Analysis of probed regions in an interactive CAD system for the detection of masses in mammograms

M. Samulski[a], A. Hupse[a], C. Boetes[a], G. den Heeten[b], and N. Karssemeijer[a]

[a]Department of Radiology
[b]National Expert and Training Centre for Breast Cancer Screening
Radboud University Nijmegen Medical Centre
Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands;

ABSTRACT

Most computer aided detection (CAD) systems for mammographic mass detection display all suspicious regions identified by computer algorithms and are mainly intended to avoid missing cancers due to perceptual oversights. Considering that interpretation failure is recognized to be a more common cause of missing cancers in screening than perceptual oversights, a dedicated mammographic CAD system has been developed that can be queried interactively for the presence of CAD prompts using a mouse click. To assess the potential benefit of using CAD in an interactive way, an observer study was conducted in which 4 radiologists and 6 non-radiologists evaluated 60 cases with and without CAD, to compare the detection performance of the unaided reader with that of the reader with CAD assistance. 20 cases had a malignant mass, and 40 were cancer-free. During the reading sessions we recorded time and probed locations which reveal information about the search strategy and detection process. The purpose of this study is to determine a relation between detection performance and time to first probe of the lesion and to investigate if longer reading times lead to more reports of malignant lesions in lesion-free areas. On average, 65.0% of the malignant lesions were found within 60 seconds and this percentage stabilizes after this period. Results suggest that longer reading time did not lead to more false positives. 74.6% of the reported true positive findings were hit by the first probe, and 93.2% were hit within 5 probes, which may suggest that many of the correctly reported malignant masses were perceived immediately after image onset.

Keywords: Observer Performance Evaluation, Image Perception, Cognitive Errors, Interactive CAD, Breast Screening

1. INTRODUCTION

In the last decade, computer-aided detection (CAD) has been the subject of much research and development, and has been widely adopted to improve reader performance. However, the effectiveness of CAD systems has recently triggered debate concerning the added value and still remains controversial. Whereas some observational studies indicate that CAD does improve the accuracy of screening mammography,[1–3] others have found that the use of computer-aided detection did not increase the cancer-detection rate significantly, and increased the number of false positive mammograms.[4,5] Therefore, further research is needed to decrease false positive marks and increase specificity while maintaining sensitivity. Computer aided detection (CAD) systems for the detection of masses in mammograms use algorithms to identify regions in a mammogram that have suspicious features. As currently used, radiologists first inspect a case, then activate the CAD display and re-evaluate the suspicious patterns marked by CAD before making their final decision. These CAD systems are mainly intended to avoid missing cancers due to perceptual oversights.

Considering that there is reason to believe that interpretation failure is a more common cause of missing cancers in screening than perceptual oversights,[6,7] methods have to be investigated to help radiologists with

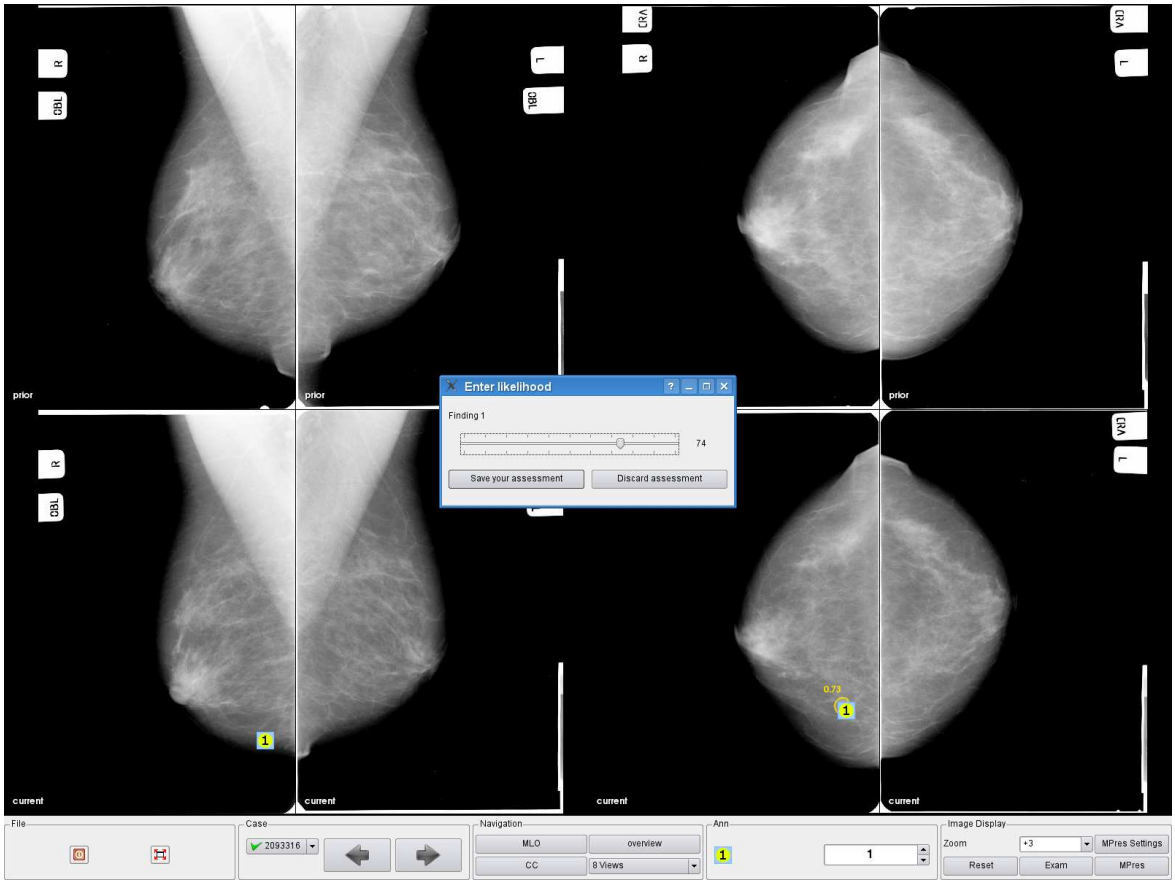


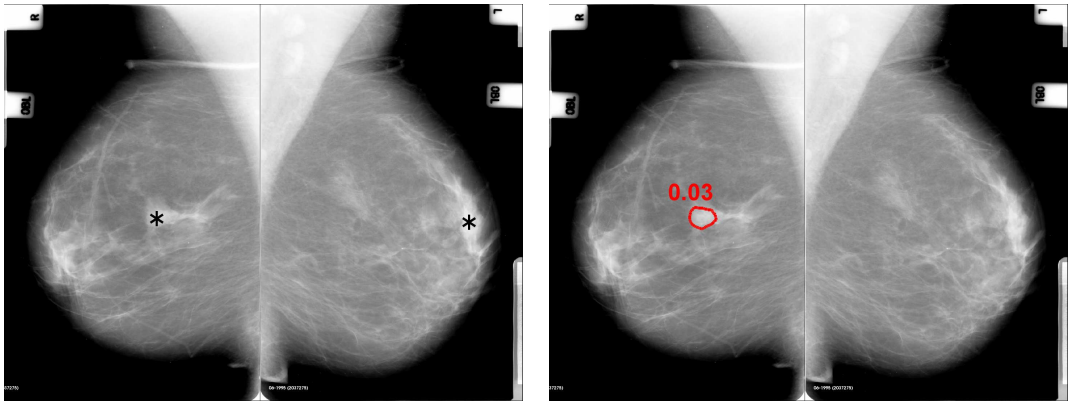
Figure 1: The graphical user interface of the developed CAD workstation that was used in the observer experiments. One finding has been annotated in both projections, and the reader is asked to assign a malignancy score between 0 and 100 to that finding. The reader has probed the region in the CC view, which had a normality score of 0.73.

decision making. To aid readers with interpretation, a dedicated mammographic CAD system is developed that can be queried interactively for the presence of CAD prompts using a mouse click. To assess the potential benefit of using CAD in an interactive way, an observer study was carried out in which readers evaluated a set of cases to compare the detection performance of the unaided reader with that of the reader with CAD assistance. During the reading sessions we recorded time and probed locations which reveal information about the search strategy and detection process. The purpose of this study is to determine a relation between detection performance and time to first probe of the lesion and to investigate if longer reading times lead to more reports of malignant lesions in lesion-free areas.

2. MATERIALS AND METHODS

2.1 Workstation

A mammographic workstation has been designed and built to investigate if detection performance of the readers can be improved by showing the probabilities of the CAD system at areas that raised the readers attention. It is shown in some studies, that giving readers additional information on the likelihood of CAD marks might be helpful in decision-making.[8,9] Our workstation can be queried interactively for the presence of CAD results by clicking on suspect regions in the mammogram (see Figure 2). This way of using CAD is conceptually distinct from regular CAD workstations that display all CAD findings exceeding a certain threshold level as prompts without displaying the probability of malignancy. For each queried location, the workstation checks if a CAD



(a) (b)

Figure 2: (a) Current practice where CAD findings are presented as prompts (b) Interactive CAD system where a region is probed. The contour of the region is colored according to its malignancy rating, and the corresponding normality score is displayed next to it.

region is available at that location. If a CAD region is available, it is presented to the reader with the computer-estimated malignancy score, provided that the CAD region exceeds a predefined malignancy threshold. The contour of the displayed region is colored based on its assigned malignancy score using a continuous color scale from red to green, for respectively high to low malignancy ratings.

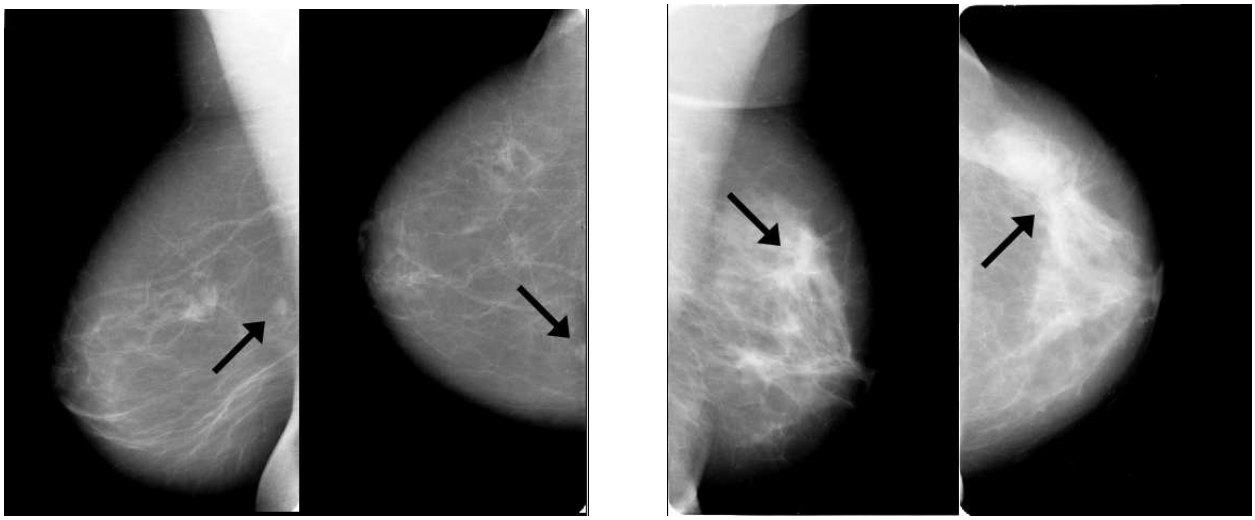
The workstation has all the basic functionality such as dedicated hanging protocols, zooming, window/level adjustments and local contrast enhancement tools. When a new mammographic case is loaded, the workstation shows the current medio-lateral oblique and cranio-caudal mammograms at the bottom, and the corresponding prior mammograms at the top of the screen, as shown in Figure 1. Using the toggle button, the prior and current mammogram of a selected projection alternates at the same display with a substantial higher resolution, which leads to improved perception of lesion growth.

For observer study purposes, the workstation is equipped with functionality to mark and score localized findings. To mark a finding, the reader drags an annotation icon from the toolbar by clicking the icon, holding the mouse button, move it to the suspicious area, and release the mouse button to drop it. This annotation icon is a yellow circle with a number inside that indicates the finding number. Annotations in different projections that are expected to be the same lesion must have the same finding number. By increasing this number, the reader can report multiple findings. When the reader has analyzed the case completely and goes to the next case, a dialog with sliders appear to let the reader assign a malignancy score to each reported finding on a continuous scale ranging from 0 to 100.

A 30 inch color LCD panel (Eizo FlexScan SX3031W) with a native resolution of 2560 x 1600 was connected to the workstation. The CAD results were obtained from the ImageChecker v8.0 by R2 Technology Inc. (A Hologic Company, Santa Clara, Calif., USA) and were available in both current and prior mammograms.

2.2 Observer study

Ten readers, of which four were certified screening radiologists, and six were non-radiologists with mammogram reading skills, participated in the study. We selected a case set of 60 anonymized screening mammograms. From these, 20 had a biopsy proven malignant mass, and 40 were cancer-free and had been stable for 2 years. All cancer cases selected were subtle cancers that were missed at the original screening and were retrospectively identified as visible. Two cancer cases are shown in Figure 3. Cases with only microcalcifications were excluded. All cases had prior mammograms available, but not always a caudal-cranial view. The digitized mammograms were obtained from the Dutch Breast Cancer Screening Program and were scanned using a Lumisys 85 digitizer at a pixel resolution of 50 μm. The mammograms were averaged down to a resolution of 100 μm, maintaining a gray level resolution of 12 bits.



(a) (b)

Figure 3: The mammograms above show two cases that were in the observer experiment test set displaying a mass lesion and architectural distortion that were missed in original screening but were judged visible in retrospect. In (a) there is an invasive ductal carcinoma close to the pectoral muscle. In the medio-lateral oblique CAD has detected the region with a normality score of 0.02, in the cranio-caudal view the region was not detected. In (b) an invasive lobular carcinoma is shown with architectural distortion as the predominant appearance. In the medio-lateral oblique CAD has detected the region with a normality score of 0.67, and in the cranio-caudal view the region was detected with a normality score of 0.27.

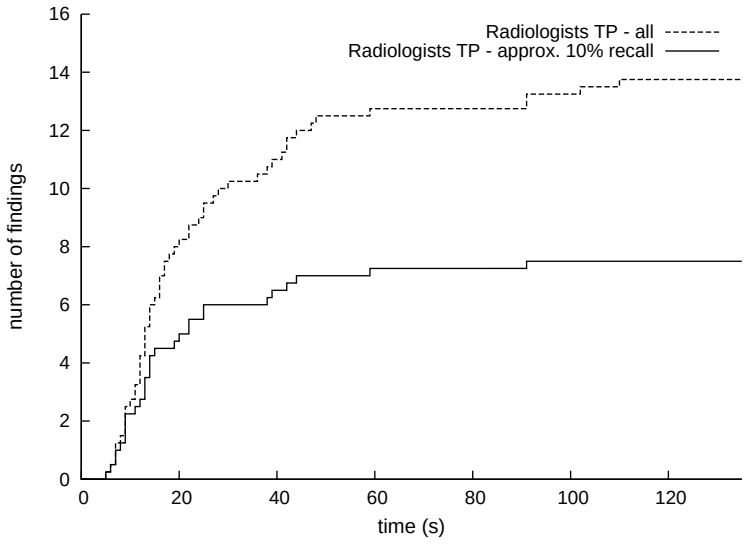
2.3 Experiment protocol

Before the observer experiment, 60 training cases (40 cases with normal findings, 20 mass cases) were presented to the non-radiologists. The radiologists were offered less training cases due to time constraints. The training cases served to acquaint the readers with the workstation, including the use of interactive CAD, and the reporting facility. After the training phase, the observers read the case set in two sessions of 60 cases each. In the first session, 30 mammograms were read with CAD and 30 without. In the second session, CAD was made available for the cases initially read without CAD and vice versa. Each session had a balanced mix of normal and abnormal cases. The order of the cases within the with-CAD and without-CAD session was randomized to minimize reading order effects. The observers were instructed to search for malignant masses only, and were told not to report any microcalcifications. They were informed that the case set contained normal and abnormal cases, and they were also told the approximate proportion of the abnormal cases. In the without-CAD session, the readers searched the mammograms for suspicious cancer regions in a manner similar to their screening practice. They were asked to mark the finding in the MLO and CC view, and assign a malignancy score on a continuous scale ranging from 0 to 100. Readers were also asked to mark at least 1 finding per case, unless a case was so obviously normal that no reasonable finding could be marked. There was no limit on the reading time. In the with-CAD session, the readers could query the CAD system by clicking on regions in the mammogram that raised their attention.

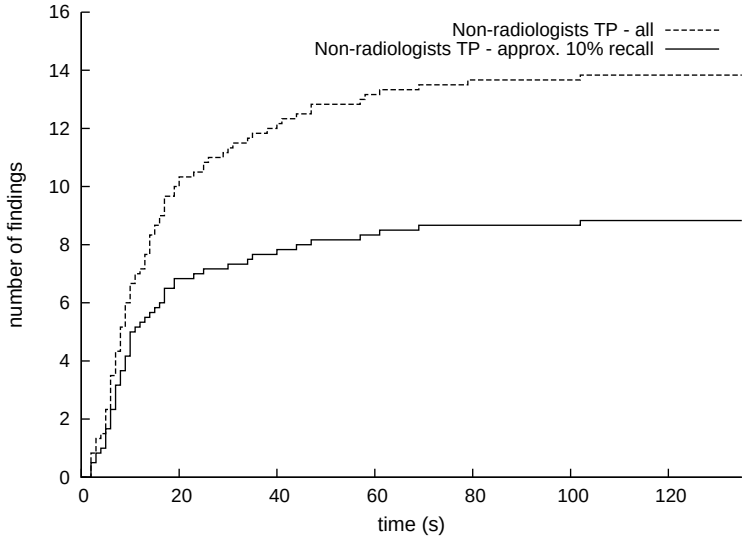
The average number of CAD regions made available is adjustable, by selecting only CAD regions with malignancy scores exceeding some threshold. In a preliminary pilot study, we had set the threshold as such that in normal cases the average number of false positive regions was five. During evaluation of the results, the number of clickable regions were considered too high. Therefore we decided to lower the threshold in such a way that in normal cases the average number of false positive regions was two.

2.4 Analysis of observers responses

A finding was considered a true-positive (TP), if the distance between the observer's marked location and the true cancer location was less than 2 cm in at least one of the views. The true cancer location is defined as the mathematical center of mass of the lesion. Marks in normal images were counted as a false-positive (FP).



(a)



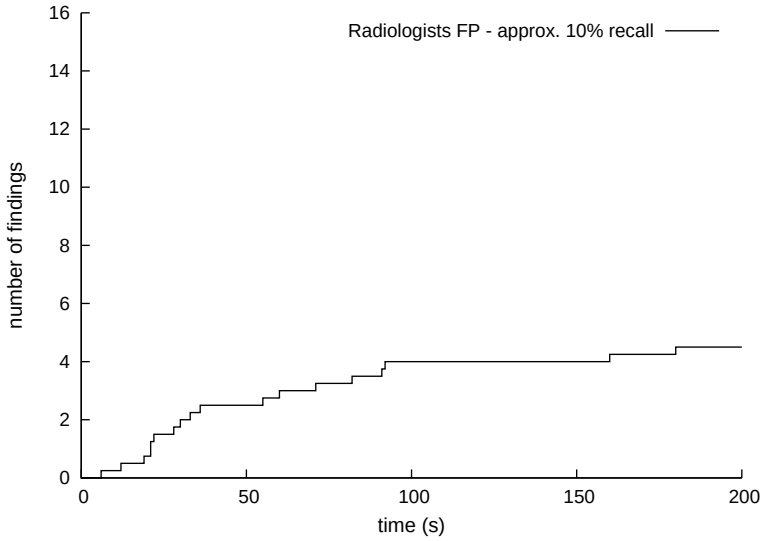
(b)

Figure 4: Graphs show the mean cumulative number of correctly reported findings (true positives) as function of time to first probe of the lesion, for the radiologists (a) and non-radiologists (b).

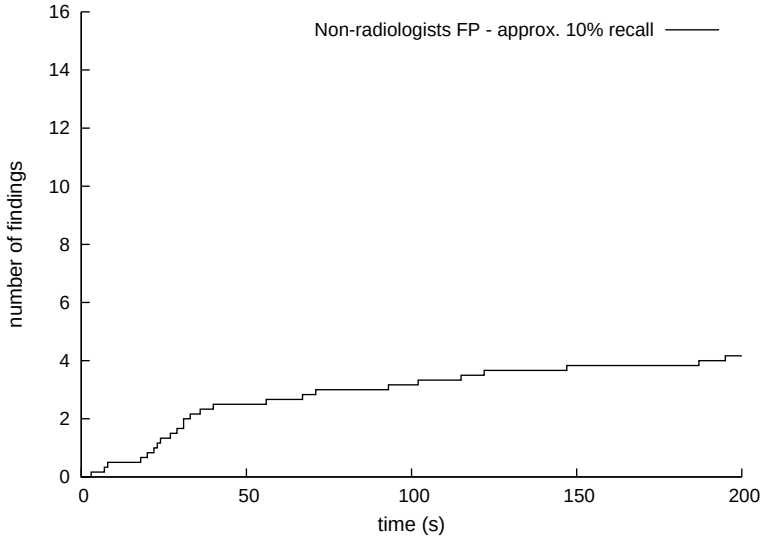
During the reading sessions the time and location of each probe for CAD information was logged. For each reported finding, we check if it was probed or not. If it was probed, we obtained the first time the finding was probed. The finding was considered to be probed if the annotation location and probe location were less than 2 cm apart. In some cases, a finding was reported for which no corresponding probe location could be found. In these cases we took the time-point when the finding was annotated.

3. RESULTS

The mean cumulative number of correctly reported findings (true positives) as function of time to first probe of the lesion is shown in Figure 4, for the radiologists and non-radiologists. A rapid increase is seen in the true positive findings for up to a first probing time of 30 seconds, followed by a much more gradual increase. On average 65.0% of the malignant lesions were found within 60 seconds when counting all correctly annotated



(a)



(b)

Figure 5: Graphs show the mean cumulative number of incorrectly reported findings on normal cases (false positives) as function of time to first probe of the reported finding, for the radiologists (a) and non-radiologists (b), at a false positive recall rate of approximately 10%.

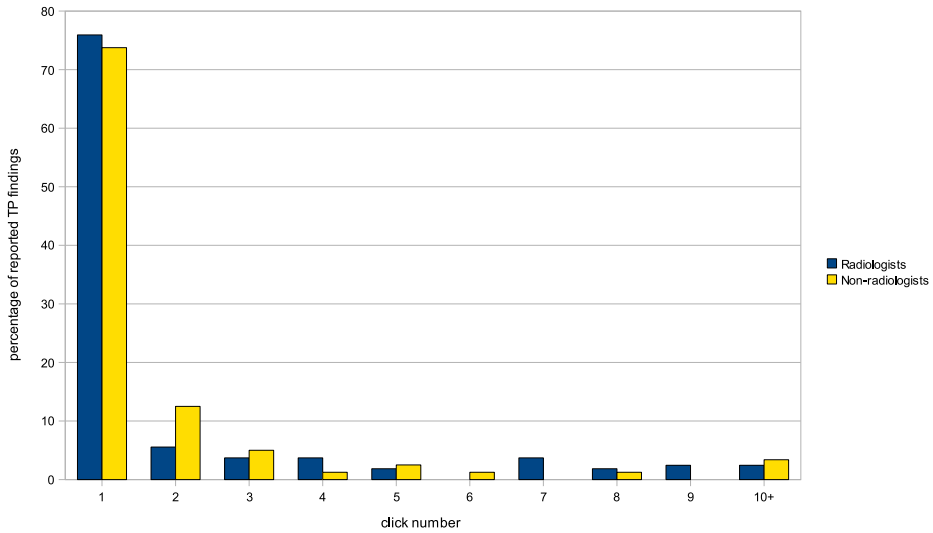


Figure 6: In this histogram the click number indicates how many locations the observer has probed to reach the location of the true positive decision. Approximately 75% of the reported TP findings were hit by the first probe, for both the radiologists and non-radiologists.

findings irrespective of the score the observer has given to the finding. After 60 seconds, the percentage of malignant lesions that were found tend to stabilize.

For every reader a localization ROC (LROC) was generated, that shows the fraction of correctly localized cancers as a function of the percentage of recalled normal cases. For every reader, we determined the cutoff point such that the false positive recall rate is 10%. This interval was chosen because in a screening setting radiologists usually have recall rates lower than 10 percent. When using that threshold on the observer score such that the false positive recall rate is 10%, the curve (shown as a solid line in Figure 4) follows the same trend as the curve of all true positives. For some readers the selected threshold resulted in a slightly higher recall rate than 10%, if they assigned the same malignancy score to multiple findings.

In Figure 5 the mean cumulative number of incorrectly reported findings in normal cases (false positives) is plotted as function of time to first probe of the reported finding, for the radiologists and non-radiologists at a false positive recall rate of approximately 10%. Results show that the number of false positive findings increased rapidly in the first 40 seconds. After this period the number of false positive findings increased slowly for both the radiologists and non-radiologists. Since the readers were instructed to mark at least 1 finding per case, we do not show all reported false positive findings in this graph, because they would include findings that were given a low malignancy score.

The histogram shown in Figure 6 indicates how many locations the observer has probed before the location of the reported true positive finding was reached. 74.6% of the reported true positive findings were already hit by the first probe, and 93.2% were hit within 5 probes. Readers probed locations for CAD information 12.1 times per case on average, 9.0 for the radiologists and 14.1 for the non-radiologists. In some cases, a finding was annotated for which no corresponding probe location could be found. In these cases we took the time-point when the finding was annotated. This occurred, on average, in 0.8 cases per reader. On average, readers probed 80.5% of the malignant lesions, and reported 71.0% of them. At a false positive recall rate of 10% only 42.5% were detected.

4. DISCUSSION AND CONCLUSION

Analyzing a mammogram involves multiple steps. Immediately after image onset, the observer gets a global impression of the image. A number of discordant regions that pop out are then individually examined using the

high resolution beam of the visual system, the fovea.[10,11] The results in Figure 6 supports the observation that the true malignant masses the observers correctly report may be perceived immediately after image onset, if we assume that the click number corresponds to the order in which discordant regions are examined in detail.

Averaged over all readers, 65.0% of the malignant lesions were found within 60 seconds and stabilizes somewhat after this period. In an eye-tracker study, Nodine et al.,[12] investigated if longer decision time negatively effects the performance in interpreting mammograms. An computer display system with an eye-head tracker mounted was used to measure the timing of decisions, where visual attention was directed, and how much time was spent fixating on a region of interest for each decision. It was found that for mammographers the false-positive responses for both abnormal and normal cases continued to increase throughout the time of viewing, but that true-positive responses continued to outweigh the false-positive responses. This pattern was different for the trainees, where the false positive responses for normal cases plus the false positive responses for the abnormal cases started to overtake the true positive responses. The performance of the trainees was approximately half that of mammographers. In our study, results show that the performance of the non-radiologists was similar to that of the radiologists when using the interactive CAD system. At a false positive recall rate of approximately 10% the performance of the non-radiologists was higher using the interactive CAD system. Possible causes are that the non-radiologists had more training with the CAD system and that the radiologists were not familiar with the computer-estimated normality scores and the numerical scale.The results also suggest that longer reading time did not lead to much more reported false positives in normal images, for both the radiologists and non-radiologists.

One of the underlying assumptions in many eye-tracker studies is that if an object receives a predetermined amount of visual attention, it is also processed cognitively. Also, an accurate and reliable calibration is required to know precisely what a subject is looking at, and certain thresholds (e.g., fixation times) need to be defined. One of the advantages of eye-trackers is that they also record unconscious fixations that play a role in the final decision.

In our model, the actual interpretation process starts when the reader actively probes a region that raised his attention during the search process, and not when a region receives a particular amount of visual attention. We assume that if the reader probes another region, he finalized his decision about the one he probed before. Interesting future research could include mounting an eye-head tracker on our interactive CAD workstation to extend the eye position and fixation data with the recorded times of active interpretation of regions of interests, to get a more accurate measurement of decision times.

In few cases the location of the reported finding could not be matched with a location that was probed, in which we took the time-point when the mark was placed on the mammogram. In future experiments, we will instruct readers to explicitly probe the suspicious locations they will report, even if the suspicious region contains an obvious malignant-looking tumor. The overall performance of the readers might seem low, but it should be noted that the readers read screening mammograms with subtle cancers that were missed at screening.

A limitation of our study is that the reading conditions were less optimal than in screening practice, because a 4M color monitor that was not specifically designed for medical purposes was used in this study. In a recent paper[13] the results of the readers using our workstation were compared to those obtained in a previous study,[6] in which ten experienced screening radiologists read original films of the same cases, to evaluate if the results of the readers were representative enough. It appeared that the performance of the readers were somewhat less on average, but high enough to draw valid conclusions.

ACKNOWLEDGMENTS

The authors would like to thank the readers that participated in the observer study. This work has been funded by grant KUN 2006-3655 of the Dutch Cancer Society and by NWO under BRICKS/ FOCUS grant number 642.066.605

REFERENCES

[1] Gilbert, F. J., Astley, S. M., Gillan, M. G. C., Olorunsolo, F., Wallis, M. G., James, J., Boggis, C., Duffy, S., and the CADET II Group, "Single reading with computer-aided detection for screening mammography," *New England Journal of Medicine* **359**(16), 1675–1684 (2008).

[2] Morton, M. J., Whaley, D. H., Brandt, K. R., and Amrami, K. K., "Screening mammograms: Interpretation with computer-aided detection–prospective evaluation," *Radiology* **239**(2), 375–383 (2006).

[3] Dean, J. C. and Ilvento, C. C., "Improved cancer detection using computer-aided detection with diagnostic and screening mammography: Prospective study of 104 cancers," *Am. J. Roentgenol.* **187**(1), 20–28 (2006).

[4] Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., and Elmore, J. G., "Influence of computer-aided detection on performance of screening mammography," *New England Journal of Medicine* **356**(14), 1399–1409 (2007).

[5] Gur, D., Stalder, J. S., Hardesty, L. A., Zheng, B., Sumkin, J. H., Chough, D. M., Shindel, B. E., and Rockette, H. E., "Computer-aided detection performance in mammographic examination of masses: Assessment," *Radiology* **233**(2), 418–423 (2004).

[6] Karssemeijer, N., Otten, J. D., Verbeek, A. L., Groenewoud, J. H., de Koning, H. J., Hendriks, J. H., and Holland, R., "Computer-aided detection versus independent double reading of masses on mammograms," *Radiology* **227**, 192–200 (2003).

[7] Manning, D., Ethell, S., and Donovan, T., "Categories of observer error from eye-tracking and afroc data," *SPIE Medical Imaging* **5372**, 90–99 (2004).

[8] Karssemeijer, N., "Use of prompt magnitude in computer aided detection of masses in mammograms," in [*Proceedings of the 8th International Workshop on Digital Mammography*], **4046**, 54–60 (2006).

[9] Gilbert, F. J., Astley, S. M., Boggis, C. R. M., Mcgee, M. A., Griffiths, P. M., Duffy, S. W., Agbaje, O. F., Gillan, M. G. C., Wilson, M., Jain, A., Barr, N., Beetles, U. M., Griffiths, M. A., Johnson, J., Roberts, R. M., Deans, H. E., Duncan, K. A., and Iyengar, G., "Variable size cad prompts and mammography film reader decisions," *Breast Cancer Research* **10**, R72+ (2008).

[10] Mello-Thoms, C., "How does the perception of a lesion influence visual search strategy in mammogram reading?," *Academic Radiology* **13**(3), 275–288 (2006).

[11] Kundel, H. L., Nodine, C. F., Krupinski, E. A., and Mello-Thoms, C., "Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms," *Academic Radiology* **15**(7), 881–886 (2008).

[12] Nodine, C., Mello-Thoms, C., Kundel, H., and Weinstein, S., "Time course of perception and decision making during mammographic interpretation," *Am. J. Roentgenol.* **179**(4), 917–923 (2002).

[13] Karssemeijer, N., Hupse, A., Samulski, M., Kallenberg, M., Boetes, C., and den Heeten, G., "An interactive computer aided decision support system for detection of masses in mammograms," in [*Proceedings of the 9th International Workshop on Digital Mammography*], **5116**, 273–278 (2008).