Computer-Supported Cooperative Learning for Mammography

Citation for published version (APA):
Hartswood, M., Procter, R., Taylor, P., Blot, L., Wilkinson, L., & Gilchrist, A. (2010). Computer-Supported Cooperative Learning for Mammography. In *Workshop on Computer-Supported Cooperative Learning at Work, ACM Group Conference*

Published in:
Workshop on Computer-Supported Cooperative Learning at Work, ACM Group Conference

Computer-Supported Cooperative Learning for Mammography

Mark Hartswood
Edinburgh University
10 Crichton Street
Edinburgh, EH8 9AB
+44 (0) 131 650 4412
mjh@inf.ed.ac.uk

Rob Procter
Manchester University
Oxford Road
Manchester, M13 9PL
+44 (0) 161 275 1381
rob.procter@manchester.ac.uk

Paul Taylor
University College
Archway Campus
London, N19 5LW
+44 (0) 207 288 3458
p.taylor@ucl.ac.uk

Lilian Blot
Durham University
South Road
Durham, DH1 3LE
+44 (0) 1913 314 735
lilian.blot@durham.ac.uk

Louise Wilkinson
South-West London Breast Screening Service, Duchess of Kent Unit, Blackshaw Road, London, SW17 OBZ
louise.wilkinson@stgeorges.nhs.uk

Alison Gilchrist
South East Scotland Breast Screening Centre, Ardmillan House, Ardmillan Terrace
Edinburgh, EH11 2JL
alison.gilchrist@nhslothian.scot.nhs.uk

ABSTRACT
In this paper, we describe the LEMI project which is developing training tools for mammography. We focus on reporting the results from a number of evaluation sessions of a tool that is intended to provide support for trainees and their mentors in the workplace.

Categories and Subject Descriptors
K3.1 [**Computer Uses in Education**]: – *Collaborative learning.*

General Terms
Computer-supported cooperative learning, human factors, evaluation.

Keywords
Mammography, training, computer-supported cooperative work, workplace studies.

1. INTRODUCTION
This paper reports on early experiences in an ongoing project to develop a computer-based training tool supporting trainee radiologists as they develop competency in reading mammograms for breast cancer as part of the UK breast screening programme. The aim is to explore different ways that a computer-based learning environment can add value to conventional training methods, for example, by allowing trainees to access rarer presentations they might not normally encounter during 'on the job', workplace-based training.

2. BREAST SCREENING IN THE UK
Breast cancer 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 20 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 (typically a radiologist). There is a very large range of normal and abnormal appearances which result from a range of different types of breast tissue and different pathological processes. Abnormalities can be very subtle and interpretation can be extremely difficult. Types of feature that are indicators of malignancy include: micro-calcification clusters are small deposits of calcium salts 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 lesions.

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.

Relatively few existing teaching aids have attempted to provide a richly interactive educational resource that is grounded in an understanding of how radiology is actually taught and how trainees learn. Those that do [2,9] focus narrowly on the formal instructive content of didactic encounters and neglect how these are parts of broader process of learning that links acquiring craft competencies with an engagement in the cultural, professional and historical regimes in which those competencies are applied [7]. One focus for the LEMI project has been to redress this omission by posing the question as to how one might design a training tool that draws upon understandings of both the perceptual challenge and contextualised professional conduct.

3. PREVIOUS WORK

Our own previous work in this field has drawn attention to how skilled radiological decision-making is not simply a function of an individual's 'image processing' abilities, but extends to take into account local and professional conventions relating to the interpretation of images, for example, in mammography, the tradeoffs between sensitivity and specificity required to achieve a manageable recall rate in a given screening centre [3,5]. Readers are not simply trained in developing their perceptual skills, but also in their professionally appropriate application. For example, our prior work in the area points to the importance of informal social practices around double reading in maintaining an appropriate balance between sensitivity and specificity within screening centres and emphasizes the importance of locating training in the breast screening workplace [4].

Much training in mammography is conducted 'on the job': an experienced mentor 'guides' a trainee's interpretation of actual screening cases. There is a formal requirement for trainee film readers to read at least 400 screening cases every month for a year. The trainee reads 'live' screening cases and makes a decision on each, but not one that influences the outcome - the cases are still read in the usual way by qualified readers. In this way trainees can participate safely in a real setting where they have access to experienced practitioners, can see them at work and examine their opinions and contrast them with their own. Some of this interaction is mediated by the dedicated paperwork on which the trainee records their decision, and upon which, at a later time, a qualified film reader will give their view of the trainee's interpretation. Whilst this provides a rich and supportive environment both for the trainee to hone their skills and for the experienced staff to measure the progress of the trainee, it also has a series of limited horizons where there are opportunities for computer-based tools to provide assistance. One is that (except where a case was recalled and the trainee involved in the assessment clinic) trainees will rarely discover whether any of the case they read actually turned out to be normal or malignant. Another is that despite the large volume of films read by the trainee, there will only be perhaps one cancer in every 200 cases they examine, meaning that they may not see very many instances of rarer presentations.

4. THE LEMI PROJECT

The LEMI project brings together research groups at University College London, Edinburgh and Manchester Universities who are vastly experienced in problems arising from the use of computers to support radiologists reading screening mammograms, and radiologists working at two UK NHS breast screening centres [6].

Building on our previous in-depth ethnographic studies of breast screening and mammography reading practices [1,4] and following a user-centred design approach, we have developed two tools for different aspects of mammographic image interpretation. One tool, provisionally termed 'Lesion Zoo' is intended to give trainees access to a large number of abnormalities [11]. A second tool is oriented to replicating the conditions of everyday mammography reading practice and enables senior radiologists (mentors) to select cases from a distributed database of digital images to meet the specific training requirements of junior colleagues (trainees) and then provides automated feedback in response to trainees' attempts at interpretation. In this paper, we will focus on our experiences of evaluating the latter and on how we are using the results to help drive its ongoing development.

5. DEVELOPMENT AND EVALUATION

We adopted a participative approach to developing the training tool involving a series of three 'design meetings' bringing together candidate prototypes, potential users and domain experts, and iteratively worked up the tool to a point where we felt it mature enough to be used by trainees. In subsequent sessions two trainees worked through a 'training roller' of 91 cases using the tool, consisting of 30 cases demonstrating an abnormality and 61 'time proven normals'. Development and evaluation were closely coupled throughout each of these activities. Even in later sessions involving actual tool use we identified problems and refined the interaction, providing a new version with improvements and fixes in time for the subsequent session. Each of the design / evaluation / use sessions were video recorded and transcripts produced for subsequent analysis.

The initial iterative design phase by no means exhausted all of the uncertainty about how the tool might best be used or configured. Many basic questions remained open, such as: How many cases a trainee could reasonably read during a session? How might the mentor best be engaged in the trainee's use of the tool? At which points should the tool provide feedback – periodically or just at the end? Which sorts of training and training materials would benefit which students? Some questions were settled more readily than others, for example, it quickly became apparent that reading roughly 20 cases over the course of an hour's session made for a good tradeoff between comfort and making progress. With respect to other questions, using the tool in practice provided a means of exploring different possibilities and gaining experience about which modes of engagement might work best and under what circumstances – part of an ongoing process of 'learning through use' how to make the tool at home as part of wider training provision. Both trainees read the training roller in four separate sessions. These were as much as 2 months apart (Christmas holidays intervened with the first trainee's use of the tool) or as little as one week apart (the second trainee completed all her sessions during the course of a month). The first trainee had four mentoring sessions, one after each of her training sessions. The second trainee completed all of her cases before having a single session with her mentor on completion.

The tool itself allows trainees to mark a case in a number of ways. They can make one or more annotations, make a comment on a case, make a recall decision, or flag the case as one that they wish to discuss. At intervals of 20 cases the tool gave the trainee the option of reviewing their decisions, and provided automated feedback based on expert annotation of the cases. Some of this functionality accumulated during the course of the sessions in the fashion described above, for example, facilities to flag cases for discussion were available to the second trainee but not the first.

The way that the trainee's score was calculated and presented in particular was revised and refined a number of times over the course of the evaluation period.

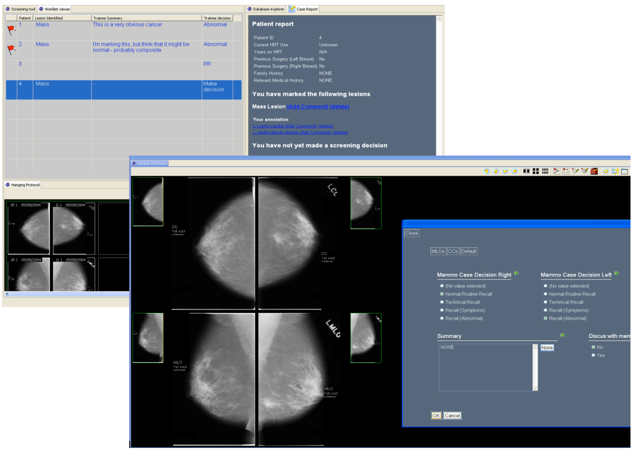

Figure 1. Training tool user interface. The interface enables browsing cases (rearmost panel), displaying the mammograms and making a decision (frontmost panel).

5.1 FINDINGS

Because the trainee continued with their 'on the job' training between each session using the digital training tool they would have accrued significant on-the-job experience of reading as a matter of routine. Often the discussion between the trainee and the mentor referred directly to the degree by which the trainee's expertise had progressed over this period. This was particularly notable for cases where an incorrect decision had been made, but where the trainee and the mentor both agreed that this was a sort of mistake that the trainee was unlikely to make now, in light of the trainee's additional experience. Given that a trainee's competence is continually developing, using the training tool in some ways provides a snapshot of their performance at a particular moment in their learning career. This then can be used as a means of measuring or demonstrating progress thus providing assurance of progress and confidence for the trainee.

Another interesting aspect of the trainees' work of reviewing their prior decisions was that of 'interpretative privilege'. Trainees, on occasion, provided an interpretative context against which their scores or marks should be judged, rather than expecting that we should take those marks at 'face value' as somehow directly representing the trainee's ability. For one trainee this was often the performance of the system itself (because flaws were ironed out and features were added to the system as the training progressed). For a second trainee it was that she often made recalls to elicit feedback, rather than because she 'really' wanted to recall the case.

On the one hand, these can be seen to be entirely reasonable considerations. For example, although it is possible to flag a case or lesion without recalling it, it is entirely likely that the second trainee only gradually became aware of these possible modes of using the training tool as the sessions progressed. Similarly, for the first trainee, improvements were made to image presentation and quality over the course of her sessions. On the other hand, though, these sorts of comments signal how:

1. trainees seek to maintain control over how their score should properly interpreted and
2. trainees sought to create a 'distance' between recorded decisions and how far those decisions can be read as indicative of the trainee's abilities.

We might see this as an important aspect of interpreting evidence relating to performance for learners. Since the general idea is for learners to progress it is perhaps important for the learner not to feel that they are 'saddled' with a particular score or mark, or to have a representation of decisions made seen to be definitive of their 'actual ability'. Any such point is always something that they need to be able to move on from or leave behind as they seek to develop their skill and expertise with any associated score or mark seen as transitional rather than definitive. In many ways this is similar to the points made above about some sorts of erroneous decisions losing their relevance as the trainee grows in competence.

Where a trainee has made a mistake, the discussion with the mentor was found often to be diagnostic, and typically involved reconstructing a plausible account of how or why the trainee might have missed or misinterpreted something, and in doing so drawing their attention to broader classes of problems or pitfalls associated with reading. Often the mentor would as well subtly point out or emphasise evidence that the trainee might have missed or weighed insufficiently, such as comparisons between views, or the possibility of teasing the lesion apart into its constituent (normal) components. It was typical also for the mentor to treat the trainee's decisions as generally credible and give the trainee room to 'argue' or 'hold out' for favored interpretations, as the mentor apparently endeavored to balance influencing the trainees' approach with allowing them flex their muscles as independent decision-makers.

There were a number of time-consuming activities associated with generating the training roller used in the evaluation. These included case selection, data entry, anonymisation, annotation, obtaining appropriate permissions for use, data correction and cleaning, and case selection. A number of steps were repeated as additional problems or demands made on the training sets were identified. This not only proved to be time consuming, but also demanding of radiology expertise, which is a very constrained resource. Some tasks proved to be very difficult to perform prospectively. For example, following the first trainee's attempt at the electronic training roller, the mentor decided that the trainee was being led to overcalling by 'difficult' normal cases in the set. The mentor decided to annotate the cases prior to the second trainee's involvement so that the tool would be able to provide feedback automatically on these 'harder' cases. The transcript of this second mentoring session shows that the mentor's annotation of normal cases wasn't exhaustive, and that the second trainee was still recalling features that the mentor had not annotated. It seems that the mentor's ability to anticipate which cases that trainees might recall is, at best, only partial.

6. DISCUSSION AND CONCLUSIONS

The development and evaluation training tools in the LEMI project has entailed exploring how a novel computer mediated training intervention can be made at home in a sophisticated regime of training delivery in a complex, collaborative professional setting. While evidently capable of generating rich opportunities for interaction between trainee and mentor, as highlighted in the sections above, use of the digital training tool also challenged trainers to develop new competencies and rethink the role and character of training delivery for different groups of students. While computer support offers flexible and configurable modes of training delivery, by the same token it raises questions as to which modes and configurations best suit which training circumstances. Issues such as establishing the most effective set composition ordering and modes of interaction for particular sorts of trainee or training objective are highly pertinent to electronic training delivery but are much less likely to arise in relation to conventional film based training rollers for reasons of impracticality.

During the development and evaluation activities outlined above many possibilities emerged for enhancing computer-support for training in mammography that we have yet to fully explore. Among these include capturing a richer account of the context in which trainees made their decisions. This could involve using audio capture, capture of tool use (e.g. mouse and keystroke events), capture by video of gestures, eye movement tracking and so on. While (as we saw above) there are cases for which there would be little interest in forensically reconstituting trainees' reading, in others it might be highly productive and generative of opportunities for additional learning. Such an approach could be extended by making available videos of trainees reading or mentoring sessions to other trainees to support 'vicarious learning' [8].

In relation to the issue of trainees measuring their progression over time there are a number of ways that computer-supported learning could assist in providing what we will refer to as 'biographical support'. Most obviously perhaps are basic numerical scores detailing correct and incorrect decisions that can be compared if this or other sets are (re) taken at a future date. The sorts of occasions detailed above, however, provide us with additional opportunities for providing richer measures of progress. For example, by marking cases so as to distinguish ones that remain hard for the trainee to interpret accurately and ones where interpretation is no longer an issue would provide a convenient way of indexing cases that the trainee might want to review and reflect upon again at some future point. Presumably, as the trainee accrues still more experience then fewer cases would remain in the latter category, and this itself provides opportunities for providing a measure of progress for trainees. Where it is possible to establish for such cases lessons that the case represents, then it would be possible to present trainees with biographical account of the sorts of lessons they have learned. Finally, knowledge of which cases the trainees still find tricky would help suggest additional training activities and cases for the trainee to complete.

Finally, our experience in the LEMI project highlights the difficulties in exploiting digital archives: selecting cases and creating training content are time-consuming tasks, heavily constrained by the availability of expertise and content must be frequently renewed. One solution might lie in the adoption of 'social media' technologies and approaches, which, in recent years, have demonstrated that the mark-up of shared content can be distributed effectively between community members, a practice often referred to as 'social curation'. It would be interesting to explore how user-generated content could incrementally add value to training sets and in doing so to shift the burden of producing, validating and assessing training materials from a small number of experts to the wider community of trainers and trainees.

7. ACKNOWLEDGMENTS

We would like to thank the UK Engineering and Physical Sciences Research Council for funding this work. We would also like to think our radiologist collaborators for giving so generously of their time.

8. REFERENCES

[1] Alberdi, E., Povyakal, A., Strigini, L., Hartswood, M., Procter, R. and Slack, R. (2005). The use of Computer Aided Detection tools in screening mammography: A multidisciplinary investigation. British Journal of Radiology, special issue on Computer-aided diagnosis, 78, 31-40.

[2] Azevedo, R. and Lajoie, SP. (1998). The cognitive basis for the design of a mammography interpretation tutor International Journal of Artificial Intelligence, 9: 32-44.

[3] Campos, J., Procter, R., Hartswood, M., Wilkinson, L., Anderson, E., Smart, L. and Taylor, P. (2005). Distributed Intelligent Learning Environment for Screening Mammography. In Proc AI in Education. p. 926-928. IOS Press.

[4] Hartswood, M., R. Procter, M. Rouncefield and R. Slack (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. 91-102.

[5] Hartswood, M, Procter, R., Rouncefield, M., et al. (2003). 'Repairing' the Machine: Case Study of Evaluating CAD Tools in Breast Screening. In Proc. European Conf on Computer Supported Cooperative Work.

[6] Hartswood, M., Blot, L., Taylor, P., Anderson, A., Procter, R. Wilkinson, L. and Smart, L. (2009). Reading the lesson: Eliciting requirements for a mammography training application. In Proc. SPIE, February.

[7] Lave, J., Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.

[8] Lee, J. (2006). Vicarious Learning and Multimodal Dialogue. In Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06).

[9] Sharples, M., Jeffery, N.P., du Boulay, B., Teather, BA, Teather, D. and du Boulay, GH. (2000). Structured Computer-based Training in the Interpretation of Neuroradiological Images. Int. Journal of Medical Informatics, 60,3: 263-280.

[10] Soutter, J., Anderson, E., Campos, J.C., Hartswood, M., Khoo, L., Procter, R., Slack, R., Smart, L., Taylor, P. and Wilkinson, L. (2003). The Role of Computer Based Training in Mammography. Royal College of Radiologists Breast Group Annual Scientific Meeting, Cardiff, November.

[11] Taylor, P., Blot, L., Hartswood, M. and Procter, R. (2010). Scoring systems in computer-based training for digital mammography. International Workshop on Digital Mammography, Spain, June.