Two complementary model observers to evaluate reconstructions of simulated micro-calcifications in digital breast tomosynthesis

Koen Michielsen[a*], Federica Zanca[b], Nicholas Marshall[b], Hilde Bosmans[b], Johan Nuyts[a]

[a]Dept. of Nuclear Medicine, KU Leuven, Belgium
[b]Dept. of Radiology, UZ Leuven, Belgium

ABSTRACT

New imaging modalities need to be properly evaluated before being introduced in clinical practice. The gold standard is to perform clinical trials or dedicated clinical performance related observer experiments with experienced readers. Unfortunately this is not feasible during development or optimization of new reconstruction algorithms due to their many degrees of freedom.

Our goal is to design a set of model observers to evaluate the performance of newly developed reconstruction methods on the assessment of micro-calcifications in digital breast tomosynthesis. In order to do so, the model observers need to evaluate both detection and classification of micro-calcifications. A channelized Hotelling observer was created for the detection task and a Hotelling observer working on an extracted feature vector was implemented for the classification task. These observers were evaluated on their ability to predict the results of human observers.

Results from a previous observer study were used as reference to compare performance between human and model observers. This study evaluated detection of small micro-calcifications ($100 - 200\ \mu$m) by a free search task in a power law filtered noise background and classification of two types of larger micro-calcifications ($200 - 600\ \mu$m) in the same background. Scores from the free search study were evaluated using the weighted JAFROC method and the classification scores were analyzed using the DBM MRMC method. The same analysis methods were applied to the model observer scores.

Results of the detection model observer were related linearly with the human observer results with a correlation coefficient of 0.962. The correlation coefficient for the classification task was 0.959 with a power law non-linear regression.

1. INTRODUCTION

Digital breast tomosynthesis (DBT) is a relatively new 3D mammography technique that improves upon digital mammography by allowing better visualization of low contrast lesions due to the removal of interference from overlapping tissues. Visualization of smaller, high contrast lesions is however slightly deteriorated.[1,2] For this reason we concentrate on micro-calcifications when developing new reconstruction algorithms.[3]

Ideally, each new optimization of reconstruction algorithms should be verified by an observer study with experienced readers, but due to the many possible parameters for most iterative reconstruction methods this is not feasible in practice. Therefore we propose two model observers that can give a first evaluation of new methods at all major steps in the development and perform a selection of the parameters prior to more in depth validation studies with human observers.

The two observers each consider a different aspect of the assessment of micro-calcifications: detection of the smallest calcifications ($100 - 200\ \mu$m) and classification of slightly larger calcifications ($200 - 600\ \mu$m). This is to make sure that when optimizing for a certain task, we don't unintentionally end up worsening performance on the other task. The model observers are described in section 2.3. Both are designed to work with the data sets and analysis methods discussed in section 2.1.

[*] koen.michielsen@uzleuven.be

Figure 1. Examples of simulated clusters containing small spherical micro-calcifications.

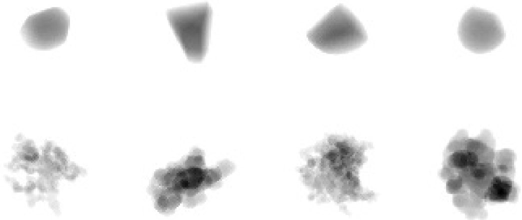

Figure 2. Examples of simulated smooth (top row) and irregular (bottom row) micro-calcifications.

2. MATERIALS AND METHODS

2.1 Phantom Creation

Two data sets that were previously used in an observer experiment with human readers[3] are now used to evaluate the model observers. The methods used to create the original data sets are summarized here.

Background images were simulated by filtering white noise with a power law filter $f(\nu) = \kappa/\nu^{\beta}$, with ν the frequency, $\beta = 3$ and $\kappa = 10^{-5}$ mm^{-1}.[4,5] The backgrounds measured 500x500x200 isotropic voxels with sides of 85 μm. These volumes were placed in one of three possible locations, always at the chest-side detector edge: central at 27 mm above the detector plane, central at 67 mm above the detector plane, or 75 mm off center at a height of 47 mm. These backgrounds were then used to generate two data sets: one containing simulated clusters of small spherical calcifications with diameters between 100 and 200 μm, and one containing a grid of irregular calcifications with diameters between 200 and 600 μm.

In the first set, each volume contained a random number of clusters (Poisson distributed with a mean of 1.0), with each cluster containing a random number of calcifications (Poisson distributed with a mean of 2.5). The clusters were placed in a random location within the volume, but not at the edge. Figure 1 shows some examples of the simulated clusters. For the second set two series of micro-calcifications (smooth, corresponding to Le Gal II and irregular, corresponding to Le Gal IV) were created according to the method of Näppi,[6] shown in figure 2.

Measurements of these volumes were simulated to match the Siemens Mammomat Inspiration system. This included supersampling the detector pixel pitch from 85 μm to 17μm pixel pitch and sampling multiple x-ray source positions for each nominal exposure angle, corresponding to a moving source with an exposure time of 120 ms per angle. X-ray energy was set to 20 keV and Poisson noise was generated with a blank scan of 1500 photons per pixel (12.5 μGy detector dose after attenuation). The simulated measurements were reconstructed with the Siemens iFBP method[7] and three iterations of two iterative reconstruction algorithms with resolution modeling,[3] one with and one without smoothing prior.

2.2 Data Analysis

The first data set was originally evaluated by human observers using the free search ROC (FROC) paradigm, where the readers have to indicate the location of each lesion they find and indicate how certain they are that the lesion is present, using scores between 1 for lowest, and 4 for highest certainty. These data were then analyzed using the weighted JAFROC method by Chakraborty.[8] It should be noted that this method only uses true positive scores in abnormal images and the maximal false positive score for normal images, which means that the model observer for this set will only need to provide these data, and does not need to score false positives in abnormal images.

The second data set was evaluated using the ROC method, with the readers classifying each calcification as either smooth or irregular and providing their certainty of this classification (low, medium or high certainty). These results were reformed to scores between 1 (smooth, high certainty) and 6 (irregular, high certainty) and then analyzed using the DBM MRMC method.[9]

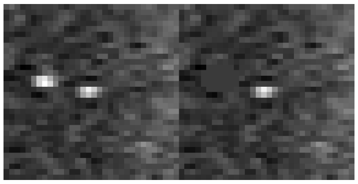

Figure 3. Masked calcification.

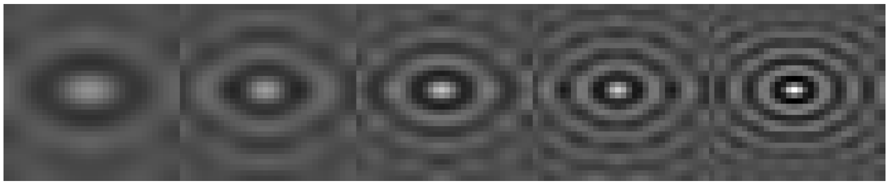

Figure 4. The model observer channels.

2.3 Model Observers

To evaluate detectability, we constructed a channelized Hotelling observer according to the msCHOc model proposed by Platiša.[10] This model uses two dimensional channels on each slice of the selected volume, and uses the combined output as input for the Hotelling observer. In our case, a surrounding region of interest (ROI) of 32x32x3 voxels (2.72 x 2.72 x 3.0 mm^3) was selected for each calcification. Because the simulated calcifications appear in clusters, each ROI centered on a single calcification could include other, nearby calcifications. Since this would negatively impact the score of the selected calcification, they were masked out of the image by replacing the pixel values with the median of pixel values at the same distance from the central calcification, as shown in figure 3.

At this point, the five channels in figure 4 were selected. These were created by applying the inverse Fourier transform to an elliptical band in the frequency domain. The frequencies are listed in table 1. The factor of (2.72 mm)$^{-1}$ is derived from the ROI size: 32 pixels of 0.085 mm. These bands were chosen to approximately match the non-isotropic point spread function of the tomosynthesis images. In practice, we found that the exact shape of the channels had little influence on observer performance.

	Tube Travel Direction	Front-Back Direction
Channel 1	[2.0 – 3.0] / 2.72 mm	[3.0 – 4.5] / 2.72 mm
Channel 2	[3.0 – 4.0] / 2.72 mm	[4.5 – 6.0] / 2.72 mm
Channel 3	[4.0 – 5.0] / 2.72 mm	[6.0 – 7.5] / 2.72 mm
Channel 4	[5.0 – 6.0] / 2.72 mm	[7.5 – 9.0] / 2.72 mm
Channel 5	[6.0 – 7.0] / 2.72 mm	[9.0 – 10.5] / 2.72 mm

Table 1. Frequency bands of the channels shown in figure 4.

Although the exact location of each calcification is known in the original phantom, they might have shifted slightly due to noise in the projection and reconstruction. Therefore each calcification is evaluated with the ROI centered on each of the voxels of a 3x3x3 region around its original location. The maximum score of these 27 evaluations is then selected as the final score.

In first instance, the observer was trained separately for five groups of lesion diameters: 5 bins of 20 μm between 100 μm and 200 μm. Each lesion was then evaluated using the matching template and false positives were scored using the template for the smallest lesion size (100 – 120 μm). However, this method resulted in very low AFROC area under the curve (AUC) for the model observer in comparison to the human observers, as shown in figure 5. Further investigation revealed that this was caused by an overestimation in the false positive detection rate by the model observer.

To solve this problem, we looked at the search results of our human observers. We found that calcifications smaller than 150 μm were only detected when they were part of a cluster, while the presence of other calcifications had no effect for calcifications larger than 150 μm, as can be seen figure 6. Therefore, we made the following assumptions about the human search model: images are first searched for larger calcifications (150 – 200 μm) which are relatively easy to spot, after which the neighborhood around these larger calcifications is searched for smaller (100 – 150 μm), lower contrast calcifications. We try to emulate this search model in the observer by creating two templates instead of the previous five: one template for smaller calcifications between 100 μm and 150 μm and one for large calcifications between 150 μm and 200 μm. All lesion locations in abnormal images and non-lesion locations in normal images were then evaluated with the larger template, except for small calcifications in a cluster with at least one large calcification, which were evaluated with the small template.

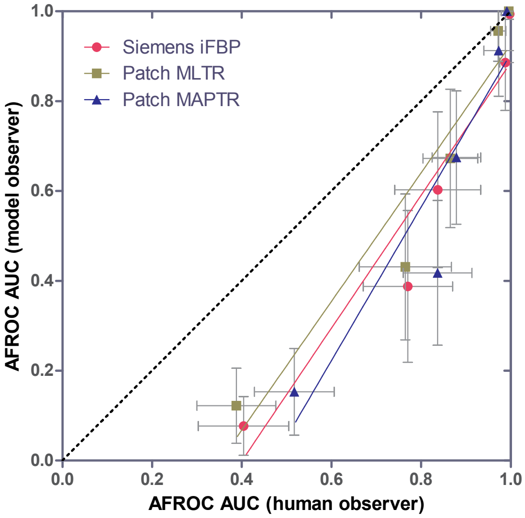

Figure 5. Underestimation of human AFROC AUC when search methodology is not taken into account.

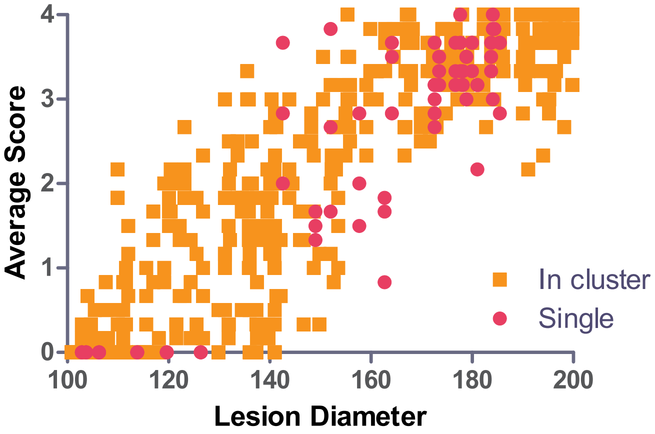

Figure 6. Effect of being part of a cluster on average lesion score for human observers.

For the second model observer, used to assess the classification of micro-calcifications as smooth or irregular, we simplified the Radon transform method described by Bocchi.[11] At first, only the neural network analysis of the final feature vector was replaced by a Hotelling observer, but these results were then very dependent on the outcome of the lesion segmentation. If the segmentation was successful, the observer scored much better than human observer, while both scored comparably when the segmented region was too large. Therefore we decided to remove the segmentation step and apply the Radon transform to a 31x31 pixel ROI centered on the calcification in its focus plane.

To calculate the 7-element feature vector, the Radon transforms of eight equally spaced angles ($\frac{n\pi}{8}$, $n \in \{0, 1, 2, 3, 4, 5, 6, 7\}$) were selected. The elements of these eight arrays were squared and added, producing an new array containing eight elements. The elements in this new array were then shifted left (with the previous first element being added as the last element of the shifted array) until the last element contained the maximum value in the array. The first seven elements of the array were then divided by the maximum value (the last element), resulting in a 7-element feature vector describing the calcification. This vector was then evaluated with a regular Hotelling observer.

The newly introduced model observers for search and classification were evaluated on their ability to predict results from the preceding human observer studies, either directly or after transformation by a monotonic function. The first model was trained on the same dataset used to train the human observers. The second model observer had access to a larger training set than the human observers (300 instead of 100 cases, which was smaller than the training set for the detection study, where the free search aspect allowed thousands of true negative cases). This was needed in order to reach the minimum number of cases to calculate the covariance matrix. None of the data in the evaluation set was used for training. Both model observers were trained for each combination of lesion diameter, position and reconstruction.

To obtain additional points for the comparisons between human and model observers, we used subanalyses based on lesion size. For the detection task, the calcifications were split in five groups: 100 – 120 μm, 120 – 140 μm, 140 – 160 μm, 160 – 180 μm and 180 – 200 μm. Each group contained 32 normal cases and 21 to 26 abnormal cases. For the classification task, each lesion diameter (200, 300, 400, 500 and 600 μm) was used as a group, and these consisted of between 28 and 34 normal cases and 28 to 32 abnormal cases. The ROC or AFROC scores and 95% confidence intervals were calculated for each group for both human and model observers. To determine if human and model observers agreed, correlation coefficients, regression lines and 95% confidence intervals were determined for each reconstruction separately and for all data combined.

3. RESULTS

Figures 7 and 8 show the area under the curve (AUC) of the model observers as a function of the human observer AUC. Linear regression lines were selected for the first model observer (search and detection) and a rescaled exponential curve ($y = 0.5 + (2x - 1)^a/2$) was chosen for the second observer (classification). The parameters of the regression lines, listed in table 2, are not significantly different for the different reconstructions (slope: p = 0.816; intercept: p = 0.156; exponent a: p = 0.706). This means it is possible to select one regression line for each data set, with the parameters for the combined data sets from the bottom row of table 2. In this case the correlation coefficient between human and model observer scores is 0.962 for the detection task and 0.959 for the classification task.

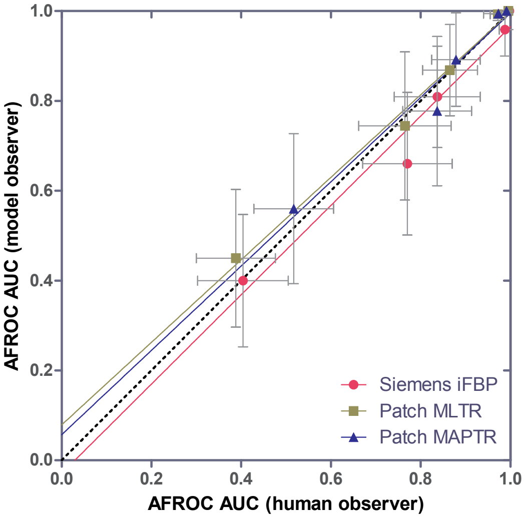

Figure 7. Detection model observer.

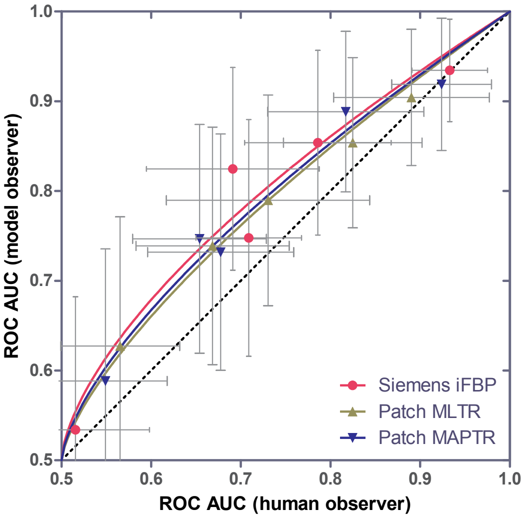

Figure 8. Classification model observer.

Reconstruction	Detection observer		Classification observer
	Slope (95% CI)	Intercept (95% CI)	Exponent a (95% CI)
Siemens iFBP	0.994 (0.653 – 1.335)	0.029 (-0.383 – 0.236)	0.639 (0.417 – 0.861)
Patch MLTR	0.918 (0.747 – 1.090)	-0.086 (-0.293 – 0.058)	0.707 (0.633 – 0.780)
Patch MAPTR	0.938 (0.587 – 1.290)	-0.061 (-0.606 – 0.191)	0.680 (0.534 – 0.817)
Combined data	0.955 (0.841 – 1.069)	-0.033 (-0.151 – 0.060)	0.676 (0.610 – 0.742)

Table 2. Parameters of the regression lines.

4. DISCUSSION

Our goal was to design a set of model observers to evaluate the performance of newly developed reconstruction methods on detection and characterization of micro-calcifications in DBT in order to narrow down the number of candidates for human observer studies. We found that our model observers closely predict results from the previous human observer study, for both aspects of the assessment of micro-calcifications.

The proposed method has some limitations: it assumes that detection and classification tasks capture the essence of the assessment of micro-calcifications. In addition, we do not take into account the visualization of masses, which would be necessary for a more complete description of the quality of DBT images. We should also note that at this time, the selection of the channels for the CHO (table 1 and figure 4) is based on the data itself which could make the validation suspect. We don't believe this to be a serious problem at this time since the exact choice of the channels mainly influenced the slope of the linear regression lines and had little effect on the correlation itself, and the fact that performance of channelized Hotelling observers is generally accepted to correlate well with human observers for simple tasks such as detecting spherical lesion in a background of power-law filtered noise.

Even considering these limitations, we still expect our model observers to perform close to humans when evaluating a range of reconstructions, since the validation already contains two very different types with an FBP based method and two iterative methods. In the end, it is important that the model observer can select a limited number of candidates out of the large parameter space available for iterative reconstruction to be included in further evaluation studies with human observers.

5. CONCLUSIONS

We proposed and validated a set of model observers to efficiently evaluate the assessment of micro-calcifications in numerous variants of our iterative reconstruction methods for digital breast tomosynthesis.

Acknowledgements

This work is supported by Siemens Healthcare and by SBO project QUANTIVIAM (060819) of the Institute for Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

REFERENCES

[1] Dobbins III, J. T., "Tomosynthesis imaging: At a translational crossroads," *Medical Physics* **36**(6), 1956–1967 (2009).

[2] Svahn, T. M., Chakraborty, D. P., Ikeda, D., Zackrisson, S., Do, Y., Mattsson, S., and Andersson, I., "Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy," *The British Journal of Radiology* **85**(1019), 1074-1082 (2012).

[3] Michielsen, K., Van Slambrouck, K., Jerebko, A. K., and Nuyts, J., "Patchwork Reconstruction with Resolution Modeling for Digital Breast Tomosynthesis," in [*Proceedings of the second international conference on image formation in X-ray computed tomography*], 21–24 (2012).

[4] Metheany, K. G., Abbey, C. K., Packard, N., and Boone, J. M., "Characterizing anatomical variability in breast CT images," *Medical Physics* **35**(10), 4685–4694 (2008).

[5] Engstrom, E., Reiser, I. S., and Nishikawa, R. M., "Comparison of power spectra for tomosynthesis projections and reconstructed images," *Medical Physics* **36**(5), 1753–1758 (2009).

[6] Näppi, J., Dean, P. B., Nevalainen, O., and Toikkanen, S., "Algorithmic 3D simulation of breast calcifications for digital mammography," *Computer Methods and Programs in Biomedicine* **66**, 115–124 (2001).

[7] Ludwig, J., Mertelmeier, T., Kunze, H., and Harer, W., "A Novel Approach for Filtered Backprojection in Tomosynthesis Based on Filter Kernels Determined by Iterative Reconstruction Techniques," in [*[IWDM '08 Proceedings of the 9th international workshop on Digital Mammography]*], 612–620(2008).

[8] Chakraborty, D. P. and Berbaum, K. S., "Observer studies involving detection and localization: modeling, analysis, and validation," *Medical physics* **31**, 2313–2330 (2004).

[9] Dorfman, D. D., Berbaum, K. S., and Metz, C. E., "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method," *Investigative radiology* **27**, 723–731 (1992).

[10] Platiša, L., Goossens, B., Vansteenkiste, E., Park, S., Gallas, B. D., Badano, A., and Philips, W., "Channelized Hotelling observers for the assessment of volumetric imaging data sets," *Journal of the Optical Society of America A* **28**, 1145–1165 (2011).

[11] Bocchi, L. and Nori, J., "Shape analysis of microcalcifications using Radon transform.," *Medical Engineering & Physics* **29**, 691–698 (2007).