# Scoring systems in computer-based training for digital mammography

Paul Taylor[1], Mark Hartswood[2], Lilian Blot[3], Rob Procter[4], Stuart Anderson[2]

[1]Centre for Health Informatics and Multiprofessional Education, University College London
p.taylor@chime.ucl.ac.uk
[2]School of Informatics, University of Edinburgh
[3]Department of Computer Science, Durham University
[4]Manchester eResearch Centre, University of Manchester

**Abstract.** A computer-based training tool was developed through a collaborative design process. The tool allows trainee radiologists to access a large number of suspicious lesions. The tool employs a certainty-based scoring system in which trainees' responses are scored not just as right or wrong but according to their confidence. Different approaches to providing trainees with feedback were considered: one based on a histogram and one using a line graph of cumulative scores. Following an initial assessment by radiologists, a revised scheme was introduced in which disagreements between trainee and expert are rated according the clinical or pedagogical significance of the error.

## 1 Introduction

Recent years have seen significant improvements in the technology used to create mammograms. Nevertheless these images are still difficult to interpret and the performance of radiologists is highly variable. [1] It is therefore important to consider whether novel tools can be used to enhance the competence of radiologists. There is considerable scope for computer-based training tools to improve the ability of radiologists to read screening mammograms. The aim of the work reported here is to explore different ways that a computer-based learning environment can add value to conventional training methods.

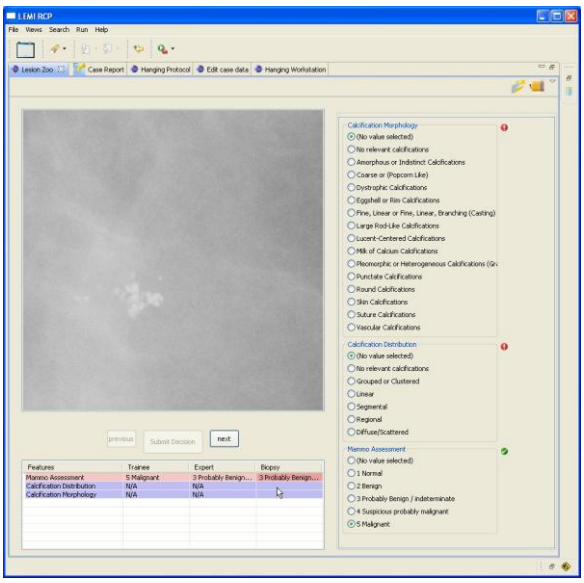
We have developed a number of tools for different aspects of mammographic image interpretation. One, provisionally termed 'Lesion Zoo', is intended to give trainees access to a large number of abnormalities. The argument is that experience of a wide range of appearances is necessary for the acquisition of visual expertise. [2] In this paper we describe this prototype, paying particular attention to the scoring of trainees' assessments, and report on expert and trainee radiologists' initial experience with the tool.

## 2 Lesion Zoo prototype

Lesion Zoo displays a sequence of selected regions of interest, each containing a lesion (either a mass or a microcalcification) and invites the user to classify the lesion using the BIRADS descriptors and to assess it, using the BIRADS assessment categories. Users are given case by case feedback on their performance and finally summary statement of their performance over the set of images (see Figure 1).

Lesion Zoo uses a database of 300 annotated images. Each of the cases in the database was selected for inclusion because it had particular value for training.[3] 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 Lesion Zoo software displays these regions of interest, automatically selected from the database images and invites users to enter a description of the lesion via a menu of descriptors based on the BIRADS terminology.

**Figure 1:** screen shot of the Lesion Zoo application. The trainee's description of the image is entered via the menu on the right. Feedback is given in the table at the bottom: the trainee disagrees with the biopsy and the expert decision (note the darker red tone for the biopsy).




### 2.1 Feedback

Feedback provided to trainees uses 'certainty-based marking'.[4] This is a scheme that aims to assess the change in confidence of a learner. Key to the approach is the collection of data not just on the accuracy of a student's responses but also on his or

her confidence in that accuracy. The point is not, as in ROC analysis, to gain a threshold-independent measure of accuracy. Instead, data about the confidence a student has in his or her clinical judgment is incorporated into the scoring in order to allow an assessment of the practical value of the judgment. Confident but inaccurate judgments are dangerous. Accurate but unconfident judgments are not useful to the trainee. Useful knowledge leads to responses that are both confident and accurate. Using this approach, responses are scored according to the scheme in Table 1 and aggregated to create a single score. The weights were chosen to reflect practice using the approach elsewhere and may need to be modified in this setting.

**Table 1;** Scheme for certainty-based marking

| Certainty | 1 (Low) | 2 (Mid) | 3 (High) |
|---|---|---|---|
| *Score if correct* | 1 | 2 | 3 |
| *Score if incorrect* | 0 | -2 | -6 |

We developed two methods of visualising the results of certainty-based marking to provide feedback in Lesion Zoo: a histogram showing how the responses are distributed across the six categories in Table 1 and a graph showing how a trainee's performance changed as cases were attempted, illustrated in Figure 2.

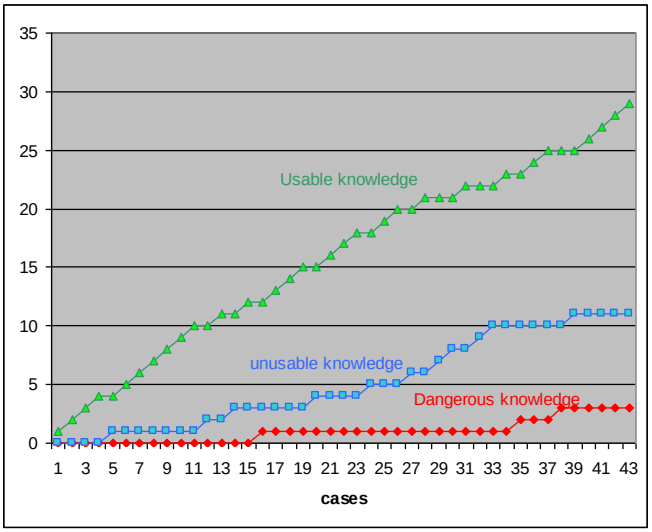
## 3 Refining the prototype

The Lesion Zoo prototype has been developed in close collaboration with staff in two UK NHSBSP screening units. It has been used on an experimental basis by expert and trainee radiologists to collect feedback on the design and the scoring system.

### 3.1 Methodology

To gather feedback on Lesion Zoo, we used a qualitative, observation-based methodology [4]. We have found in previous studies of the use of computer-based tools in mammography that this approach is very effective for understanding how people use these tools [5].

**Figure 2:** Number of cases classified as useful knowledge (mid to high confidence, correct response), unusable knowledge (low confidence) and dangerous knowledge (mid to high confidence, wrong response) as a function of number of cases attempted. The x-axis corresponds to the cases done so far by the trainee. Ideally, the green line should rise to a slope of 45° while the red and blue should approach the horizontal.




We gathered qualitative data over four use sessions (two observed and written notes taken, one videoed and one where feedback was elicited from the user after use) and the design team examined the data for emerging themes. Typically the sessions involved encouraging users to work through a set of 10 or so lesions and to verbalize their impressions. These sessions provided the basis for iterative refinement to the lesion zoo tool. Although informal and lightweight we found the approach to be helpful and informative, with users demonstrating consistent and coherent views of the tool's more or less appealing characteristics.

### 3.2 Results

Some participants responses indicated the sorts of reflection and deliberation that we hoped would be elicited by Lesion Zoo. On other occasions, participants' responses were not so straightforwardly positive. In particular there were a number of occasions where they disagreed with the characterizations of lesions, and were concerned that the scoring mechanism was unfair when penalising and rewarding certain sorts of responses. One participant using the Lesion Zoo pointed out how it can be hard to make a forced choice between BIRADS descriptors when these may not necessarily be mutually exclusive.

"Lesions can have two appearances – it can be spiculated in one part and ill-defined in another"

That a lesion might not always unambiguously be assigned to a single category, and to penalise disagreement with the expert on the basis of what is seen as a judgement call is seen as unfair. The tool should not be unnecessarily discouraging, for example, by penalising the trainee for trivial classification errors when, otherwise, they are demonstrating competence.

It became increasingly obvious that Lesion Zoo would need to have a notion of which distinctions are clinically important, and which not, and to be able to moderate the marking accordingly. For example, conflating round and oval as descriptors of mass shape is a less critical error than calling a malignant presentation benign. It also became clear that there can be a number of dimensions against which the appropriateness of a response might be judged. For example, while confusing two different benign descriptors would not adversely consequential for the patient, it may indicate problems with the trainee's conceptual grasp of the subject matter.

One trainee radiologist categorically disagreed with the expert, stating that she had recalled a lesion similar to the one presented that had turned out to be a cancer, whereas the expert opinion on the lesion presented was that it was benign.

"I saw a cancer the other day that looked just like that – I disagree with the expert"

The point here is not whether on this case the classification was appropriate, or the trainee's memory exact, but that there should be some room to allow for disagreement between professionals to be voiced where feeling is strong.

Feedback from the participants suggested that one assumption behind certainty-based marking, that there was a simple right answer in the interpretation of these cases, was inappropriate. We subsequently undertook an exercise to establish what sorts of deviation from expert opinion might be more or less consequential to enable us to moderate the marking appropriately. Three consultant radiologists were asked to complete a confusion matrix for each of the BIRADSs rating questions, scoring the severity of the disagreement for each combination of trainee and expert opinion.

The experts' rankings for BIRADs assessment are shown in Table 2. Similar tables of rankings were obtained for the different descriptors used to characterize masses and microcalcifications. These will be used to weight the scores assigned to disagreements.

**Table 2:** Experts' (n=3) rankings for the significance of disagreements between expert and trainee over BIRADs assessments

| | | Trainee | | | | |
|---|---|---|---|---|---|---|
| | | **M1** | **M2** | **M3** | **M4** | **M5** |
| Expert | **M1** | | 2 | 3 | 4 | 5 |
| | **M2** | 2 | | 2 | 4 | 4 |
| | **M3** | 3 | 3 | | 2 | 3 |
| | **M4** | 5 | 5 | 2 | | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | **M5** | 5 | 5 | 4 | 1 | |

## 4 Discussion

Our study provided insights into how the tool would be used in practice and revealed a number of ways in which it could be improved. We have provided a facility for trainees to express their opinion as a free text comment. We also solicited free text comments from the experts assessing the lesions so that they also had room to express how the lesion matched the BIRADS descriptors. Comments could be used to state, for example, that the lesion is typical or atypical member of the chosen category. Expert comments are made available to the trainee after they have made their own assessment of the lesion.

Another issue concerned radiologists' application of the five-point BIRADS scale for suspicion (1. Normal to 5. Benign). One junior film reader suggested that her more experienced colleagues tended to be more confident in calling a lesion M5. A senior radiologist (independently) made the same point when she said that junior readers are more cautious in how they grade lesions. At screening there is no difference in clinical outcomes between grading lesions as M4 or M5, as both will result in a recall for further assessment, thereby providing leeway for junior radiologist to make conservative assessments without impacting on patient care. The less experienced radiologist drew attention to this, partly because she thought she might improve her score if she rated lesions in a way that emulated how she presumed the expert would rate them. In other words, she was attempting to 'second guess' the expert.

This exemplifies a more general problem observed with using game-like environments for education, neatly expressed by Conati and Lehman, who devised a 'micro-world' game to teach principles in physics [5]: *"Our observation of student players indicates that it may be possible for a student to become skilled in solving a problem in game terms, i.e. without significantly improving their physics knowledge*". So, by using a game format, one may introduce elements of fun, competition, ease of playing and focus on a specific task, but risk a worse than might hoped for transfer of skill to the real world domain.

We have learned lessons experimenting with confidence based marking (CBM) in a novel domain. While sympathetic to CBM's goals of reducing trainees' motivation for simply guessing when uncertain, and encouraging reflection on, and awareness of, lacunae in trainees' understanding of a topic, we found its application in a mammography task somewhat tricky in a couple of respects. Part of this stems from the fact that CBM was designed to be applied to domains where it is possible to unambiguously distinguish between correct and incorrect answers. This is evidently not the case in mammography, where variation in lesion appearance in continuous, rather than discrete and always neatly categorisable, as the BIRADs classification scheme might imply. Thus, the evident discomfiture of the participants in the exercise above where reasonable or non-consequential disagreements were penalised, and of one consultant radiologist who completed a session only to have her decision-making described as 'dangerous' by the tool.

## 5 Conclusions

The evaluation of the Lesion Zoo has enabled us to have a better understanding of how it would be used and to identify criteria for a successful marking scheme for trainees' responses.

First, the scheme must be seen to be fair. BIRADS descriptors are not all mutually exclusive, assigning a lesion to a single category is something of a judgement call. Penalising the trainee for errors in such cases is seen as unfair. It became increasingly obvious that the system had to have a notion of which distinctions are the important ones, and which less so, and to moderate the marking accordingly. For example, conflating 'coarse' and 'eggshell' calcifications is a less critical error than calling a malignant presentation benign. However, while confusing two very different benign descriptors (e.g. 'skin' and 'suture' calcifications) would not have any consequence, it may indicate problems with the trainee's grasp of the subject matter.

Second, we must also allow for disagreement between professionals. We have provided a facility for trainees to express their opinion. This provides room for trainees to express disagreement with the expert classification or a justification for their own. It is hoped that this also will provide a means of softening the impact of the marking scheme by providing an opportunity for readers to voice professional disagreement, as well as creating a valuable resource for understanding how trainees interpret lesions. We also collected a free text comment about each lesion in the zoo from three expert radiologists. The aim of the comment is for the expert to be able to provide a rationale for their classification and rating decisions, for example, if the lesion exemplifies the category, or if its classification is problematic.

It is anticipated that these expert comments (made available to the trainee after completing each lesion) will help reinforce the trainee's grasp of the relation between the lesion's appearance and its classification, as well as alerting them to less clear cut cases where their own opinion might more reasonably deviate from the expert.

A collaborative design process has resulted in a novel tool that includes a sophisticated assessment of the significance of a trainee's disagreement with an expert and assesses this in conjunction with a measure of the trainee's confidence in order to provide a measure of how useful knowledge is developing.

## References


1. Goddard P, Leslie A, Jones A, et al. Error in radiology. *Br J Radiol*. 2001 Oct; 74 (886):949-51
2. Taylor,P. A Review of Research into the Development of Radiological Expertise: Implications for Computer-based Training. *Academic Radiology* 2004 ; 14, 1252-63
3. Brady M, Gilbert F, Lloyd,S., et al. eDiaMoND: the UK's Digital Mammography National Database. in Pisano E. (ed) Proceedings of the International Workshop on Digital Mammography, 2004
4. Gardner-Medwin AR (1998) Updating with Confidence: Do your students know what they don't know? Health Informatics 4:45-46.
5. Bryman, A. Social Research Methods. Oxford University Press 2008.

6. Alberdi E, Povyakal A, Strigini L , et al. The use of Computer Aided Detection tools in screening mammography: A multidisciplinary investigation. Br J Radiol (2005), special issue on Computer-aided diagnosis, 78, 31-40.
7. Conati C and Fain Lehman,J EFH-Soar: Modeling Education in Highly Interactive Microworlds . In Lecture Notes in Artificial Intelligence. Advances in Artificial intelligence, AI-IA '93 Springer Verlag, Berlin 1993.