe) tool that is intended to support the work of radiologists working in breast screening.

The aims of this study are twofold. First, we are interested in understanding the impact of CADe tools on the situated, collaborative practical actions of reading mammograms – actions that we argue constitute radiologists' *professional vision*, i.e., "socially organized ways of seeing and understanding events that are answerable to the distinctive interests of a particular social group" (Goodwin, 1994:606). Elsewhere, we have argued for the importance of professional vision for the maintenance of radiologists' decision-making performance, and documented the various ways in which they act to sustain professional vision through the organisation of reading and the artefacts used for reporting this work (Hartswood, Procter, Rouncefield and Slack, 2002). One of the issues that we wish to examine in this paper is how the adoption of CADe tools might – or might not – mesh with these practices.

Second, by bringing the situated and practical actions of reading work to the fore, we aim to question the value of the 'gold standard' for medical technology evaluations: the quantitative, randomised, control clinical trial. In particular, we would stress the importance of complementing the clinical trial's quantitative emphasis with qualitative investigations of the impact of technological interventions on the everyday working and mundane interactional practices of medical workers.

Breast screening in the UK

Breast cancer is the most common non-skin related malignancy and accounts for one-fifth of deaths among women from all forms of cancer in the UK, and is the second leading cause of cancer death among women in the US and Europe. A screening programme, based upon mammography, has been in operation in the UK for more than 10 years. The initial screening test is by mammography, where one or more X-ray films (mammograms) are taken of each breast by a radiographer. The usual types of mammogram taken are mediolateral oblique (Oblique) and craniocaudal (CC). Each mammogram is examined for evidence of abnormality by at least one trained reader[1]. Types of feature that are indicators of malignancy include: micro-calcification clusters are small deposits of calcium visible as tiny bright specks; ill-defined lesions are areas of radiographically-dense tissue appearing as a bright patch that might indicate a developing tumour; stellate lesions are visible as a radiating structure with ill-defined borders. Architectural distortion may be visible when tissue around the site of a developing tumour contracts; asymmetry between left and right mammograms may be the only visible sign of some features.

---

[1] Most, but not all, readers are qualified radiologists. For the sake of simplicity, we will use the more general term of reader.

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. Two reader performance parameters are particularly important: specificity and sensitivity. A high specificity (high true positive rate) means that few women will be recalled for further tests unnecessarily; a high sensitivity (low false negative rate) means that few cancers will be missed. Achieving high specificity *and* high sensitivity is difficult.

The goal of screening is to achieve a high, reliable and controlled cancer detection rate. Current UK breast screening practice is that each mammogram should be 'double read', i.e., assessed independently by two readers (Blanks, Wallis and Moss, 1998). Superficially, this suggests that each reading of a mammogram is the work of the individual reader. Our earlier studies reveal, however, that reading mammograms is a thoroughgoingly social enterprise that is achieved in, and through, the making available of features that are relevant to the community of readers (Hartswood, Procter, Rouncefield and Slack, 2002). This involves a number of formal and informal collaborative practices. As an example of the latter, through the use of annotations on the screening reporting form, readers contrive to use double reading in order to make their work observable-reportable as they read, thereby enabling them to intersubjectively calibrate their performance without sacrificing their independence as decision-makers.

Because of the growing shortage of trained readers, there is interest in the UK breast screening programme in using CADe tools to replace double reading with a single reader using a CADe tool. The principle of CADe is to apply image analysis algorithms to identify target features in each mammogram and draw these to the reader's attention through the use of prompts. The prompts act as an attention cue, and so counteract the effects of variability in concentration and, more generally, make the visual search pattern more systematic and complete. A number of CADe tools have now been developed, but the practical realisation of their potential benefits is not easy (Hartswood, Procter and Williams, 1998). As Warren-Burhenne et al. (2000) comment, while CADe tools appear useful "we must also realize the possible drawbacks and fully understand the proper use of such a device".

The implications of CADe for readers' work practices is one of the issues that we focus on in this paper. In particular, as we will show, the proposed reconfiguration of readers' work is problematic by virtue of the very manner in which the CADe tool under trial works.

Ethnomethodologically-informed ethnography and the evaluation of healthcare technologies

While this paper is primarily concerned with presenting an empirical study of a technology in use, it also necessarily problematises some of the issues involved in the evaluation of healthcare technologies in general. It thereby documents some of the real difficulties of any evaluative exercise, addressing the concern raised by Bannon (1996) that "evaluations are important yes, but it is also important to be aware of the quality of the evaluation, and of what can legitimately be learned from any particular study". Bannon (1996) goes on to suggest:

> "a careful systematic account of what happens in particular settings when a prototype or system is installed, and how the system is viewed by the people on the ground, can provide useful information for 'evaluating' the system and the fitness for the purpose for which it was designed." (Bannon, 1996:427)

The gold standard for evaluation of healthcare technologies is the randomised control clinical trial. The method is increasingly seen as problematic for evaluating computer-based systems, however, since while these may perform well under trial conditions, they are nevertheless often are found wanting in use (e.g., Hartland, 1993).

> "By insisting on evidence from randomised control trials we waste precious resources on evaluation work that is methodologically flawed and impractical and at best provides results that are difficult or impossible to generalise." (Heathfield and Buchan, 1996:1008)

Following Heathfield and Wyatt (1993) and Heathfield and Buchan (1996), we argue that while the traditional clinical trial methodology may provide useful measures of *efficacy*, as measure of *effectiveness* it is entirely inappropriate. The problem is that, for the sake of statistical repeatability, the randomised, control trial glosses the way in which the work that the technology is intended to support is actually done and so fails to get to grips with understanding (and evaluating) technologies in their social and organisational circumstance of use.

As an attempt to address this problem, we have sought to complement the clinical trial methodology with ethnomethologically-informed ethnographic investigative and evaluative techniques (Hughes et al., 1994). The main virtue of ethnography lies in its ability to make visible the real-world sociality of a setting and efforts to incorporate ethnography into IT systems development processes stem from the realisation that the success of design has much to do with the social context into which systems are placed. 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. As Suchman argues:

> "… ethnographies provide both general frameworks and specific analyses of relations among work, technology and organization. Workplace ethnographies have identified new orientations

for design: for example, the creation and use of shared artifacts and the structuring of communicative practices." (Suchman, 1995: 61)

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 the design of technologies resonates with the circumstances of use. As a method of evaluation, ethnography attends to the haecceities[2] of the setting, showing in this study, for example, how practical actions such as mammogram arrangement, gesturing and pointing to features on mammograms, manipulating mammograms, and annotations are all components of the lived work of doing reading.

The CADe machine was evaluated using the conventional clinical trial methodology in order to quantify its differential impact on reader performance, i.e., on their sensitivity and specificity. This quantitative evaluation was complemented by ethnographic studies of its use under trial conditions with the aim of contextualising and explaining the performance data.

Evaluating the machine

The CADe machine consists of two components, a digitising and image analysis unit and an optical mammogram viewer with two built in computer screens to display any prompts generated by the analysis (see Figure 1). Up to twenty cases (sets of mammograms for an individual woman; typically four mammograms per case, i.e., Oblique and CC views of each breast) can be digitised in a single 'session', although the machine can store up to 1000 cases. When the mammograms have been digitised, analysed and loaded onto the viewer, moving on to the next set of mammograms automatically triggers the display of the appropriate prompts.

Once digitised, analysed and loaded, the mammograms are arranged in the following order: Right-Oblique Left-Oblique; Right-CC Left-CC – mirroring the way the prompts appear on the computer displays. Mammograms on the viewer are scrolled up and down. When the button used to scroll the next set of mammograms into view is pressed then the prompts screens are 'switched off' and a further button needs to be pressed to see the prompts. In this way readers are encouraged to examine the mammograms prior to looking at the prompts.

---

[2] By this we mean that it is important to attend to the 'just this-ness' of the setting, just what it is, here and now with these members and this assemblage of technologies and so forth. See Garfinkel and Wieder (1992) for more on this topic.
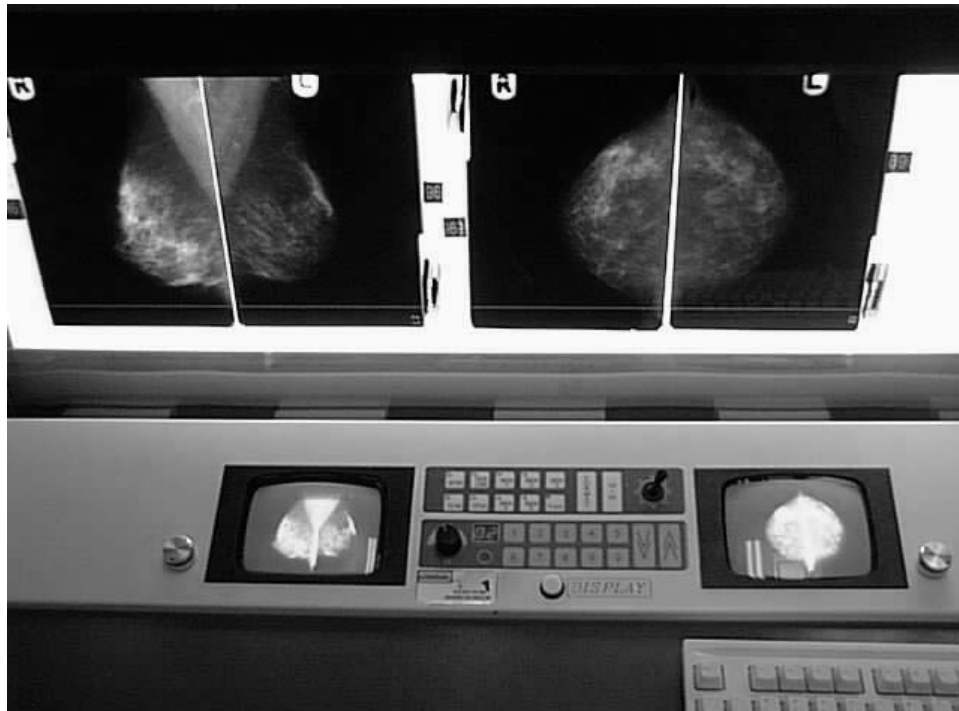
Figure 1: The CADe machine showing the mammogram viewer and prompt displays.

The CADe machine targets ill-defined and speculated lesions and micro-calcifications. Calcification clusters are marked by a shaded triangle. Ill-defined lesions are marked with an asterix and a circle is drawn around either prompt type if the machine's confidence is high. The machine does not perform a comparison between left and right views (i.e., for asymmetry). The machine's image analysis algorithms cause it both to prompt features that are not cancer as well miss some obvious cancers. As an example of the former, 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. As an example of the latter, the machine will miss masses that may be obviously cancers because they are either under 10mm or over 20mm in size, the machine's preset range for mass detection.

Following conventional clinical trial design, three sets of 60 prompted and unprompted (control) 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 – with decisions 1 and 2 treated as recall decisions for the purpose of analysis. Before the trial was

run, each reader was given a brief explanation of how the machine worked, emphasising that the machine was merely for detection, not diagnosis. Readers were told that the machine 'spotted' masses and calcifications and about the appropriate prompts. They were also advised that the threshold of sensitivity of the machine had been set such that there would inevitably be many false prompts; and warned that since this was a trial set there would be more cancers than in a 'normal' reading session.

Trial observations

As part of the ethnographic evaluation, readers were observed doing the various test sets and then asked about their experiences of using the CADe machine. The readers were also taken back to cases identified in the test sets 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 a reading and the test, the fieldwork extract below gives some idea of the process observed:

Simplified fieldwork extract 1:

Case 1: Gets blank film to mask area of the film ("so I can concentrate on it" ... these are set up differently from the way I usually look at them … so I have to train my eye each time."). Using magnifying glass. Marking on booklet. Looking from booklet to scan. Homing in on an area – "I'd say it's benign"

Case 2: Using blank film. Takes film off roller and realigns. Magnifying glass. Looking from booklet to film. "I'd not recall ... what the computer has picked up is benign ... it may even be talcum powder."

Case 10: Looking at film – using blank film to mask area. 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.

Case 15: Looking at film – aligns on roller – gets blank film to mask – gets magnifying glass. Looking at booklet prompts - looking at film – back to booklet - looking at film. "There's quite a suspicious mass on the CC – I'm surprised it didn't pick it up on the oblique." Marking booklet – makes decision.

It was observed that, as with 'everyday' reading, readers used a repertoire of manipulations to make certain features 'more visible'. 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. Where a reader wished to attend to a particular segment of the mammogram, a blank film was used to blank off a part of the mammogram (see Figure 2).

In cases where a suspicious feature was seen on one view, readers used their fingers or an object such as a pen for measurement and calculation so as to check for the feature's appearance in the other view. As we discuss later, these

repertoires of manipulations are an integral part of the embodied practice of reading mammograms (see Figures 3 and 4).
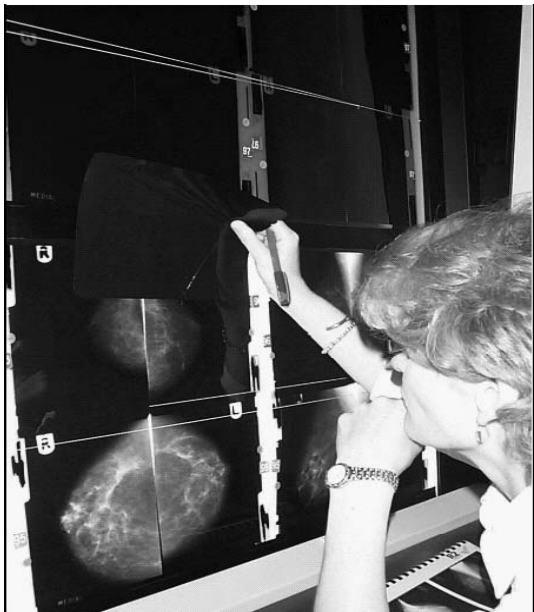

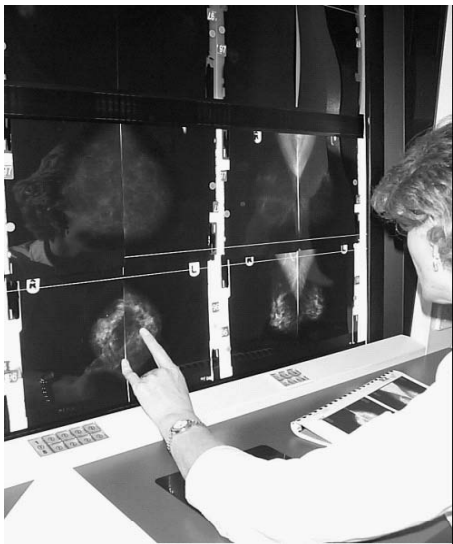Figure 2: Using an opaque film to block off areas of a mammogram.

Figure 4: Using fingers for measuring and comparing.

Figure 3: Using a magnifying glass.

The main strengths of the CADe machine 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 prompting 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 judgment ... but if the prompt is pointing things out I will go and look at it again but I make my own judgment ... it might bring up abnormalities that I haven't seen whether they're benign or ... nothing but its still nice to have a prompt so that I can look again."

There was also a perception that the CADe machine was more consistent than readers might be:

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

Readers 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 machine 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."

> "I'd made up my mind about where the cancer was ... but I was looking at all these other areas … because one has to look at the other breast from experience because one has to look for the second cancer that maybe difficult to see ... and also you're looking for multi-focal cancer."

Amongst the weaknesses identified by readers was the distracting effects 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 machine was also seen to prompt the 'wrong' things – benign features or artefacts of the mammogram production 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 it's an artefact on the film (a thin line at the edge of the film)."

At the same time, the machine was seen to be missing obvious prompts that raised wider issues to do with trusting and 'understanding' the machine:

> "That's quite a suspicious mass on the CC … I'm surprised it didn't pick it up on the oblique. (Points to area) I'm surprised the computer didn't spot it ... it's so spiky … I'd definitely call that back."

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

In documenting the reactions to the prompts above, it is interesting that the need to *account for* the prompt – even if it is dismissed – distracts the reader. Evidently, the machine does not appear to be capable of reciprocating the perspective of the skilled practitioner. In other words, its 'docile' prompts call attention to features that the readers have decided are not important enough to

merit attention. Nevertheless, for readers, the fact that the machine prompts for features other than candidate cancers is still in need of account.

Reading work: 'worldly interpretations' and 'professional vision'

We have noted elsewhere (Hartswood, Procter, Rouncefield and Slack, 2002) how readers make use of 'worldly interpretations' (Driessen, 1997) of the significance of the object – through ideas about 'territories of normal appearance' and 'incongruity procedures' (Sacks, 1972). Just as Sacks explicates how police officers spot criminals attempting to 'pass' as normal persons, readers are able to spot abnormalities within mammograms: what is normal is contingent on its presence in a particular site and time. Just as there are people who do not belong in a certain part of the city, there are objects which do not belong in a 'normal' (i.e., non-recallable) mammogram. A reader faced with a set of mammograms where the pectoral muscle is not seen equally on both mammograms will be able to marshal a candidate version of why this is the case – they will be able to suggest, for example, that the woman was difficult to position and then seek confirmatory sources within the ensemble of evidence at their disposal (equally, they might find no such reason and hold the staff responsible for making the mammograms accountable).

Thus, 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 interpet and readers particularly orient to them in their reading. 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 ... with some breasts they're very fatty and you can just look and not see something but in other breasts ... so I try to look at the whole film, because I know if I just glance at it and don't make that conscious effort I don't look ... and also these rather big breasts where all the breast tissue gets scrunched up ... that's sometimes difficult."

Readers are sensitive towards the set of criteria for correctness and what is required for the satisfaction of the maxims that constitute it:

"My approach tends to be to look (positively?) 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."

Readers are also aware of the work that they have done over the course of a day and the reasonableness of their finding or not finding recallable objects therein:

"Paranoia can set in if I have a large number of films that have passed as normal – might think 'what have I missed?'"

"If you get to the end of a session, the end of a pile of reporting and you haven't recalled anything, then you think 'this is ... maybe I've missed something' then in the next bunch you find that you will recall every other one. So, it averages out."

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).

As so far described, the work of reading mammograms has the appearance of an individualised task. However, our earlier studies have shown that it has significant collaborative dimensions that are important for its reliable performance (Hartswood, Procter, Rouncefield and Slack, 2002). 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. It is for this reason that we turn to the work of Goodwin and, in particular, his notion of 'professional vision' (1994), to explicate the social, intersubjectively available nature of reading work. Goodwin suggests that professional vision might be thought of as follows:

"Discursive practices are used by members of a profession to shape events in the domains subject to their professional scrutiny. The shaping process creates the objects of knowledge that become the insignia of a profession's craft: the theories, artifacts, and bodies of expertise that distinguish it from other professions. Analysis of the methods used by members of a community to build and contest the events that structure their lifeworld contributes to the development of a practice-based theory of knowledge and action." (1994, p. 606).

Goodwin's analysis is based on an ethnographic study of field archaeologists and shows how, in what is a complex 'semiotic field' (Goodwin, 2000a; 2000b), archaeologists are able to find and make observable-reportable features such as post holes through their action on and scrutiny of a piece of ground. How one "establishes a figure in what is quite literally a very amorphous ground" (Goodwin, 1994:610) requires that features in that ground are made visible and accountable, in the case of archeology, by circling the surrounding dirt with a trowel. "It is in this way that the perceptual field provided by the dirt is enhanced in a work-relevant way by human action on it" (*op cit*:611). As Goodwin notes, it is through coding, highlighting and "producing and articulating material representations" that participants build professional vision. Professional vision is a way of seeing, a technique for making work relevant features available, making them stand out from the rest of the field and accounting for just what they are in and as a part of this or that domain of scrutiny.

As with archaeology, so with 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 employ a number of techniques for

making features visible: these 'repertoires of manipulation' make the area one on which to perform an analysis of the feature and come to a formulation of what it is or might be. Methods for doing this include using the magnifying glass; 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."

As we saw in the summary of the trial observations, 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. 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. That is to say, such practices are constitutive of the discipline. We should keep in mind Garfinkel et al's notion of the 'potters object' here: it is the 'intertwining of worldly objects and embodied practices' (Garfinkel et al., 1981:165) that realises – literally makes real for all practical purposes – the accountable decisions of the reader. In and of itself, the mammogram is not enough – it is a start, but it is only realised as an 'increasingly definite thing' (*idem*) through the practical actions undertaken. That is to say, it is "real in and as of inquiry's hands on occasions." (*idem*).

Such manipulations should not be thought of as 'private' – merely there to facilitate the single reader's interpretation of the case at hand – but rather, and also, they are the means that readers routinely use to make features within, and interpretations of, the mammogram available and accountable to others. This is what we mean when we say that they are 'constitutive of the discipline' – they are significant in particular ways that are readily intelligible to other film readers in and through their common socialization as readers. Similarly, and as we noted above, readers use annotations as a means of communication with each other through the work of double reading. That a first reader can make an economic mark such as 'calc?' on the reporting form to suggest that a feature has been seen as an example of a calcification, whilst, at the same time, signaling the first reader's uncertainty and prompting the second reader to examine the region, demonstrates the richness and power of these manipulations given readers' professional vision.

Finding order in the machine

It is important to note that the CADe machine should not be taken to make things less uncertain – decisions still have to be made and these fall to the readers. The prompts are docile in that their character is simply to prompt as opposed to say what should be done. In the trial observations we see that readers attempt to ascertain what a prompted feature is. That a prompt occurs is a meaningful thing,

but what to do about it is still a readers' matter. There is still a deal of sense-making to be done in order to account for what the machine is showing as accountably this or that feature, warranting this or that decision. In other words, the machine still requires the professional vision of the reader to remedy prompts as what they accountably are. The following extracts of readers' commentaries give some indication of this concept:

> "I'm having trouble seeing the calc it's picked up there (pointing). I can only think its an artifact on the film (a thin line at the edge of the film)."

> "… just making sure there's nothing the other side (using fingers) and there is ... a bit of chalk but it's harmless.

> (aligns scans) (using fingers) "so what I thought was an asymmetry is probably completely OK."

In each case, we find that the 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 machine knows nothing of what it is to be a competent, professional reader and what it is to look for features in a mammogram beyond its algorithms – that is self evident – and the reader must 'repair' what the machine shows, making it accountable in and through their professional vision. Readers' professional vision turns on their being a competent practitioner, their ability to distinguish between 'normal' and 'abnormal' features of a mammogram. This is, as we have shown, a thoroughgoingly social procedure and as such something that the machine cannot be a part of.

Beyond its algorithms, the machine cannot account for what it has and has not prompted, it cannot be queried as colleagues (informally or formally) routinely are in normal screening work. It is here, then, that the machine runs into problems – its docile prompts render it a 'dumb colleague'. Put simply, when the machine draws attention to something, what it has prompted for may be clear, but the basis for saying what it has drawn attention to is different: the machine is limited to its algorithms, while readers can make sense of the 'prompt' of a colleague through their intersubjectively constituted professional vision. The machine is impoverished in terms of what it can do *qua* colleague – it is not just that a colleague can be interrogated, but that the manner in which they came to their reading can be ascertained, for example, through the repertoires of manipulations. The point is that a colleague's reading is artfully accomplished and accountably artful in character whereas the machine is limited to its algorithms. Added to this is the notion that colleagues have an idea of the artful practices of finding features in mammograms, whereas the machine's findings are relatively opaque precisely because they rely upon algorithms.

The fact that the CADe machine did not always behave as readers expected, directs us to try to understand how readers made sense of its actions. We found a variety of responses. Readers were sometimes baffled by false prompts, others they were able to rationalise by devising explanations of the machine's behaviour that were grounded in the properties of the mammogram image – e.g., that it was talcum powder, or an artefact of the developing process. In yet other cases, readers came up with explanations of the machine's behaviour that were grounded in incorrect notions of its capabilities. As we have argued elsewhere, how readers make sense of the CADe machine's behaviour influences how they use its prompts to inform their decision-making (Hartswood and Procter, 2000) and may have implications for its dependable use. This, in turn, points to general issues concerning trust – users' perception of the reliability of the evidence generated by decision aids – and how trust is influenced by users' capacity for making sense of how the decision aid behaves.

CADe tools draw attention to features on the mammogram by use of a prompt. The aim is then for the reader to examine the prompted region to see if the prompt indicates something they may have overlooked. The reader should use her own judgment (and was instructed to do so as part of the trial protocol) as to whether the prompted region contains a feature and whether that feature is benign or malignant. For a CADe tool to be used in this way, two conditions must hold. First, readers should not use the machine to inform their judgment as to whether a feature exists or is benign or malignant. Second, the reader should take the given prompts seriously enough to warrant examining the prompted regions. It would appear that this account encompasses a straightforward type of trust in relation to the machine, however, in practice we find that the situation is more complex. If we unpack the work involved in trusting the machine there are various ways in which the prompts are made accountable. First, there is the machine's biography – readers' accumulated experience of how the machine responds to different sorts of objects within the image – and second, there is what readers might reasonably expect the machine to prompt if they were to be given some understanding of how the machine works. Relying solely on either of these ways of accounting can lead to mistaken views about how prompts should be interpreted.

The question, of course, is how *do* readers construct, achieve or *make* sense of the machine? 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." (Schutz, 1964:73). 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 machine 'meant', what the machine 'is thinking', and so on. The problem is, of course, that the machine provides no such account of its actions. As Dourish (2001) writes:

> "In just the same way as they approach all other activities, they (users) need to be able to decide what to do in order to get things done. In everyday interaction ... accountability is the key feature that enables them to do this. The way that activities are organised makes their nature available to others; they can be seen and inspected, observed and reported. But this feature – the way that actions are organized – is exactly what is hidden by software abstractions. Not by accident either but by design ... the information that is hidden is information about how the system is doing what it does, how the perceived action be organised." (Dourish, 2001: 83)

> "It requires a technical approach that provides three primary features. First we need to find a way to ensure that the account that is offered of the system's behaviour – a representation of that behaviour – is strongly connected to the behaviour that it describes ... Second, we need to find a way to allow this representation to be tied to the action in such a way that the account emerges along with the action rather than separately from it ... Third, we need to ensure that the account that is offered is an account of the current specific behaviour of the system." (Dourish, 2001:85)

While it might be desirable to make a CADe machine 'self-accounting', or 'technomethodological' (Dourish and Button, 1998), in practice this may be the more complex route. While it is certainly possible to conceive of richer representations of the machine's behaviour than the bare prompts it currently furnishes, it is an open question as to whether such representations could be sufficiently contexted in the manner that would enable readers to use them in any meaningful sense. It seems to us that such representations are not accounts in themselves but *resources* for the realisation of accounts by society members. We argue that even a series of representations from which readers could choose may not provide sufficient detail to answer 'why that now?' types of question. At the very least, the provision of such accounts must be complemented by the engagement of readers and CADe developers with the machine over time – where a formulation of what has happened and how things came to this at this time is provided.

In this trial, readers were given only a brief explanantion of the image analysis algorithms used in the machine and, arguably, a more detailed explanation would have been beneficial. It is common for such algorithms to consist of a number of stages where the outputs of one feed into the inputs of the next. An ill-defined lesion detection algorithm, for example, might look for 'blobs' in an image and then decide on some other criteria to actually prompt. Clearly, a negative decision by the machine may be due to the failure of either of these stages. So, if a reader has identified a feature that they are then unsure about recalling and then make a decision on the basis of the absence of the prompt, then the weight that could be

given to machine's 'decision' in this respect could depend on whether the candidate blob was discarded at stage 1 or 2. In other words, the 'evidence' that a reader presumes they are making use of may be more or less strong in a particular instance. A case may be made for providing readers with additional information about, e.g., how a ill-defined lesion prompts are arrived at – perhaps by making available to the readers details about which features in the mammogram were detected as blobs, so the reader might be able to disambiguate the above situation in specific circumstances. This does not mean that machine would be providing an account for prompts as one might expect an algorithm designer to do – but rather furnishing the resources for a reader to be able to realise an account for themselves for all practical purposes.

Conclusions

Our ethnographic study of the use of the CADe machine in the context of a clinical trial has raised significant questions as to their impact on screening work and, therefore, as to how healthcare technologies should be evaluated. In particular, the artificial character of the clinical trial, divorced from the lived reality of everyday medical work and the various affordances of the workplace casts some doubt on its value for determining the value of healthcare technologies in a (very different) real world setting.

   In our study, we find there a number of threads to this argument. First, in everyday practice, readers use more than the mammograms provided on the viewer. They will consult previous mammograms and various documents in the patient's record. Thus, it makes sense to regard the decision as being achieved through the coherent marshalling of ensembles of evidence. That is to say, reading takes place within a familiar, 'known in common' territory of appearances – not just normal appearances, but appearances that will come to this or that, occasioning a recall or no recall decision. There are territories of known appearance which readers encounter during their training and their everyday work: further, these are known about in terms of their implications, but also in terms of the circumstances of their production. Readers are skilled in marshalling these ensembles over time, they know and use information within the corpus and know how it came to be produced: it is precisely because of this diachronically achieved familiarity that they can rely on the information.

   Second, when we look at the nature of reading work, we see that it is profoundly social in character: readers interact and come to a collaboratively realised diagnosis of the feature. Readers formulate the essential reasonableness of their work and reflect upon its achievement as a part of the day's work. Calculations of this type – formulating 'where we are' and 'what we have done' – are members' judgments, and we might ask how the CAD machine impacts on this. The CAD machine is engaged in its own algorithm based calculation, but

that this is of a very different order to the calculation work readers are engaged in – the machine cannot 'know' that it is prompting too many features and cannot be socialised as a new reader can into 'how we do things around here'. In an attempt to cope with this, readers develop biographies for the machine, suggesting what features it might be good and bad at prompting for: that is to say, readers use their professional vision to account for the machine's strengths and weaknesses by assembling preliminary accounts based on their interactions with the machine over time. It is through practices such as this that the machine might be integrated into the workplace – although how far this can happen in the limited time of clinical trials is a moot point.

Third, the docile nature of the prompts has the consequence of a reader having to make two readings – their own and one of just what the machine might have intended by that which it has prompted. Just as importantly, there is the need to provide an account for what the machine has missed and why. Readers cannot know of the machine's reasons for this as, of course, it has none: the machine is not accountable like human colleagues are and readers need to render it accountable, in and through their own practices of looking at what has been prompted and what this has come to. The work of the CADe machine is stubbornly opaque and in need of remedy to decide what it amounts to for the diagnosis.

Fourth, when considering how trust might be achieved between reader and machine, it is instructive to consider how readers trust one another. Readers are trusted to act in a professionally adequately way, and as part of that their recall decisions are credited as having value and are taken seriously. Thus, when a reader recommends recall, this is taken seriously – and other readers will treat this in a professionally proper way as having the status as a potentially recallable patient. This does not mean to say that the decision would not be contested – but that any contestation would have to be accountable – reasons would have to be furnished that both accord with readers' professional vision in such a way that signal that the candidate diagnosis has been taken seriously. That is: a candidate decision made by a colleague cannot be lightly dismissed in the same way as it might be for a novice. Within readers' professional vision there is space for versions to be contested in a way that does not challenge a reader's status as an adequately competent practitioner. Trust here is not a binary value, but rather it is fine grained. It is not just that decisions by other readers are taken seriously, but that this is done in light of a reader's biography – what it is known that a reader is more or less competent at detecting ('blobby' people, good at calcs etc). Similarly, readers show an awareness of their own strengths and weaknesses and place trust in their own ability accordingly. Trust in other readers is not something that is a given, but rather, it is an ongoing social achievement and is established afresh in and as a part of doing the work of reading. This is not to say that a reader is only treated as good as their last decision (slips and lapses are

accountable, but not necessary fatal to judgments that a reader is adequately competent as performance is accepted as being for all practical purposes), but rather in order to be treated as competent, a reader has to continuously demonstrate their competence.

Fifth, there is an important sense in which the readers use the prompts in a way that turns on a biographical familiarity with the machine that has yet to be achieved. What a prompt might mean, what it might come to in certain circumstances is an achievement of knowing the biography of the machine – something that can only be done diachronically. Practically speaking, CADe tool developers must appreciate how it is used in practice, while readers must understand the ways that the machine works to produce the prompts such that they can say 'ah yes, it is doing that because of that reason', for all practical purposes. There is no need for the readers to know in full technical detail the algorithms but, conversely, descriptions of the character 'it just does that' are not satisfactory either. The descriptions of the machine's behaviour might arise in dialogue between developers and readers around the situated uses of the machine. Given that readers develop some understanding of the machine, we may also need to consider and understand how use of machine changes over time. Reader training programmes for CADe tools may need to be designed to provide not only a resource for initial familiarisation, but also to support continued learning and evolving of practices. The issue of change over time also raises some wider issues of evaluation in that many of the benefits of CADe technology are unlikely to be evident for a substantial time after its introduction and adaptation to the particular circumstances of use. Yet, as a part of their 'grammar', clinical trials are set up to investigate evidence of *immediate* benefits.

Sixth, perhaps the most fundamental way that the clinical trials paradigm is divorced from the reality of screening practice is that no account is taken of the routine ways in which readers intersubjectively 'calibrate' decision-making within the clinic, through, *inter alia*, annotations made on the screening reporting form. The impact of this cannot be overstated; it is the intersubjective character of things such as annotations that achieves the practical work of doing mammography. It is precisely here that the CADe machine cannot take part by its very nature; that is to say it cannot collaborate with a reader except in the manner that a signpost collaborates with a traveler in pointing the way. Important questions then are how replacing double reading with CADe assisted single reading will effect the diachronic maintenance of performance previously achieved in large by the informal collaboration between the first and second reader, and what alternative strategies might be employed to facilitate maintenance of professional vision in this scenario?

Finally, our findings raise questions as to whether practice changes new technologies are intended to support are actually achievable. Related to this is the wider issue of how such machines should be designed and implemented (Berg,

1997). As IT systems and artefacts become ubiquitous, and as design becomes more entwined with the complexities of organisational working, so the challenges facing systems designers correspondingly increase. The 'design problem' becomes not so much concerned with the simple creation of new computer tools as it is with the effective integration of IT systems with existing and developing localised work practices. Such 'socio-technical' systems are mutually constituting and adaptive. This effectively takes the 'design problem' beyond the design phase to implementation and deployment, where users must try and apply any new system to their work practice (Hartswood, Procter, Rouncefield and Sharpe 2000).

Acknowledgements

References

Bannon, L. (1996): 'Use, design and evaluation: Steps towards an integration', in D. Shapiro, M. Tauber and R. Traunmüller (eds.): *The Design of Computer Supported Cooperative Work and Groupware Systems*, North-Holland, pp 423-444.

Berg, M. (1997): *Rationalising Medical Work: Decision Support techniques and Medical Practices*, Cambridge: MIT Press.

Blanks, R., Wallis, M. and Moss, S. (1998): 'A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme', *Journal of Medical Screening*, vol. 5, no. 4, pp. 195-201.

Dourish, P. and Button, G. (1998): 'On "Technomethodology": Foundational relationships between Ethnomethodology and System Design', *Human-Computer Interaction* vol. 13, no. 4, pp. 395-432.

Dourish, P. (2001): *Where The Action Is: The Foundations of Embodied Interaction*, MIT Press, Cambridge Mass.

Driessen, J. (1997): 'Worldly Interpretations of a Suspicious Story', *Ethnographic Studies*, vol. 1, no. 2.

Garfinkel, H, Lynch, M. and Livingston, E. (1981): 'The Work of a Discovering Science Construed with Materials from the Optically Discovered Pulsar', *Philosophy of the Social Sciences*, vol. 11, pp. 131-158.

Garfinkel, H. and Wieder, L. (1992): 'Two Incommensurable, Asymmetrically Alternate Technologies of Social Analysis', in Watson, G. and Seiler, S.M (eds.): *Text in Context: Contributions to Ethnomethodology*, New York: Sage, pp. 175-206.

Goodwin, C. (1994): 'Professional Vision', *American Anthropologist*, vol. 96, pp. 606-633.

Goodwin, C. (2000a): 'Action and Embodiment within Situated Human Interaction', *Journal of Pragmatics*, vol. 31, pp. 1489-1522.

Goodwin, C. (2000b): 'Practices of Seeing: Visual Analysis: An Ethnomethodological Approach', in T. van Leeuwen and C. Jewitt (eds.): *Handbook of Visual Analysis*, London: Sage, pp. 157-82.

Hartland, J. (1993): 'The Use of Intelligent Machines for Electrocardiograph Interpretation', in G. Button, (ed.): *Technology in Working Order*, London: Routledge.

Hartswood, M., Procter, R. and Williams, L. (1998): 'Prompting in practice: How can we ensure radiologists make best use of Computer-Aided Detection Systems?', in Karssemeijer, N. et al. (eds.): *Proceedings of the Fourth International Workshop on Digital Mammography*, Nijmegen, Netherlands, June 7th-10th. Kluwer Academic Publishers.

Hartswood, M., Procter, R., (2000): 'Computer-aided Mammography: A Case Study of Error Management in a Skilled Decision-Making Task', *Topics in Health Information Management*, vol. 20, no. 4, pp. 38-54.

Hartswood, M., Procter, R., Rouncefield, M. and Sharpe, M. (2000): 'Being There and Doing IT in the Workplace: A Case Study of a Co-Development Approach in Healthcare', in T. Cherkasky, J. Greenbaum, J. and P. Mambery (eds.): *Proceedings of the CPSR/IFIP WG 9.1 Participatory Design Conference,* New York, November 28th-December 1st, pp. 96-105.

Hartswood, M., Procter, R., Rouncefield, M. and Slack, R. (2002): 'Performance Management in Breast Screening: A Case Study of Professional Vision and Ecologies of Practice', *Journal of Cognition, Technology and Work*, vol. 4, no. 2, pp. 91-100.

Heathfield, H. and Wyatt, J. (1993): 'Philosophies for the design and development of clinical decision support systems', *Methods of Information in Medicine*, vol. 32, no. 1, pp. 1-8.

Heathfield, H. and Buchan I. (1996): 'Letters: Current evaluations of information technology in health care are often inadequate', *BMJ* 1996, vol. 313, pp. 1008.

Hughes, J., King, V., Rodden, T. and Anderson, R. (1994): 'Moving Out from the Control Room: Ethnography and Systems Design', in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, ACM Press, pp. 429-439.

Sacks, H. (1972): 'Notes on the Police Assessment of Moral Character', in D. Sudnow (ed.): *Studies in Social Interaction*, NewYork:Free Press, pp. 280-93.

Schütz, A. (1964): 'The Problem of Rationality in the Social World', in *Collected Papers*, *Volume Two, Studies in Social Theory*, Den Haag: Martinus Nijhoff.

Suchman, L. (1995): 'Making Work Visible', *Communications of the ACM*, vol. 38, no. 9, pp. 56-64.

Warren-Burhenne, L., Wood, S. and D'Orsi, C. (2000): 'Potential contribution of computer-aided detection to the sensitivity of screening mammography', *Radiology*, vol. 215, pp. 554-62.