Reading the lesson: Eliciting requirements for a mammography training application

Citation for published version (APA):

Published in:

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6B0] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Reading the lesson: eliciting requirements for a mammography training application

*aSchool of Informatics, University of Edinburgh, Edinburgh, United Kingdom; 
bCHIME, University College London, London, United Kingdom; 
cCentre for e-Social Science, Manchester University, Manchester, United Kingdom; 
dDuchess of Kent Screening Unit, St Georges Hospital NHS Trust, London, United Kingdom; 
eSouth East Scotland Breast Screening Centre, Edinburgh, United Kingdom;

ABSTRACT

Demonstrations of a prototype training tool were used to elicit requirements for an intelligent training system for screening mammography. The prototype allowed senior radiologists (mentors) to select cases from a distributed database of images to meet the specific training requirements of junior colleagues (trainees) and then provided automated feedback in response to trainees’ attempts at interpretation. The tool was demonstrated to radiologists and radiographers working in the breast screening service at four evaluation sessions. Participants highlighted ease of selecting cases that can deliver specific learning objectives as important for delivering effective training. To usefully structure a large data set of training images we undertook a classification exercise of mentor authored free text ‘learning points’ attached to training case obtained from two screening centres (n=333, n=129 respectively). We were able to adduce a hierarchy of abstract categories representing classes of lesson that groups of cases were intended to convey (e.g. Temporal change, Misleading juxtapositions, Position of lesion, Typical/Atypical presentation, and so on). In this paper we present the method used to devise this classification, the classification scheme itself, initial user-feedback, and our plans to incorporated it into a software tool to aid case selection.

Keywords: Mammography, Computer-Based Learning

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 encounter during ‘on the job’ training.

The first report in the research literature of an intelligent tutoring system for mammography was RadTutor [1]. The investigators combined a cognitive model of diagnostic reasoning with methods identified in an analysis of teaching practice and a number of well-known instructional principles. Although this tool showed promise, it did not make the transition into routine training practice. Much training in mammography is conducted ‘on the job’: an experienced mentor ‘guides’ a trainee’s interpretation of actual screening cases. While an intelligent learning environment should provide the means for trainees to learn the necessary perceptual discrimination tasks, training also involves the assimilation of professional conventions relevant to the context in which these skills will be applied [2]. To address this issue, it is critical to understand the ways in which the mentor guides and advises the trainee and to gauge the sorts of automated feedback and advice that might best support this.

In our project, we seek to extend this and subsequent work with new insights, providing a novel basis for reasoning about trainees’ performance and guiding subsequent interactions. Our objective is to determine what form of learning is likely to be most effective and which specific categories of cases readers require practice on. Our approach is to support
existing training practices by devising ways in which tools can facilitate interaction between mentor and trainee, for example, allowing each to post and respond to comments made at different levels of granularity (by lesion, by mammogram, by case and so on). We will also investigate how collaborative training might be facilitated – for example: pooling performance data, the accumulated annotations created by mentors and trainees (within and between centres); accumulating performance data, improving the accuracy of automated advice given by the system; and providing easier access to a wider community of expertise, creating a richer learning experience and evaluate the benefits.

We previously worked together in the eDiaMoND project, a major UK e-Science project that set out to demonstrate the benefits of a GRID infrastructure for the UK’s national breast cancer screening programme [3]. We developed a basic prototype computer-based and Grid-enabled training tool for mammography [4]. The initial training application, developed in eDiaMoND consists of four main components:

1. A Grid-enabled training set builder: this enables clinicians to query a federated database of cases in order to assemble training sets aimed at delivering specific learning objectives.

2. An administration tool: this allows mentors to assign training sets to trainees and to subsequently review their decisions.

3. A configurable marking engine: this asks questions of trainees and compares their responses (on a per case and per lesion basis) to the original screening decision, radiologist annotations, and biopsy data. It then scores the trainee decision and provides feedback.

4. A presentation engine: built on top of a prototype screening workstation, this provides the ability for trainees to interact with the system, for example, to work through a training set and make decisions on lesions and cases.

![Fig 1](image.png)

**Fig 1.** Screenshot of the eDiaMoND training tool.

In this paper we report two distinct pieces of work. The first is a user evaluation of the eDiaMoND prototype. The second is an analysis of the learning points. In the CONCLUSIONS section we indicate how the lessons learnt are shaping the design of a new version of the software.
2. METHODS

2.1 User Evaluation

The eDiaMoND prototype was used as a probe to engage clinicians in the process of working up requirements for more sophisticated (and useful) versions that the current project aims to deliver.

Four evaluation sessions were held where the software was demonstrated to radiologists and radiographers working in the breast screening service. Three of these sessions were with staff from one screening unit, and one was carried out as part of a project meeting attended by staff from the other. The comments of the clinicians were recorded on audiotape. In addition, one participating radiologist was involved in an extended evaluation where she worked through putting together and reporting a training set, reporting on the difficulties she encountered. The authors analysed the audiotape to identify important points and common themes, which are reported here in the RESULTS section. These observations have been used to determine the requirements for the next version of the system.

2.2 Classification of Learning Points

Each of cases in the eDiaMoND training tool database was selected for inclusion because it was identified as having particular educational significance, being of particular value for training. All the selected cases were annotated by an expert radiologist. The radiologist viewed the original films with the associated clinical information and then entered relevant information into the database via a bespoke annotation tool. One element in this involved marking the centre and diameter of any regions of interest. The annotation also involved a substantial element of structured data entry. The entered information included: Patient information e.g. age at the time the mammograms were taken, date of mammograms, interval since previous mammograms, duration of Hormone Replacement Therapy (HRT), relevant past history of surgery; Radiological appearance included what abnormalities are present in what location and what form they take (mass, calcification, distortion, asymmetry, or other). The level of suspicion associated with the abnormality was also recorded; Pathology the method used to detect the pathology of the abnormality was recorded as was the result.

Radiologists were also encouraged to enter free text observations describing the educational significance of the case. These observations were termed ‘learning points’ and their content represents a particularly valuable element in the eDiaMoND database. As part of our current project these ‘learning points’ have been reviewed independently by two members of the research team who attempted to identify a useful scheme for the classification of these points. Two schemes resulting from these analyses were then reconciled through a consensus process. The final scheme obtained is presented below under RESULTS and discussed further in the final section of this paper.

3. RESULTS

3.1 User Evaluation

The eDiaMoND evaluation sessions helped elaborate requirements in a number of ways. We report the feedback in four broad topic areas:

Basic usability issues

Many of the users’ comments concerned general design questions, not of specific interest to the digital mammography community. The way advice, trainee annotation and ‘ground truth’ were related in the feedback presentation was confusing for users. This partly arose from additional complexity introduced following an earlier evaluation where radiologists were keen to have a ‘fresh look’ at the case prior to viewing the trainee’s decision and the lesion annotation. Stronger links needed to be established between marks on the image and related textual elements.

Automated marking and feedback

A key component of the prototype compares trainees’ decisions to an ‘expert’s opinion’, or to ground truth established by biopsy. The module is configurable allowing different sorts of question to be posed about a case (where the appropriate case data is available). Trainee decisions can be recorded for lesions they identify, as well as for the case
overall. Feedback from early evaluation sessions indicated we should simplify the question regime included in our demonstration prototypes, which asked trainees to identify the type of lesion (mass, calc etc) as well as their degree of suspicion on a five point scale.

A number of radiologists suggested that the marking regime was too strict, the system expected an exact match between the trainee and expert. There was also a requirement for a metric that a trainee or mentor could use to judge improvements in performance. One radiologist suggested that while small differences in suspicion for individual lesions would not be significant, routinely under-calling or over-calling certain lesions might be. Making available aggregate scores, for example, by lesion category, might be enlightening for both the trainee and the mentor.

Integrating with existing training practices

The importance of ‘discussion’ between trainees and mentors crops up repeatedly. This sits well with our aim to complement rather than replace existing mentoring arrangements. Users stressed that providing a facility for the production of performance metrics is not always seen as a key part of training which is more about teaching and discussion. Relevant suggestions for supporting such communication included: providing comment facilities that can be attached to decisions, cases, lesions and the ability to ‘bookmark’ cases for further discussion.

With increasing experience of the trainee, training regimes (for film reading) become more focussed with more emphasis on formal assessment, and more complex training tasks (such as interpretation rather generating a binary ‘recall’ or ‘no recall’ decision). The task shifts from supporting film reading to supporting diagnostic practice. Some training needs can be supported by training sets with differing compositions (e.g. sets of ‘exemplars’ versus sets of subtle cases). Users also suggested that different ‘training modes’ might allow, for example, different sorts of questions to be posed to the trainee, and provide different sorts of feedback. For example: tutorials (providing textbook teaching), didactic modes where feedback is given per case (perhaps allowing a trainee to try again a number of times), test modes where feedback is give in summary, a blend of these where one can get case by case feedback on completion of a set; provision of links between appearance at screening, assessment and pathology, revealing progressively greater detail as a trainee ‘works-up’ a virtual case.

A further issue concerns the status of the training tool use – that some distinction should be made between different occasions of use, some of which might be more ‘serious’ than others. It was felt that trainees should be given the opportunity to have ‘exploratory’ sessions using the system that would not be scored in a formal way, nor necessarily involve training sets prescribed by a mentor.

Searching and browsing cases to build training sets

The capacity to construct training sets appropriate for different teaching goals is a key element in a useful and usable training tool. We have an archive of annotated cases that can be drawn upon. A question then is how best these might be searched. The tool developed for building training sets supported a series of parameterised queries to the Grid, as well as a facility to retrieve cases from a local drive. Query results were listed with a description of main radiological features and an indication of whether the case was normal or abnormal. Full case details could be viewed by selecting an item on the list. A facility was provided for selecting a sub-set of cases that could then be exported as a named training set, and assigned a level of difficulty on a three point scale. Again, this tool proved valuable as an initial probe for requirements for a case selection utility.

The tool did not allow the clinician selecting cases to see images at the moment of making the selection, and all clinicians saw this as a key requirement. While in some ways this is an obvious requirement, it does emphasise that training set selection is likely to remain a labour intensive task if each candidate case is to be individually scrutinised for inclusion. For a larger database of cases the process will be increasingly dependent on the quality of queries that return candidates.

3.2 Classification of Learning Points

The different dimensions of the classification scheme are interesting in that they reveal what can be construed as possible learning goals. As well as learning points that simply demonstrate typical or atypical mammographic presentations (figure 1), there are those that related to reading technique (figure 2):
- Temporal changes or stasis
- Divergence of appearances in two views
- Location within the breast
- Misleading appearances
- Juxtapositions of benign and malignant presentations
- Interpretation in the context of background tissue / other features

Figure 1: Classification of learning points relating to lesion appearance

Figure 2: Classification of learning points corresponding to aspects of reading technique
The scheme provides a catalogue of exceptions, interpretive practices and pitfalls that situate the application of skills in perceptual discrimination in the wider and more messy world of disease progression, imperfect representation, natural variation and coincidental juxtaposition. The catalogue is probably not complete and further revisions may be required. Other types of learning point were categorised as pertaining to developing an appropriate differential diagnosis, work-up strategies and resolution through use of multiple imaging modalities and further investigations, and classified under the general heading of ‘management’ (figure 3). These were more process oriented, often concerning how a normality and abnormality where finally demonstrated. These ‘complex’ learning points are discussed in more detail below.

Informal evaluation

In an informal evaluation session two radiologists were invited to comment on the two initial classification schemes we had produced. They both responded positively to hierarchies and categories presented, providing reassurance that the lesson categories we had inferred are sensible. They both immediately grasped how the classification scheme might be used:

“If we decide we want to look at spiculated masses that have not had any temporal change that are of poor technical quality we should be able to get those”

“My impression is that you’re going to code all these in all these particular categories and then we’re going to create the test roller out of this”

Discussion ensued about how the hierarchies and how they related to screening practice as well as how they might relate to training. The readers attempted to work out a top-level conceptual scheme for training that fitted with the sorts of lesson categories in the prototype classifications. Loosely speaking this consisted of three components, which roughly correspond to the distinctions drawn in relation to the more or less complex learning points described above.

**Process:** first to final mammogram. i.e. appropriate comparisons between views, over time, with additional work-up films and appropriate selection of evidence to consider next.

**Appearance:** identifying masses, calcifications and asymmetries, being able to identify typical and atypical presentations.

‘Interesting cases’ (for advanced learners): Relating pathology back to initial mammographic appearances especially for cases where the pathology result might not normally be anticipated for that presentation. More generally, exceptions and interesting cases.
4. DISCUSSION

In the course of the User Evaluation our radiologist colleagues identified different training scenarios where different training aims are appropriate, which imply different sorts of training mode that a system might support.

Table 1: Learning goals and training approaches suggested for different categories of staff

<table>
<thead>
<tr>
<th>Role</th>
<th>Learning goals</th>
<th>Training approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiographers</td>
<td>Producing diagnostic films. Identifying aspects of poor technique. Encourage reflection on own performance.</td>
<td>Examples of what was missed because of poor technique, what was found because of good technique. Rating films.</td>
</tr>
<tr>
<td>Radiographer Continuing Professional Development (CPD)</td>
<td>Tutorial based teaching of basic principles of mammographic interpretation.</td>
<td>Training sets: exemplars of each feature type. Self assessment with obvious cases. Discussion.</td>
</tr>
<tr>
<td>Junior Registrars</td>
<td>Train to a higher level with more subtle cases.</td>
<td>Metric to feedback on test rollers. Allow trainees to get a sense of their progress.</td>
</tr>
<tr>
<td>Senior Registrars</td>
<td>Conducting choice of views and action to take at review clinics. Interpretation of ultrasound.</td>
<td>For this group of trainees: “I would want to bring this lady back, do cone views, grade it as such”</td>
</tr>
<tr>
<td>Consultant (CPD)</td>
<td>Working through sets of interval cancer cases.</td>
<td></td>
</tr>
</tbody>
</table>

While there are many possibilities, it is useful to think of the problem generally as one of providing different ways of staging the presentation of case materials to trainees in ways that can be configured to support training goals appropriate to their ability. Understanding precisely which of these configurations would be most effective is essentially an empirical matter, and while discussion with radiologists such as those outlined above can be generally informative, neither they nor we know which will prove to be successful in practice. The solution to this problem can only be to follow an iterative and user-driven design and development process where requirements are gathered, prototypes are constructed and then evaluated by clinicians.

While we will not necessarily be in a position to provide a platform that can fully support all of the above training activities, an understanding of the diversity of training practice provides a useful underpinning for subsequent platform development. We can draw a number of general principles from early discussions with users that tell us in broad terms the sorts of functionality we should be aiming for, as well as the sorts of impacts on practice we might expect to see.

An important element in our work is the evaluation of the content of the ‘learning points’.

Representativeness

Some types of lesson are better represented than others within the classification scheme, some categories have a handful of learning points, while others contain tens of learning points. At the moment it is difficult to determine how important this is. It may be that by chance some lessons are underrepresented in the sample selected or it may be that some occur naturally with a low frequency. One further possibility is that there are biases in the sort of lesson descriptions made by radiologists - often single cases can be used to demonstrate more than one learning point, but radiologists may have had a preference for certain types of learning point when undertaking the mark-up. There may also be biases in the classification process. For example, whether a specific learning point is interpreted as providing a single lesson with incidental detail, or whether it could be used to illustrate two different lessons depending on the prevalent learning goals, would depend in part on the classification criteria applied. The question of representativeness could turn out to be an important one for training, because it is not obvious that one would want lessons to be represented in training at the same
frequency at which the problems they describe occur in actual practice. On the contrary, it may be useful to bias training sets towards uncommon circumstances to more quickly aid developing expertise in these areas (with the proviso that one would want to avoid biasing trainees’ expectations about frequency of occurrence). Another, more mundane, reason for wanting categories to be reasonably populated is to be able to provide sufficient resource to train and test trainees on a number of separate examples at different times.

**Complexity**

It is interesting that many of the learning points have been classified as demonstrating two or more types of lesson. The following is an example of a learning point demonstrating a single lesson:

“Malignant type calcification.”

In this case, classified as an ‘exemplar’. Here is an example of a learning point that could be interpreted as demonstrating two lessons:

“Tiny invasive ca, more visible on CC”

The case could be seen as demonstrating both a lesion with ‘subtle signs’ as well one that is best seen in the Cranio Caudal view (as opposed to in Medio-Lateral Oblique, the other view routinely taken at a screening visit) – i.e. pointing out that lesion can sometimes be seen better in one view than another. The following learning point is more complex still:

“Multiple benign opacities of both breasts. Case recalled dues to appearance of a large somewhat indistinct opacity, new since previous screen and clinically palpable. Palpable lesion aspirated (5mls cyst fluid), but persisting mixed echo lesion on ultrasound, warranting biopsy”

One can find several distinct lessons from within this learning ‘point’ (that there is something suspicious that might be confused with benign presentations, that there has been change over time, appropriate clinical work-up (palpation and aspiration), ambiguous ultrasound and the resulting need for biopsy), that are woven into a narrative describing how the various clinical and diagnostic problems (that we have picked apart to correspond to individual lessons) were identified and partially or finally resolved. This narrative character of more complex learning points is something that our classification scheme does not capture well (if at all), but is something that the radiologists continually stress as important: training should encompass and integrate the different interpretative activities both in reaching a screening decision and as well as subsequently working up the cases at assessment. That the training points themselves exhibit differing degrees of narrative quality reflects perhaps the distinct but interrelated tasks of learning specific interpretive skills (e.g. correctly assessing a lesion’s visual appearance) and appropriately combining them into a problem solving diagnostic process (e.g. selection of appropriate work-up strategy to demonstrate cancer or normality). It is likely that it will be easier to deliver automated training for distinct interpretative tasks such as the knowledge representation problem is likely to be more tractable (for example, using classifications schemes like the one presented here). Tellingly, learning points that refer explicitly to process in this classification scheme have been largely lumped together and not differentiated so well as those for lesion appearance. It is also worth noting that more complex learning points are complex often because there is (one or more) unusual aspect to the case with the lessons revolving around handling exceptions appropriately. One would then hypothesise that generally cases with simpler learning points would better aligned to the learning goals of novice trainees, whereas more complex ones would support better more experienced learners.

One of our aims in classifying learning points is to provide a naturally derived hierarchical set of categories that can aid readers in selecting from a large volume of cases in order to build tailored training sets. At the point of selection it may be useful to provide a rough index of learning point complexity by indicating the number of different categories and given learning point exemplifies.
5. CONCLUSIONS

As stated above, the aim of developing an intelligent computer-based training tool for mammography is to enhance existing training practices. We are near the beginning of the development process, and are experimenting with different modes of interaction and support provision in three broad areas:

- providing the ability to draw upon large datasets for training materials,
- providing automated feedback and advice based upon trainee decisions and performance data, and
- exploiting rich data of individual cases and training sets provided by an accumulated history of trainee decisions and annotations (that allow inferences to be made, for example, as to the difficulty of the case).

This paper presents the first iteration in this process. We have evaluated and refined an early training prototype. Initial data has concerned usability and presentational aspects of feedback and advice, as well as highlighting to sorts of knowledge required by the system in order to provide automated assistance. This has lead to a refinement of the underlying data model, and the development an experimental classification of cases based upon ‘learning point’ data. Further iterations will involve in-situ use by trainees and mentors that allow us to explore the impact and effectiveness of the above sorts of provision in a clinical context. We plan to broaden our evaluation to examine how use of the training tool impacts upon and intersects with conventional training practice. In particular, we are keen to learn precisely how the tool might support the acquisition of both ‘cognitive’ and ‘professional’ expertise (where we define the former as the ability to make a particular perceptual discrimination, and the latter as knowing what precisely to do about it in a given clinical context).

6. ACKNOWLEDGEMENTS

This work is supported by the UK Engineering and Physical Sciences Research Council. The contribution of radiologists and radiographers at the Duchess of Kent and Ardmillan Screening Units is gratefully acknowledged.

REFERENCES