

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

Mark Hartswood<sup>1</sup>, Rob Procter<sup>1</sup>, Mark Rouncefield<sup>2</sup>, Roger Slack<sup>1</sup>, James Soutter<sup>1</sup> and Alex Voss<sup>1</sup>

<sup>1</sup>Social Informatics Cluster, School of Informatics, University of Edinburgh

<sup>2</sup>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 ethnomethodologically-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<sup>1</sup>. 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.

---

<sup>1</sup> 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 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 ethnomethodologically-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<sup>2</sup> 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.

---

<sup>2</sup> 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 asterisk 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



























