Learning from experience: maintaining flexibility in the production of training materials to support breast cancer screening.

Mark Hartswood[1], Stuart Anderson[1], Lilian Blot[2], Alison Gilchrist[3], Rob Procter[4] and Paul Taylor[5]

[1] School of Informatics, University of Edinburgh

mjh|soa@staffmail.ed.ac.uk

[2] School of Engineering & Computing Sciences, Durham University

lilian.blot@durham.ac.uk

[3] South East Scotland Breast Screening Centre, Edinburgh

Alison.Gilchrist@nhslothian.scot.nhs.uk

[4] Manchester e-Research Centre, University of Manchester

rob.procter@manchester.ac.uk

[5] Centre for Health Informatics and Multiprofessional Education, UCL

p.taylor@chime.ucl.ac.uk

Abstract. This paper is about our experiences of working with clinicians to create and use an archive of 'training cases' as part of a research project aimed at developing training tools to support cancer detection in the UK breast screening programme. As the project progressed we became increasingly interested in how the clinicians learned about the character of the archive as they worked with its contents, and the ways that this work shaped the archive in ways that typically lead to heterogeneity and uneven data quality. Critically examining our approach to training cases curation and use with resilience engineering concepts in mind revealed significant bottlenecks rooted in normative assumptions relating to authority over content, completion of training sets and batch modes of working. We argue that social approaches to curation offer more promising ways of mobilising complex data that can both overcome bottlenecks associated with conventional approaches as well as capitalising on learning in use.

1   INTRODUCTION

In the UK current mammography training highlights the constraints of radiology apprenticeships. Trainees make 'practice' decisions on live screening cases and receive feedback from their mentor, but they rarely learn whether cases actually turn out to be benign or malignant. Trainees and mentors alike often lack a formal means of measuring and evaluating progress over the course of a training rotation and, despite reading a large

volume of films, the trainee will only see perhaps one cancer in every 200 cases they examine, making encounters with rarer presentations unlikely [Soutter et al., 2003]. Furthermore, there is significant variation in radiologist performance across the breast screening programme [Duijm1 et al., 2009] indicating an important role for ongoing training as part of continual professional development.

Access to training libraries drawn from digital case archives promises to address all of these shortcomings, and although recent technological advances make this a tantalising prospect, a number of significant barriers still hinder progress. Our research shows that selecting cases and creating training content are time-consuming tasks, heavily constrained by the availability of expertise, that high standards of data quality are difficult to maintain, a fact compounded by the need to frequently renew content as mammography practice changes. This leads us to propose that social curation (i.e. pooling the efforts of multiple contributors [e.g. De Roure et al., 2010; Lui, 2010] providing feedback (e.g. annotations, ratings, etc.) based on their experiences of using the content) may provide a solution.

This paper reports our experiences of data curation during an EPSRC-funded project (LEMI), which explored the use of digital clinical materials (mammograms, images) to deliver training via a computer-based training environment. The project entailed a variety of activities where participants handled and worked up the clinical materials, fashioning them into training packages via a number of distinct steps which included case selection, mark-up, cleaning and allocation into discrete training packages. The majority of these activities were video recorded and transcribed to track the unfolding work of generating and using training materials. Our experiences of undertaking this work alongside mentors and subsequent analysis of the transcripts revealed the following:

1. Each data handling step provided opportunity for learning, both about individual cases, specific sets, and the archive as a whole.

2. Use of the data created the requirement to change, amend or mark-up the data, which in turn, because of constrained resources led to heterogeneity in data quality across the archive, and contrary to our intention.

3. We naturally fell into working in sessions on subsets of the data, organised around the completion of specific tasks (for example, mark-up, correction, training) with a division of labour where certain participants had authority over particular aspects of the data.

Although we organised a series of data handling sessions sequentially with the completion of one feeding into the beginning of the next (set compilation following data cleaning, for example) we found that the actual process was much less linear, and far more iterative than we anticipated. For example, downstream use in training and mentoring triggered further rounds of selection, cleaning and mark-up. We also found that our careful step-wise approach to crafting training packages lead to significant delays in their production. Our approach involved a number of (up to 2 hour) sessions in which one or more clinicians attended to distinct tasks (e.g. annotation) for a particular

batch of cases. This approach lead to significant bottlenecks because of dependencies (e.g. set compilation following annotation), the emerging need for iteration (e.g. discovering that further mark-up or data cleaning was required), coupled with the difficulty in scheduling access to scarce and expensive clinical expertise.

Taking these findings together we argue in this paper that more robust training materials can be generated more effectively if we rethink our approach to their production in ways that take into account of the non-linear character of the process and take advantage of the learning that data handling occasions. More effective ways of transforming clinical materials into training packages is important because many of the anticipated benefits depend upon being able to mobilise large numbers of examples and embed these into everyday practice.

2 LEARNING ABOUT THE ARCHIVE

As outlined above, the LEMI project involved a number of different activities where archived clinical materials were handled, modified or viewed by project participants, who included the training environment developers, project radiographers, radiologists and trainees. Below we outline three distinct but related sorts of learning that occurred as a side effect of handling training materials, and born out of the necessary intimacy participants had with the materials they were working up and using.

First, our participants learned which cases and case materials it was necessary to include in training packages, how packages should be structured and presented, and to what standard. Although radiologists already had a sophisticated understanding of counts as 'good examples' for different learning outcomes, these understandings were continuously elaborated as they re-encountered cases in different contexts of mark-up, selection, training delivery etc. Second, mentors learned how the trainees responded to the cases; the sorts of mistakes they were prone to making and the sorts of feedback that the system could usefully provide. Third, mentors learned something about the properties of the training materials themselves; through their own repeated exposure and through the trainee's responses they began to gauge the quality and appropriateness of the archive available to them to select from, and the difficulties posed by the set the trainees had completed. The following example illustrates some of the character of this type of learning. The trainee (Tina) and her mentor (Jane, a radiologist) are discussing a training case that Tina had read while using our training environment:

Tina: Maybe that's one to take out because it is misleading … would be my comment.
Jane: Not, not, not necessarily I mean a learning point there you see if you'd had the previous [films]... as a learning point I would probably have put on that "this ladies come off HRT, look at the difference".
Tina: Mmmmm or something like that. Which would be quite nice. Learning point about HRT [Hormone replacement therapy].

The case referred to above was found to be confusing when used to actually deliver training, despite it having previously survived rounds of scrutiny and review to be included, first, in the archive itself and, second, in this training package aimed at

teaching film reading. The trainee's advice is to omit the case from the set in the future. The mentor, however, sees an opportunity to adapt the case to provide a different sort of lesson based around the (possible) cause of the original confusion, namely to demonstrate what happens to the appearance of the imaged breast tissue when a women stops taking hormone replacement therapy between screening rounds. The participants have learned that this case has the potential to be confusing in its current use context, but that with a bit of work, might be an excellent exemplar in a different one. We found many similar examples of 'learning in use', but also found that we typically failed to capitalise on them. Our tacit adoption of a linear approach to data curation meant that certain sorts of mark up were only possible during specific stages in the training cases' production. So, at the point where trainees or mentors discover it would be useful to add something new to a case, the window of opportunity for doing so had long since passed.

3   SHAPING THE ARCHIVE

As well as learning about the archive, various data correction and use activities shaped the archive in different sorts of ways. As those parts of the data that survive the initial selection process go through multiple cycles of use, repair and work-up they also become increasingly familiar to and better understood by users. Conversely, this meant that a large part of the data archive became increasingly neglected. In the context of this short project, with limited resources, the effect has been to 'funnel down' from a large inconsistent, untidy, unkempt, largely unfamiliar resource to well understood, worked up, better ordered and tidier, but more compact one. Our starting point was a digital case archive from a prior project (eDiaMoND [Power et al. 2004]), half of which was annotated and all of which was in need of cleaning. Our initial plan was to correct all of the eDiaMoND data, but time constraints led us to focus our efforts, and in the end we strove to clean those cases in the dataset that had been already been annotated. Focussing attention on subsets of the eDiaMoND database in response to resource constraints became something of a pattern, leading to heterogeneity in data quality as some cases received more attention, becoming better 'worked-up' and 'cleaner', whereas others languished unused in the archive. Cases that receive less attention become increasingly dispreferred (for example, as outlined above, our cleaning only of annotated data), despite representing a significant and potentially useful resource. Similar decisions were made on other occasions. For example, in the following transcript, Mark (a training environment developer) is working with Jane, Mary (radiologists), and Ann (the project radiographer) to select cases from the archive for inclusion in a training set:

Mark: Case number 1 doesn't have any pathology – if we had a look at the pathology then we could always come back to it and find out what it was.
Jane: Yeah, we don't need to necessarily use it.
Mary: we don't – we got so many to choose from.
Ann: yeah, that's right, we've got so many to choose from we don't…

Here we see efficiency being traded for thoroughness in the handling of case materials, as it is far easier for Jane and Mary to search the archive for another example, rather than fixing the example at hand. On the other hand, the tools provided do little to support

thoroughness as there is no facility to flag the case as being in need of attention, an omission arising from our presumption that data emerging from prior preparation steps would now all be clean.

4 PROCESS AND AUTHORITY IN DATA CURATION

'Sets' and 'sessions' dominated our initial approach to curating training cases. 'Sets' are collections of cases that share some common feature, or serve some common purpose. 'Sessions' are meetings where a range of trained staff work with a set for a variety of ends including: repair or mark-up of cases, training delivery, or set compilation. This approach was thorough and emphasised the depth and sophistication of data handling amongst practitioners, the heterogeneity of medical roles, and the complexity of data quality issues. These features all posed significant challenges for the creation of training sets. For example, a sophisticated working division of labour emerged over who had authority over which parts of the data in the archive:

Project radiographer: allowed to: withdraw cases from the archive, and select films for inclusion in the database; see pre-anonymised cases and undertake anonymisation work; make some interpretative judgements (e.g. 'breast density', image quality); make decisions about how the 'normal' set might be compiled, and undertake case selection.

Radiologists: maintained a large degree of interpretative privilege over the image, including tasks such as annotation, and descriptions of location, degrees of suspicion or difficulty. They also took responsibility for selecting abnormal cases, compiling training packages and producing pedagogic content.

The training environment developer (Mark): performed occasional correction of anonymised data, such as assigning films into prior and current screening rounds, but was explicitly barred by the project's ethics approval from handling non-anonymised data. Performed tasks at the user interface under direction, but would hand the controls back to the radiologist to perform certain operations, such as marking the lesion.

Production and maintenance of training sets involves time consuming and costly, iteratively applied procedures, bringing together scarce expert labour, hard to retrieve documentation and knowledge of context. Part of the time-consuming nature of these tasks revolves around an implicit, a-priori orientation to "perfecting" the set and linear approaches to data curation. This led us to reconsider how set, session and divisions of labour could be reorganised for the more efficient production of robust training cases:

Rethinking the session: Sessions typically involve a single task (annotation, training, data collection, mark-up) performed by one or more people at a particular time, typically working through a batch of cases. It may be possible to 'farm out' component tasks to a broader community of experts and trainees, break them into smaller components, and interleave them with other activities. For example, rather than attempting to annotate normal breast features that a trainee might naively interpret as abnormal, why not only mark up those regions that the trainee have erroneously recalled?

Rethinking divisions of labour: Some interpretative judgements could be left to emerge as a function of multiple uses of the set made by trainees. Thus the 'difficulty' of

a case could be established 'empirically' by how hard it proves to be for trainees to interpret correctly. While there are some aspects of interpretative privilege that it would not be sensible to remove from the hands of the radiologists, there are others that arguably would be better performed by drawing on experience in use.

Rethinking the set: Fixed sets provide a way of making performance comparisons and have known properties in terms of training outcomes. Allowing the trainee to access a stream of training cases, the composition of which is adjusted continuously to address specific aspects of the reader's performance, is a radical approach to achieving efficiency while potentially maintaining an appropriate coverage of cases.

The data handling practices outlined in this paper reveal a situation where, in practice, data is continuously re-elaborated with each use. If we relax the stricture of attempting to get all of the data 'perfect' prior to use; and, instead, treat each use event as an opportunity to elaborate and add value to the data archive, we might be in a position to more effectively mobilise complex data such as mammograms for sophisticated use settings such as training. This approach would need the provision of appropriate tools to support a variety of forms of data elaboration (e.g. annotation, capturing history of use) at each use encounter.

5 CONCLUSIONS

One way of conceptualising learning and shaping activities outlined above is to consider how the cases feature serially in a number of 'use contexts', starting with their origins as a clinical artefact, their transformation into a training resource through selection and abstraction, their allocation to specific training sets, and their use in multiple rounds of training and mentoring. On each occasion of use there is the potential for learning, the possibility of repairing or elaborating the case, be it the trainee's decision, or the mentor's comments on that decision, and utility in accessing prior use contexts (e.g. is this a case that trainees more frequently struggle with; what have mentors said about this feature before?). So, training cases can be thought of as artefacts in a continual state of production, where each subsequent use provides an opportunity for discovery or confirmation. This provides a powerful argument for capturing and providing a trace of the training case's 'biography', that is, for opening up opportunities to add to or elaborate the training case data *at the point of use*. One way of visualising a case biography would be analogous to the audit trail of revisions for wiki pages. This would support exploration of how the case faired in earlier use contexts when users make decisions about the current one. To give some examples, the training case 'biography' could reveal which training packages the case belongs to; how easy it has been for trainees to interpret the case in each; the features within the case that trainees tend to call; the comments and mark-up the case has accumulated in use, and so on.

In effect, we are advocating a social approach to curation whereby curation activities are broken up into smaller component tasks, distributed between a large constituency of users, and where each encounter with the training cases is taken as an opportunity to undertake a small (but collectively significant) amount of curation 'housekeeping'.

We believe that our insights, while pointing to more efficient and robust approaches to curating training cases, also correspond to the principles of resilience engineering [Hollnagel & Woods, 2006]. A linear approach to curation, coupled with the presumption of success at each step led to us losing *control* of the curation process; we had to *reactively* undertake iterations to effect repairs that meant we never were completely on top of data quality issues. And while the archive users had *knowledge* and *competence*, they lacked the appropriate *resources* (tool support) to effectively capitalise on their learning through use and handling of training cases. This led to degradation, both of our ability to generate training cases and of their quality. In turn, this limits their potential for impacting positively on the performance of breast cancer screening.

While resilience engineering principles helped inform our understanding of the problem, we turned to social curation approaches for potential solutions, and would like to draw attention more generally to the synergies between 'social' technologies and media and resilience engineering precepts relating to anticipation, knowledge and learning.

6 ACKNOWLEDGEMENTS

REFERENCES

De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., Lin, Y. & Poschen, M. (2010) Towards Open Science: The myExperiment approach. Concurrency and Computation: Practice and Experience, Vol 27(17), July, 2335-2353.

Duijm1, E.M., Louwman, M.W. J., Groenewoud J.H., van de Poll-Franse, L.V., Fracheboud, J. & Coebergh, J.W. (2009) Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome *British Journal of Cancer, 100*, 901–907.

Liu, S.B. (2010) Trends in distributed curational technologies to manage data deluge in a networked world. *Upgrade, 11(4),* 18-24.

Hollnagel. E., Woods, D.D. (Eds.) (2006) Resilience Engineering. Ashgate.

Power, D., Politou, E., Slaymaker, M., Harris, S. & Simpson, A. (2004) A relational approach to the capture of DICOM files for Grid-enabled medical imaging databases. ACM Symposium on Applied Computing, Special track on Computer Applications in Healthcare. Cyprus, March

Soutter, J., Campos, J., Hartswood, M., Jirotka, M., Procter, R., Slack, R. & Taylor, P. (2003) Grid-Based Mammography Training. *Hospital Radiologist*, *5(6)*.