



City Research Online

City, University of London Institutional Repository

Citation: Alberdi, E., Taylor, P. and Lee, R. (2004). Elicitation and representation of expert knowledge for computer aided diagnosis in mammography. *Methods of Information in Medicine*, 43(3), pp. 239-246. doi: 10.1267/METH04030239

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/1586/>

Link to published version: <http://dx.doi.org/10.1267/METH04030239>

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

TITLE:

Elicitation and Representation of Expert Knowledge for Computer Aided Diagnosis in Mammography

AUTHORS & AFFILIATION:

E. Alberdi, P. Taylor & R. Lee - CHIME, University College London, UK

CORRESPONDENCE ADDRESS:

Eugenio Alberdi, CSR, City University, Northampton Square, London EC1V 0HB, UK
Tel: +44 20 7040 8424 Fax: +44 20 7040 8585 Email: e.alberdi@csr.city.ac.uk

SUMMARY

Objectives: To study how professional radiologists describe, interpret and make decisions about microcalcifications in mammograms. The purpose was to develop a model of the radiologists' decision making for use in CADMIUM II, a computerized aid for mammogram interpretation that combines symbolic reasoning with image processing.

Methods: In Study 1, eleven radiologists were asked to 'think out loud' as they interpreted 20 sets of calcifications. Participants used 159 terms to describe calcifications. We used these data to design a scheme with 50 descriptors. In Study 2, ten radiologists used the scheme to describe 40 sets of calcifications. We assessed the capacity of the terms to discriminate between benign and malignant calcifications, testing them against radiologists' assessments of malignancy and follow-up data.

Results: The descriptors that were found to be the most discriminating in Study 2 were included in CADMIUM II's knowledge base. They were represented as arguments for either a benign or a malignant diagnosis. These arguments are the central component of the decision support provided by the system. Other components are: image processing algorithms for the detection and measurement of calcifications and a set of rules that use the measures to decide which of the arguments apply to a given set of calcifications..

Conclusions: Preliminary evaluations of the CADMIUM II prototype reinforce the value of representing explicitly decision making processes in computer aided mammography and of deriving these processes from image processing measurements. Decision support is presented here at a level of description that is both relevant and meaningful to the user.

Keywords: computer aided diagnosis, radiology, knowledge elicitation, knowledge representation, image processing

1. Introduction

Advances in digital mammography have led to the development of a variety of computer decision aids for mammogram interpretation [1-2]. The most common approach is the use of image processing algorithms for the detection of abnormalities [3]. Many systems also assist in providing a diagnosis for the detected findings by using classifiers such as neural networks or k nearest neighbour classifiers, for example [4-7]. Many such systems obtain remarkable results in terms of diagnostic accuracy. However, the diagnoses they produce are based on complex numerical processing whose rationale will not be apparent to the user. In contrast, human factors research suggests that computer aided diagnosis would benefit from explicit explanations linking diagnostic decisions to the abnormalities detected in the mammograms [8].

We are following this approach in CADMIUM II, a system for the diagnosis of mammograms that combines image processing with symbolic representations of clinical decisions [9]. The system uses symbolic reasoning to relate information obtained from image processing to the decisions that radiologists take. The system is an advanced version of CADMIUM, a prototype that was successfully evaluated with radiographers trained to interpret mammograms [10].

CADMIUM II explicitly represents the decision making process involved in the interpretation of mammograms, including the candidates (options) and the arguments that would support each candidate. Up to now, we have concentrated on a single problem: the differential diagnosis of calcifications. For this decision, the candidates are terms that are indicative of the risk of malignancy and the arguments are statements about confirmed properties of the observed calcifications.

To our knowledge, no other system for mammography has been developed that combines symbolic decision making and imaging. Some decision aids do use symbolic reasoning and can provide useful information about diagnoses [11-13]. However, such systems require the features and measurements of mammographic appearances to be introduced manually by a human expert. In CADMIUM II, in contrast, the process is fully automated as such measurements are obtained from the results of image processing operations.

This requires three elements: 1) a knowledge base about radiological decisions; 2) image processing to detect and characterize calcifications; and 3) a mapping between imaging measurements and the symbols in the knowledge base. In this paper we are

concerned only with the first part: the acquisition and representation of the relevant radiological knowledge to be incorporated into the knowledge base.

2. Acquisition of Radiological Knowledge

The aim of knowledge acquisition is to obtain a body of knowledge which is as complete, consistent and correct as possible [14]. Most common methods of knowledge acquisition involve either eliciting information from human experts or extracting knowledge automatically from data. Methods of knowledge elicitation from humans range from informal or semi-structured interviews and observations to more structured methods like the transcription and analyses of verbal reports or conceptual techniques such as laddering, hierarchical sorting, graph construction, etc. [15]. Some of these elicitation techniques can be conducted through the interaction between human experts and purpose built computer tools [16]. Examples of methods of automatic knowledge extraction from data are machine learning (or data mining) [17] and natural language understanding techniques [18].

An alternative approach is the extraction of knowledge from written documentation [15]. This approach was used in the development of the original CADMIUM prototype. Statements from review articles, textbooks and monographs were transcribed and incorporated as rules. However, the resulting knowledge base was considered by radiologists to be confusing and in some respects erroneous [10]. One problem was that many statements were ambiguous if presented without other contextual information provided in the texts. We also found that there is considerable variation amongst authors in terms of what aspects of the calcifications they consider typical or discriminating.

Recent years have seen an advance in the development of standardized health terminology, with notable implications for computerized medical systems [19]. In the context of mammography, a well established reporting scheme such as the American College of Radiology's BI-RADS (Breast Imaging Reporting and Data System) [20] is a common choice by system developers [4, 13]. However, BI-RADS has not been universally accepted and there is no clear evidence that its use improves consistency and accuracy in mammogram interpretation [21-23]. Additionally, there are indications of its limited expressiveness for encoding mammography findings in computerized data bases [24].

A difficulty in the development of a mammographic knowledge base is the variability amongst mammogram readers in the use of descriptors, which has led some authors to

conclude that the standardization of descriptive terms for mammogram interpretation may be an unattainable goal [25]. However, various studies of radiologists' judgments have positively contributed to the definition of useful sets of mammographic features for inclusion in decision aids [26-27]. These studies typically start with a set of pre-specified ad-hoc mammographic terms that experts are asked to rate. The terms with highest inter-rater agreement are then extracted and incorporated in the decision aid. Again, however, the sets of terms used by these studies are fairly limited.

In the studies described in this paper, we sought to establish a terminology that would capture the reasoning of radiologists in a relatively naturalistic setting. To this end, we started by using protocol analysis techniques, followed by a more focused conceptual approach. The purpose was to elicit, from scratch, a new set of terms from expert radiologists (rather than, e.g., testing a predefined set of descriptive terms) as we deemed existing reporting schemes did not provide the level of detail to explain the basis for the assessment of calcifications.

More specifically, the goals of our two studies were: a) to identify a core set of salient features that are actually used by radiologists when making diagnostic decisions about calcifications (in Study 1); and b) to determine which of these descriptors are useful to discriminate between benign and malignant appearances (in Study 2).

3. Knowledge Elicitation Studies

3.1 Study One

The main goal of Study 1 was to gain an understanding of the terminology used by radiologists when making decisions about calcifications. An exploratory approach was followed at this stage and think-aloud protocols were analysed. Preliminary results of this study were presented in [28].

3.1.1 Method & Materials

Eleven consultant radiologists from 6 different hospitals and screening centres took part. They interpreted mammograms from 20 symptomatic patients: 4 with no reported calcifications or abnormalities, 8 with reported malignant calcifications, and 8 with reported benign calcifications. Four films were used from each patient, namely, the standard mediolateral-oblique and craniocaudal views of the left and right breasts.

The study was conducted individually with each radiologist at her own workplace. The participants were presented with the 20 sets of mammograms, one set at a time, on conventional light boxes. All radiologists saw the same mammograms, though in different sequences. The participants were asked to read each mammogram as they would in a normal clinical situation and to “think aloud”, reporting everything that went through their mind. More specifically, they were instructed to note all the mammographic features they saw on the image (especially calcifications) and to describe them in detail. They were also asked to provide a tentative diagnosis, suggest a course of action and rank their confidence in their decisions. Their verbal reports were recorded on audio-tape.

3.1.2 Results

The resulting audio-recorded think-aloud reports were transcribed. As an illustration of the verbal reports generated by the participating radiologists, we include below an excerpt of one of the transcripts:

"In the left breast superiorly and... laterally, there is an area of increased density with associated barn-door malignant calcification. And there is no question that this [calcification] is malignant. It's segmental, it's going down toward the nipple in a big segment, it's got branching, it's heterogeneous, differing in density, differing in sizes, and is slightly jagged, Chinese letters, looks nasty."

A record was then kept of all the different findings reported by each radiologist (in particular, calcifications). The subsequent analyses focused on the terms used by the radiologists to describe the reported calcifications.

The participants used 159 different terms to describe the calcifications in the study. These terms were grouped in categories corresponding to twelve different dimensions or properties, such as shape, size, density, etc. (see headings in Table 2). Six ‘composite’ descriptors referred to more than one property at the same time. For example, the term “ring” or “ring-like” makes reference to the shape (round), density (lucent centre) and contour (a rim). Other composite terms were: coarse, fine, punctate, popcorn, tubular, and needle-like. These were replaced with the more “nuclear” terms to which they refer. Additionally, synonymies were established. This resulted in a descriptive scheme comprising 50 descriptors grouped along the 12 different properties.

The most commonly used properties and descriptors for each of the main diagnostic categories of calcifications (benign and malignant) are summarized in Table 1. The table includes those properties and values that were noted by six or more participants. A term is

included in each diagnostic category if a radiologist used the term to describe a calcification that she believed belonged to the category¹.

3.2 Study Two

The analyses carried out in Study 1 yielded a potentially useful descriptive scheme. However, the completeness of the scheme was questioned by several of the participants, who noted that some specific appearances (e.g., benign "micro-cystic" calcifications and some indeterminate types) were missing from our set of mammograms. Therefore, we decided to conduct a second more focused study with a larger data set including as many different calcification types and morphologies as possible. The main goals of Study 2 were: a) to validate the set of descriptors obtained from the first study; and b) to obtain more data about their capacity to discriminate between benign and malignant interpretations.

3.2.1 Method & Materials

An experienced radiologist selected mammograms containing 40 cases of calcification. Each case consisted of a pair of mammograms (craniocaudal & lateral-oblique views of the same breast). All calcifications were from cases for which the diagnosis had been confirmed at biopsy or through follow-up. There were 29 cases of confirmed benign calcification and 11 cases of confirmed malignant calcification.

Ten consultant radiologists were shown the 40 cases, one set at a time. On each case, an area of calcification was highlighted on a transparent overlay. With each pair of mammograms a sheet of paper was presented containing the descriptive scheme with the 50 terms obtained from Study 1. They were asked to tick all those descriptors that they felt applied to the highlighted calcification(s). If they thought that the descriptors on the sheet were not sufficient to characterize the calcification(s), they had the option to add any other terms that they may find appropriate. They were also asked to note the level of suspicion attributed to the highlighted calcifications (on a 5 point scale) and to provide, if possible, a tentative diagnosis/etiology for the marked calcifications.

¹ It is worth emphasizing that the "diagnosis" (benign/malignant) associated with each descriptor refers only to how the radiologists categorized the calcification, but not necessarily to their diagnosis of the whole breast; e.g. it is possible for a radiologist to consider that a set of calcifications is benign even if it is associated with other appearances that she considers malignant (see e.g. associated "malignant mass" in the "benign" column in Table 1)

3.2.2 Results

All the descriptors in the original scheme were ticked at least once by at least 50% (5) of the radiologists; 39 of the 50 descriptors were used by at least 80% (8) of the radiologists. The radiologists suggested many other descriptors not included in the original scheme. However, most of these descriptors were used by fewer than three of the participants. Only 5 were suggested by more than 5 of the radiologists, namely: blood vessel (as associated finding), widespread (distribution), intermediate/medium/moderate (cluster size), vessel (location), in mass/density/opacity (location). The five new descriptors were added to the revised scheme and 'irregular' was moved to appear under the heading 'shape' rather than 'contour'. This revised terminology is summarized in Table 2.

The following analyses were conducted to ascertain which descriptors served to discriminate between diagnostic categories. We wanted to assess the discriminatory power of the terms against both the radiologists' assessments of risk and against the follow-up data for these cases. First we measured the correlation between the number of radiologists who applied a descriptor for a given case of calcification and number of radiologists who included the case in each of the risk categories. The sensitivity and specificity of each descriptor were also calculated. Table 3 summarizes the results of these analyses. The terms are listed with the ones with highest specificity appearing first. Only the most discriminating terms are included. For example, the descriptor "big" was used only to describe those calcifications that most radiologists considered benign, never for the malignant ones (hence it has 100% specificity and is highly discriminating), but it only accounts for 28% (8) of the total set of "benign" calcifications (i.e. the sensitivity of the descriptor is low; not all benign calcifications are big); there was also a fairly high (and statistically significant) correlation (0.60) between the number of participants who considered a calcification benign and the number of those who described it as "big".

3.3 Discussion

Based on the analyses of the transcripts of think-out loud reports in Study 1, we derived a descriptive scheme for the characterization of calcifications. This scheme was, to a great extent, backed up by the results of the second study, which led to some minor revisions. An interesting outcome of our work is that the radiologists used a far richer vocabulary for describing calcifications than existing reporting schemes. For example, BI-RADS contains 22 terms to characterize calcifications, whereas the radiologists in our study referred to at least 50 different descriptors. A number of descriptors that we found to be

highly discriminating in our study are absent from the BI-RADS scheme; for example, descriptors relating to "size" and "density", "well" versus "ill defined" contour, and "variable density" versus "homogeneous" particles². See more details in [29].

Furthermore, Study 2 yielded a set of descriptors that are potentially able to discriminate between benign and malignant appearances. Many of the descriptors obtained from the studies were found to have high specificity. Additionally, most of these terms were highly correlated with radiologists' assessment of risk. The sensitivity of most descriptors was low but this is unsurprising since both the benign and malignant categories encompass a wide range of calcification types and morphologies.

Whatmough and colleagues followed a similar approach to ours in a study which looked at radiologists' agreement on the predictive value for malignancy of mammographic features; their results are consistent with ours [26]. Furthermore, our results are consistent to some extent with studies that have looked at the predictive value of mammographic appearances by comparing radiological descriptors with biopsy diagnoses. Table 4 summarizes the results of five such studies featuring those calcification characteristics which were reported to be highly predictive of malignancy or benignity [16, 30-33]. Many features which we found to be of diagnostic value were not used in these studies. It is worth noting that each study tested a different set of radiological features and not all the authors explain the reasoning behind their selection. See more details of our comparisons in [29].

A subset of the most discriminating descriptors obtained in our studies was used subsequently to inform the representation of radiological knowledge in CADMIUM II. This is detailed in the next section.

² Interestingly, there are also descriptors in the BI-RADS scheme that were never mentioned by the radiologists in either study; for example, "spherical", "milk of calcium", "suture", "dystrophic" and "regional" distribution. Arguably, the participating radiologists did not consider these properties to be relevant. But another possible explanation is that the data we used in our study did not contain calcifications with those characteristics. This was a major concern in our data selection. During Study 1, several participants noted that some specific appearances were missing from our set of mammograms. We decided to overcome this difficulty by including a wider range of cases in the second study. We are quite confident that the appearances in Study 2 are fairly extensive (this was reinforced by the participants' comments on the matter). But we are aware that the generalizability of our results is an issue worth exploring and plan further studies with larger number of cases.

4. Knowledge Representation in CADMIUM II

Radiological knowledge in CADMIUM II is represented as a set of arguments, that is, statements that relate characteristics of the calcifications to the risk of malignancy associated with the calcifications. Informally, we could describe the arguments as the bits of information that a radiologist would weigh up when trying to decide whether a calcification is either benign or malignant.

In CADMIUM II, we represent as arguments those descriptors that were found to be most useful to discriminate between benign and malignant appearances in Study 2. We combine both information about the radiologists' "subjective" assessment (as reflected by the correlation values in Table 3) and "objective" information about the specificity of the descriptors, which was based on follow-up data for the cases in the study (see Table 3). Only those descriptors that showed a positive correlation with the radiologists' assessment and a specificity of 50% or above for each diagnostic category (benign and malignant) were included as arguments for either candidate.

It is obvious that not all the descriptors provide the same degree of supporting evidence, so we made a distinction between "strong" arguments and "weak" arguments. *Strong arguments* are those that, if proved true, are considered to be of (almost) sufficient evidence to approve a diagnostic option. For example, if the size of calcifications is "big", it is almost certain that the calcifications are benign; similarly, segmental distribution is a very clear indication of the malignancy of a group of calcifications (see Table 3). These could be considered strong arguments. *Weak arguments*, on the other hand, are those that add support for a diagnosis but are not sufficient individually to approve it. For example, the fact that the calcifications in a region have a linear shape is not enough to decide that they are malignant; however if, in addition, the calcifications have low density and, e.g., are branching and clustered, the probability that they are malignant increases (see again Table 3). These are descriptors that contribute to a malignant diagnosis but are not sufficiently discriminating on their own.

In the current implementation, the strong arguments for either diagnostic category are those descriptors that were found to have a specificity of 80% or above and also showed a statistically significant correlation with the participants' assessments. The weak arguments are those descriptors that either: a) have a specificity of 80% or higher *and* show a positive correlation; or b) show a statistically significant correlation *and* a

specificity of 50% or higher. Table 5 presents the arguments selected following these criteria.

5. Decision Support in CADMIUM II

The arguments just described form the basis for the decision support provided by CADMIUM II. However the effective provision of decision support, requires that the system identifies, automatically, when each of the arguments applies to a set of detected calcifications. Therefore, in addition to the descriptive scheme and the knowledge base described in the previous sections, CADMIUM II incorporates the following elements:

1. an *algorithm* for the detection and segmentation of calcifications; this part of the work is described in a previous publication [34].
2. a set of *image processing measures* to characterize the calcifications detected by the algorithm; specifically, we were interested in identifying image processing measures that could be used to determine when a descriptor used in an argument applied to any set of one or more calcifications.
3. a set of decision rules to establish a mapping between the selected image processing measures and each of the symbolic descriptors used in the arguments; these rules define each descriptor in terms of combinations of imaging measurements.

Figure 1 presents an example of the type of decision support provided by CADMIUM II. The figure is a screendump of the user interface of the system. The user interface displays the digitized mammograms associated with a particular case. The user can display a higher resolution image of a particular segment of a mammogram by manipulating a mouse. In Figure 1, the region contains what is known as “pleomorphic” calcification (often associated with malignancy). If the user requests decision support, a popup window is displayed containing the decision support advice for that region. The advice consists of a display of the regions identified by our algorithm as calcifications together with a set of statements describing the calcifications. In particular, the statements contain the arguments that connect the characteristics of the calcifications to a benign or malignant diagnosis.

In a preliminary evaluation of the CADMIUM II prototype, we compared the performance of our prototype with the R2 ImageChecker 2000 (a commercially available CAD tool) [3]. Such CAD tools are highly sensitive, detecting almost all cancers, but

produce large numbers of false prompts, averaging around 0.5 prompts per film. We were granted access to a set of films being used in an evaluation of R2, but which only included a limited number of calcifications. The ImageChecker produced 27 prompts for calcification on these films. We looked at the arguments generated by CADMIUM II for these regions. Our aim was to see if adding the arguments generated by CADMIUM to the prompts provided by R2 could help radiologists identify false positive prompts. For six of the prompts, CADMIUM II generated equal numbers of benign and malignant arguments (i.e., the diagnosis was indeterminate). The data for the remaining 21 regions are presented in Table 6. Fifteen of these prompted regions were "false positives" (i.e., contained benign calcifications that R2 marked incorrectly); for seven of these, CADMIUM II provided a correct benign diagnosis. In one prompted region containing clear benign and subtle malignant calcification, CADMIUM II failed to diagnose correctly the malignant calcification leading to the single "false negative" in the table. In addition to the cases in Table 5, the data set contained three cases of malignant calcification that R2 failed to prompt; when these cases were processed by CADMIUM II, our system correctly identified for them a malignant cause. This (limited) evaluation suggests a role for CADMIUM II as an adjunct to a prompting system.

A further evaluation is still in progress. We hope to interview eight radiologists who will all have an opportunity to familiarize themselves with the prototype before answering questions about the approach and the advice given.

6. Conclusions

Two empirical studies looking at the descriptive terms that expert radiologists use when making decisions about calcifications have yielded a set of salient features with a potential value for discriminating between malignant and benign mammographic appearances. These features have been used to inform the argumentation used in the decision support component of CADMIUM II, a computer aided diagnosis tool for mammography that combines symbolic reasoning with image processing.

An issue currently being explored is the generalizability of our results as both the data and subject samples in the studies were fairly limited. As noted, we are quite confident that the cases in the second study contained an extensive range of calcification appearances (this was reinforced by the participants' comments). However we feel additional studies with larger numbers of cases and radiologists are needed to investigate

further the validity of the implemented descriptive scheme and arguments. Additionally, we are considering the potential for using more sophisticated knowledge acquisition techniques, such as automatic rule induction (i.e., machine learning), to derive the symbolic explanations of diagnoses from the image processing data.

Another issue we are investigating is the representation of the uncertainty associated with the arguments. As noted, not all the descriptors represented as arguments are unequivocal indicators of risk of malignancy. At the moment CADMIUM II implements simply a binary distinction between “weak” and “strong” arguments. We are exploring the possibility of associating the representation of the arguments with numerical weights that reflect their comparative strength or predictive value. The idea is to provide advice that indicates, for example, the probability that a calcification has of being benign or malignant if it possesses a particular characteristic³.

The results of the preliminary evaluation of our prototype seem to back up the notion that computer support in mammography would benefit from an explicit representation of the decision making process. The results show the potential of our tool to improve on the performance of the current market leader in this field. Another potential application of CADMIUM II, which we are currently exploring, is its use in supporting the training of less experienced mammogram readers.

Although it was not the purpose of our studies, we believe that the set of descriptors we have developed has potential to be used as a reporting scheme in mammography, as it is arguably more complete and consistent than existing terminologies. This view is supported by feedback we have received from domain experts. Our descriptive scheme is limited to mammography and to a particular subset of appearances (calcifications) and we do not think it can be generalized to other domains. However, we believe that the methodology we have used (possibly in combination with automatic methods of rule induction) is a fruitful approach for the development of computer advice in a variety of medical applications. An advantage of this approach is that the advice provided is presented at a level of description that is both relevant and informative for the user.

³ We must note that our findings are unlikely to be independent predictors and it would be therefore inappropriate to apply our data within a Bayesian network, for example. It might nevertheless be interesting to ascertain if the non-independence was sufficient to lead a Bayesian system into correct predictions.

Acknowledgments

This work is funded by the Engineering and Physical Sciences Research Council through the Health Informatics programme and by the Imperial Cancer Research Fund. The co-operation of Drs R. Given-Wilson, J. Davies, A. Schneider, J. Cooke, S. Rankin, I. Nockler and their colleagues is gratefully acknowledged.

References

1. Taylor P. Computer aids for decision making in diagnostic radiology – a literature review. *Brit J Radiol* 1995; 68: 945-957.
2. Roehrig J, Castellino RA. The promise of computer aided detection in digital mammography. *Eur J Radiol* 1999; 31 (1): 35-39.
3. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001 220:781-6.
4. Schmidt F, Sorantin E, Szepesvári C, Graif E, Becker M, Mayer H, Hartwagner K. An automatic method for the identification and interpretation of clustered microcalcifications in mammograms. *Phys Med Biol* 1999; 44 (5): 1231-43.
5. Hara T., Yamada A., Fujita H. et al. Automated classification method of mammographic calcifications by using artificial neural network and ACR BIRADS criteria for microcalcification distribution. In: Yaffe M, ed.: *Proceedings of the 5th International Workshop on Digital Mammography*; Medical Physics Publishing, 2000: 198-204.
6. Kaufmann G.H., Salifty M.F., Granitto P. and Ceccatto H.A. Automated detection and classification of clustered calcifications using morphological filtering and statistical techniques. In: Yaffe M, ed.: *Proceedings of the 5th International Workshop on Digital Mammography*; Medical Physics Publishing, 2000: 253-8.
7. Markopoulos C, Kouskos E, Koufopoulos K et al. Use of artificial neural networks (computer analysis) in the diagnosis of microcalcifications on mammography. *Eur J Radiol* 2001; 39:60-5
8. Hartswood M, Procter R, Willimas J. Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening

- mammography? In: Karssemeijer E, Thijsen M, Hendriks J, eds.: *Digital Mammography*; Kluwer Publishing, Dordrecht, 1998: 363-370.
9. Taylor P, Alberdi E, Lee R et al. Incorporating Image Processing in a Clinical Decision Support System. In: Insana MF, Leahy RM, eds.: *Information Processing in Medical Imaging*; Springer Verlag, 2001: 134-140.
 10. Taylor P, Fox J, Todd-Pokropek A. The development and evaluation of CADMIUM: a prototype system to assist in the interpretation of mammograms. *Medical Image Analysis* 1999; 3: 321-337.
 11. Cook HM and Fox MC. Application of expert systems to mammographic image analysis. *Am J Physiol Imag* 1989; 4: 16-22.
 12. Stewart BD, Cairns AY, Ricketts IW. The analysis of breast abnormalities by computer. In: Gale A, Astley S, Dance D, Cairns A, eds.: *Digital Mammography*; Elsevier Science, Amsterdam, 1994: 251-261.
 13. Floyd CE, Lo JY, Tourassi GD. Case-based reasoning algorithm that uses mammographic findings for breast biopsy decisions. *AJR* 2000; 175: 1347-52.
 14. Fox J, Myers CD, Greaves MF, Pegram S. Knowledge acquisition for expert systems: Experience in leukaemia diagnosis. *Methods of Information in Medicine* 1985; 24: 65-72.
 15. Cooke NJ. Varieties of knowledge elicitation techniques. *International Journal of Human Computer Studies* 1994; 41: 801-849.
 16. Khan AS, Hoffmann A. building a case based diet recommendation system without a knowledge engineer. *Artificial Intelligence in Medicine* 2003; 27(2): 155-179.
 17. Guimaraes G, Peter JH, Penzel T, Ultsch A. A method for automated temporal knowledge acquisition applied to sleep-related breathing disorders. *Artificial Intelligence in Medicine* 2001; 23(3): 211-237.
 18. Ceusters W, Rogers J, Consorti F, Rossi-Mori A. Syntactic-semantic tagging as a mediator between linguistic representations and formal models: an exercise in linking SNOMED to GALEN. *Intelligence in Medicine* 1999; 15(1): 5-23.
 19. Chute CG. Clinical classification and terminology: Some history and current observations. *JAMIA* 2000; 7 (3): 298-303.
 20. American College of Radiology. *Breast imaging reporting and data system (BI-RADS)*. 2nd ed. Reston, VA: ACR, 1995.

21. Baker JA, Kornguth PJ, Floyd CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *AJR* 1996; 166: 773-8.
22. Liberman L, Abramson AF, Lassman JR, Morris EA, Dershaw DD. The breast imaging reporting and data system: Positive predictive value of mammographic features and final assessment categories. *AJR* 1997; 171: 35-40.
23. Kerlikowske K, Grady D, Barclay J, Frankel SD, S. H. Ominsky SH, Sickles EA, and Ernster V. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting Data System. *Journal of the National Cancer Institute* 1998; 90 (23): 1801-1809.
24. Starren J and Johnson SM. Expressiveness of the Breast Imaging Reporting and Database System (BI-RADS). *JAMIA* 1997; Suppl. S: 655-659.
25. Simpson W, Neilson F, Kelly PJ. Descriptive Terms for mammographic abnormalities: Observer variation in application. *Clin Radiol* 1996; 51: 709-71.
26. Whatmough P, Gale A, Wilson ARM. Do radiologists agree on the importance of mammographic features? In: Doi K, Giger ML, Nishikawa RM, Schmidt RA, eds.: *Digital Mammography'96*. Elsevier Science, Amsterdam, 1996; 111-116.
27. D'Orsi CJ, Getty DJ, Swets JA, et al. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. *Radiology* 1992; 184: 619-622.
28. Alberdi E, Taylor P, Lee R et al. CADMIUM II: Acquisition and representation of radiological knowledge for computerized decision support in mammography. In: Overhage JM ed.: *Proceedings of the American Medical Informatics Association Symposium, 2000*; 7-11.
29. Alberdi E, Taylor P, Lee R et al. Eliciting a terminology for mammographic calcifications. *Clinical Radiology*, 2002; 57: 1007-1013.
30. Skinner MA, Swain M, Simmons R, McCarty KS, Sullivan DC, Iglehart JD. Nonpalpable breast lesions at biopsy. *Ann Surg*, 1988; 208(2): 203-208.
31. Franceschi D, Crowe J, Zollinger R, Duchesneau R, Shenk R, Stefanek G. Biopsy of the breast for mamographically detected lesions. *Surgery, Gynecology and Obstetrics*, 1990; 171(6): 449-455.
32. Monostori Z, Herman PG, and Carmody DP, Eacobacci TM, Capece NR, Cruz VM, Fentin S, Vernace FM. Limitations in distinguishing malignant from benign lesions

of the breast by systematic review of mamograms. *Surgery, Gynecology and Obstetrics*, 1991; 173: 438-442.

33. Harkins K, Tartter PI, Hermann G, Squitieri R, Bower ST, Keller RJ. Multivariate analysis of roentgenologic characteristics and risks factors for nonpalpable carcinoma of the breast. *Journal of the American College of Surgeons*, 1994; 178: 149-154.
34. Lee R, Alberdi E, Taylor P et al. A Comparative study of four techniques for calcification detection. In: Yaffe M, ed.: *Proceedings of the 5th International Workshop on Digital Mammography*; Medical Physics Publishing, 2000: 264-271.

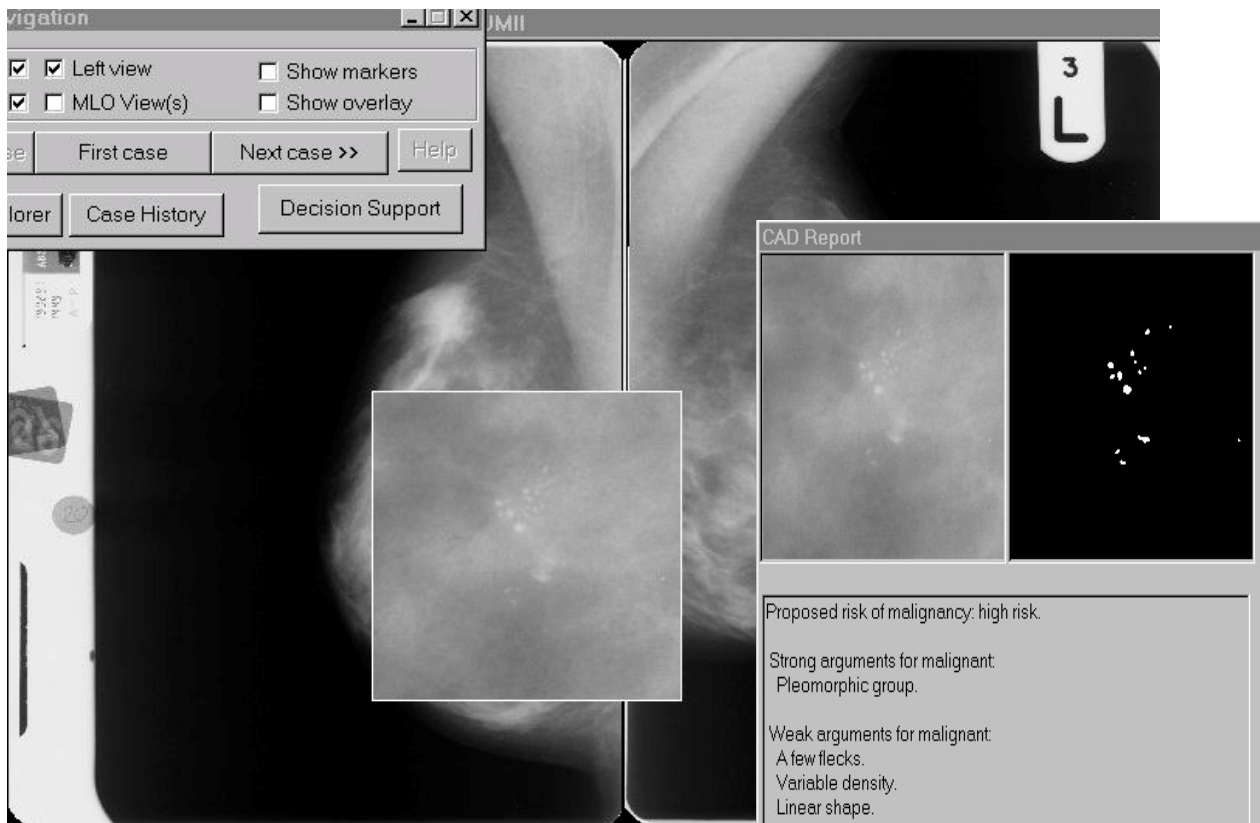


Figure 1. User interface in CADMIUM II. It displays: a) in the background, digitized mammograms associated with a particular case; b) in the top left, a popup window that allows the user to select a case and various bits of information associated with it, including the decision support generated by the tool; c) towards the center, a higher resolution image of a particular mammogram segment selected by the user (i.e., the user's current region of interest); d), bottom right, a popup window containing the decision support advice for that region, namely, a display of the regions identified by our algorithm as calcifications together with a set of statements describing the calcifications and risk of malignancy.

Table 1. Most frequently used properties and values to describe calcifications in Study 1.

Benign			Malignant		
Properties and Descriptors	R's	% (N=527)	Properties and Descriptors	R's	% (N=242)
Size	11	26% (139)	Size	11	21.5% (52)
large	11	14% (72)	small	11	21% (51)
small	11	13% (67)	Shape	10	20.5% (50)
Density	11	23% (120)	branching	8	11% (28)
high density	9	11% (59)	linear	6	7% (17)
low density	10	6% (31)	Variability	10	20% (49)
lucent centre	11	5% (26)	pleomorphic	10	12% (30)
Distribution	11	20% (108)	Associated findings	10	14% (35)
scattered	11	9% (49)	malignant mass	10	9% (22)
vascular	11	6.5%(35)	Density	8	10% (24)
Shape	11	19% (100)	low density	8	9.5% (23)
round	11	15% (77)	Distribution	9	10% (24)
linear	8	3% (14)	Contour	6	4.5% (11)
Number of Flecks	11	10% (52)	irregular	6	4% (10)
1-5 specks	11	8% (44)			
Contour	8	8% (41)			
well defined	6	5% (28)			
Associated findings	11	7% (36)			
opacity	10	4% (22)			
malignant mass	6	2% (10)			

NOTE: Properties are shown in bold type and descriptors are shown in regular type.

R's indicates the maximum number of radiologists who referred, at least once, to each property or property value.

N in each diagnostic category (i.e., benign & malignant) indicates the total number of instances in which any term was used to describe a calcification that a radiologist considered belonged to that category.

Table 2. The revised descriptive scheme following the evaluation in Study 2.

SHAPE	NUMBER OF FLECKS	CLUSTER/AREA SIZE
branching	1 fleck	very small
linear	few flecks (≤ 5)	small
streaming	several specks ($>5, <X$)	medium
oval	multiple specks ($>X$)	big
round		N/A (single)
curvilinear	ASSOCIATED FINDINGS	BETWEEN-FLECK VARIABILITY
irregular	benign opacity/mass/density	homogeneous
	malignant opacity/mass/density	pleomorphic
SIZE	distortion	variable size
big	blood vessel	variable density
medium	no associated finding	variable contour
small		similar density
	DISTRIBUTION	N/A (single)
DENSITY	isolated	LOCATION
high density	scattered	in axilla
low density	widespread	in skin
lucent centre	adjacent to each other	within fat
	clustered	within glandular stroma
CONTOUR/EDGE	segmental	opacity/mass/density
well defined	ductal/linear	vessel/artery
ill-defined	vascular	
with a rim	ORIENTATION	VIEW DIFFERENCES
	towards nipple	cc:
		l-o:

Table 3. Most discriminating descriptors for cases in Study 2.

Benign					Malignant				
N	Descriptor	Correl	Sens	Spec	N	Descriptor	Correl	Sens	Spec
8	big	0.60*	28%	100%	3	segmental	0.68*	27%	100%
8	within fat	0.59*	28%	100%	1	malig. finding	0.52*	9%	100%
4	similar density	0.49*	14%	100%	1	distortion	0.36	9%	100%
3	curvilinear	0.44*	10%	100%	1	few specks	0.14	9%	100%
1	with a rim	0.42*	3%	100%	6	pleomorphic	0.87*	55%	86%
4	isolated	0.39	14%	100%	6	variable density	0.77*	55%	75%
4	lucent centre	0.39	14%	100%	2	towards nipple	0.46*	18%	66%
12	1-5 flecks	0.38	41%	100%	5	ill-defined	0.55*	45%	63%
6	scattered	0.37	21%	100%	9	variable size	0.68*	82%	60%
2	vascular	0.27	7%	100%	3	several specks	0.28	27%	60%
1	in skin	0.21	3%	100%	4	linear	0.58*	36%	57%
3	adjacent	0.08	10%	100%	2	branching	0.73*	18%	50%
4	oval	0.07	14%	100%	1	ductal/linear	0.54*	9%	50%
22	well defined	0.73*	76%	92%	6	clustered	0.53*	55%	50%
7	homogeneous	0.65*	24%	88%	5	low density	0.37	45%	50%
20	no finding	0.24	69%	80%	7	small	0.54*	64%	41%
19	high density	0.26	66%	76%	11	stroma	0.53*	100%	41%
13	round	0.24	45%	76%	6	multiple specks	0.27	55%	40%
7	big (cluster)	0.18	24%	70%	1	variable contour	0.59*	9%	33%
3	medium	0.13	10%	60%	6	small (cluster)	0.18	55%	29%

Note. N refers to the number of calcifications (or clusters) that have been described by a majority of readers with each descriptor in each category.

Table 4. Studies of the correspondence between mammographic appearances and diagnosis

	Lieberman et al. [22]	Skinner et al. [30]	Franceschi et al. [31]	Monostori et al. [32]	Harkins et al. [33]
MALIGNANT					
branching	✓	✓	✓	✓	
pleomorphic					✓
number		✓	✓	✓	
segmental	✓				
small			✓		✓
scattered					✓
BENIGN					
round				✓	✓
coarse					✓
solid				✓	
packed					✓
curvilinear					
irregular				✓	

Table 5. List of arguments for the diagnostic categories “benign” and “malignant” as implemented in CADMIUM II.

	Benign		Malignant
Strong	big within fat similar density curvilinear with a rim well defined homogeneous	Strong	segmental pleomorphic
Weak	isolated lucent centre 1-5 flecks scattered vascular in skin adjacent oval no assoc. finding	Weak	few specks variable density towards nipple ill-defined variable size linear branching clustered

Table 6: The balance of arguments generated by CADMIUM II compared to the actual diagnosis, on 21 image regions identified by the R2 ImageChecker 2000

CADMIUM II	Actual diagnosis	
	Malignant	Benign
Malignant	5	8
Benign	1	7