



Universiteit  
Leiden  
The Netherlands

## Signal template generation from acquired images for model observer-based image quality analysis in mammography

Balta, C.; Bouwman, R.W.; Veldkamp, W.J.H.; Broeders, M.J.M.; Sechopoulos, I.; Engen, R.E. van

### Citation

Balta, C., Bouwman, R. W., Veldkamp, W. J. H., Broeders, M. J. M., Sechopoulos, I., & Engen, R. E. van. (2018). Signal template generation from acquired images for model observer-based image quality analysis in mammography. *Journal Of Medical Imaging*, 5(3).  
doi:10.1117/1.JMI.5.3.035503

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/79160>

**Note:** To cite this publication please use the final published version (if applicable).

# Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

## Signal template generation from acquired images for model observer- based image quality analysis in mammography

Christiana Balta  
Ramona W. Bouwman  
Wouter J. H. Veldkamp  
Mireille J. M. Broeders  
Ioannis Sechopoulos  
Ruben E. van Engen

**SPIE.**

Christiana Balta, Ramona W. Bouwman, Wouter J. H. Veldkamp, Mireille J. M. Broeders, Ioannis Sechopoulos, Ruben E. van Engen, "Signal template generation from acquired images for model observer-based image quality analysis in mammography," *J. Med. Imag.* **5**(3), 035503 (2018), doi: 10.1117/1.JMI.5.3.035503.

# Signal template generation from acquired images for model observer-based image quality analysis in mammography

Christiana Balta,<sup>a,b,†</sup> Ramona W. Bouwman,<sup>a,†</sup> Wouter J. H. Veldkamp,<sup>c</sup> Mireille J. M. Broeders,<sup>a,d</sup> Ioannis Sechopoulos,<sup>a,b</sup> and Ruben E. van Engen<sup>a,\*</sup>

<sup>a</sup>Radboud University Medical Center, Dutch Expert Centre for Screening (LRCB), Nijmegen, The Netherlands

<sup>b</sup>Radboud University Medical Center, Department of Radiology and Nuclear Medicine, Nijmegen, The Netherlands

<sup>c</sup>Leiden University Medical Centre, Department of Radiology, Leiden, The Netherlands

<sup>d</sup>Radboud University Medical Center, Radboud Institute for Health Sciences (RIHS), Nijmegen, The Netherlands

**Abstract.** Mammography images undergo vendor-specific processing, which may be nonlinear, before radiologist interpretation. Therefore, to test the entire imaging chain, the effect of image processing should be included in the assessment of image quality, which is not current practice. For this purpose, model observers (MOs), in combination with anthropomorphic breast phantoms, are proposed to evaluate image quality in mammography. In this study, the nonprewhitening MO with eye filter and the channelized Hotelling observer were investigated. The goal of this study was to optimize the efficiency of the procedure to obtain the expected signal template from acquired images for the detection of a 0.25-mm diameter disk. Two approaches were followed: using acquired images with homogeneous backgrounds (approach 1) and images from an anthropomorphic breast phantom (approach 2). For quality control purposes, a straightforward procedure using a single exposure of a single disk was found adequate for both approaches. However, only approach 2 can yield templates from processed images since, due to its nonlinearity, image postprocessing cannot be evaluated using images of homogeneous phantoms. Based on the results of the current study, a phantom should be designed, which can be used for the objective assessment of image quality. © 2018 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.5.3.035503](https://doi.org/10.1117/1.JMI.5.3.035503)]

Keywords: mammography; model observers; image quality; template; nonprewhitening model observer; channelized Hotelling observer.

Paper 18097R received May 7, 2018; accepted for publication Aug. 13, 2018; published online Sep. 8, 2018.

## 1 Introduction

Model observers (MOs) are currently being evaluated for use in image quality assessment of x-ray imaging systems.<sup>1–6</sup> In these studies, the nonprewhitening (NPW) and/or the channelized Hotelling observer (CHO) are the MOs most commonly used to evaluate the detectability of an object in images acquired with a particular system. These MOs are chosen because of their relatively limited computer processing power requirements. More importantly, the evaluated MO-based detection task is assumed to be, at least partly, representative of the radiological task.

In screening mammography, the task of the radiologist is identifying the very few cases containing (subtle) signs of malignancy in a large set of images. The presence of structures like masses, calcifications, and/or architectural distortions are features, which might be signs of malignancy.<sup>7</sup> Therefore, it could be argued that the probability of detecting such features when present is a measure of image quality. In mammography, it has been demonstrated that the ability to detect calcifications can be predicted, to some extent, by evaluating the detection of small disk-shaped objects embedded in homogeneous phantoms.<sup>8</sup>

Breast images are processed to optimize the display of images on softcopy reading stations for reading by radiologists.<sup>9,10</sup> However, due to their nonlinear nature, the impact of this image postprocessing on image quality cannot be evaluated using images of homogeneous phantoms. This is because the behavior of these algorithms may be different when applied to these phantom images as opposed to real patient breast images, due to differences in image characteristics (for example, in cases of histogram-based contrast adaptation). Therefore, after applying image processing, the system may no longer be assumed to be linear. MOs do not require the system to be linear nor the use of homogeneous phantoms. Therefore, MOs, in combination with anthropomorphic breast phantoms, have been proposed as appropriate for assessment of image quality.<sup>1,2,4,11,12</sup> However, to the best of our knowledge, there has not yet been a proposal on the actual procedures required to use MOs as part of a quality control (QC) process.

In this study, we focused on the detection of a 0.25-mm diameter gold disk (the signal), which is evaluated by the nonprewhitening MO with eye filter (NPWE) and the CHO. The NPWE correlates the images with the expected signal and requires that the signal to be detected is known. For the CHO, the images are divided into two sets: a training set and a test set. The evaluation is then performed by correlating the channelized covariance matrix of the training set with the test set. If the pixel

\*Address all correspondence to: Ruben E. van Engen, E-mail: [r.vanengen@lrcb.nl](mailto:r.vanengen@lrcb.nl)

<sup>†</sup>Both authors contributed equally.

correlation is independent of the signal and equal for the images with and without the signal, the training set could be limited to images without signal and a template of the expected signal.

In the previous work, we proposed a method to construct a template of the expected signal from acquired images for use in the NPWE.<sup>13</sup> The proposed procedure was subsequently used to evaluate the detection probability of disk-shaped objects in images of two digital mammography (DM) systems with and without image processing applied.<sup>1</sup> The latter study demonstrated that the proposed procedure has potential for objective evaluation of processed images. However, the proposed procedure<sup>13</sup> to construct the template is not yet optimized for implementation in QC procedures. The number of exposures needed for this approach was too large to be used in QC procedures. Therefore, the goal of this work is to optimize the template construction from acquired images such that the number of acquisitions required to generate a signal template is as low as possible while resulting in a sufficiently accurate prediction of the values of the area under the (receiver operating) curve (AUC).

## 2 Methods

### 2.1 Introduction of NPWE and CHO

MOs are used to determine a decision variable  $\lambda$  for images with ( $i = 1$ ) and without ( $i = 2$ ) the signal to be detected. For linear MOs, this  $\lambda$  is estimated from a linear transformation between an observer template ( $\mathbf{w}$ ) and the image vector ( $\mathbf{g}$ ) via:

$$\lambda_i = \mathbf{w}^t \mathbf{g}_i, \quad (1)$$

where  $t$  is the transpose, and  $i$  is the image class of two-dimensional images, which are treated as one-dimensional vectors (vectors and matrices are annotated using bold symbols).

Using Eq. (1), the decision variables based on the images with and without signal,  $\lambda_1$  and  $\lambda_2$ , respectively, are estimated. Subsequently, the performance of the MO is evaluated using the AUC estimated using the methodology described by Gallas.<sup>14</sup>

$$\text{AUC} = \frac{1}{N_1 N_2} \sum_{j=0}^{N_1} \sum_{k=0}^{N_2} \Delta(\lambda_{1j} - \lambda_{2k}), \quad (2)$$

where  $\Delta(\cdot)$  is a step function, which equals 1 when  $(\lambda_{1j} - \lambda_{2k})$  is positive, 0 when  $(\lambda_{1j} - \lambda_{2k})$  is negative and 0.5 when  $\lambda_{1j} = \lambda_{2k}$ .  $N_1$  and  $N_2$  are the total number of signals present and signal absent images, respectively. The variance of the AUC is estimated using the one-shot estimate for a single observer<sup>14</sup> based on the estimated decision variables.

#### 2.1.1 NPWE

For the NPW, the signal template ( $\mathbf{w}_{\text{NPW}}$ ) equals the expected signal ( $\mathbf{s}$ ) to be detected:

$$\mathbf{w}_{\text{NPW}} = \mathbf{s}. \quad (3)$$

To account for the frequency response of the human eye, an eye filter is included. We refer to this MO as the NPWE MO.<sup>15</sup> After inclusion of the eye filter ( $\mathbf{E}$ ), the decision variable is estimated using the following:

$$\lambda_i = [\mathbf{E}^t \cdot \mathbf{E} \cdot \mathbf{s}]^t \cdot \mathbf{g}_i, \quad (4)$$

where the eye filter is given in line pairs per degree ( $\phi$ ) and is defined as in our previous work:<sup>1</sup>

$$\mathbf{E}(\phi) = \phi^{1.4} \cdot e^{-0.013 \cdot \phi^{2.6}}. \quad (5)$$

These eye-filter parameters were chosen because they were deemed appropriate for this type of background, and they were found to be able to predict human detection performance, based on the work of Bouwman et al.<sup>16</sup>

#### 2.1.2 CHO

For the CHO, the dimensionality of the image data  $\mathbf{g}$  is reduced using channels. The channel set used in this study is the dense difference of Gaussian as defined by Abbey and Barrett<sup>17</sup> is a matrix  $\mathbf{U}$ :

$$\mathbf{U}_j(f) = e^{-\frac{1}{2} \left( \frac{f}{Q\sigma_j} \right)^2} - e^{-\frac{1}{2} \left( \frac{f}{\sigma_j} \right)^2}, \quad (6)$$

where  $f$  is the spatial frequency in lp/pixel,  $Q$  is the bandwidth of the channels, and  $\sigma_j$  is the width of the Gaussian function of the  $j$ 'th channel, and  $\sigma_j$  is defined by

$$\sigma_j = \sigma_0 a^j, \quad (7)$$

where  $a$  is a scaling parameter. The signal template of the CHO ( $\mathbf{w}_{\text{CHO}}$ ) is defined by

$$\mathbf{w}_{\text{CHO}} = \mathbf{K}_c^{-1} [\overline{\mathbf{g}_{c1}} - \overline{\mathbf{g}_{c2}}], \quad (8)$$

where  $\overline{\mathbf{g}_{ci}}$  is the mean channelized image vector of class 1 and 2, respectively, and  $\mathbf{K}_c$ , the interclass covariance matrix, is described by

$$\mathbf{K}_c = \frac{1}{2} [\mathbf{\kappa}_{c1} + \mathbf{\kappa}_{c2}], \quad (9)$$

where  $\mathbf{\kappa}_{ci}$  is as follows:

$$\mathbf{\kappa}_{ci} = \text{cov}(\mathbf{U}^t \mathbf{g}_i). \quad (10)$$

This covariance calculation for CHO experiments is found in the literature, for example, in the work of Platiša et al.<sup>18</sup> and Yu et al.<sup>19</sup> In this study, this method is referred to as ‘‘SP-trained.’’

If the signal ( $s$ ) can be assumed to be independent of the structure inside the images, then the observer template ( $\mathbf{w}_{\text{CHO}}$ ) can be estimated using only the channelized covariance matrix of images without signal ( $\mathbf{\kappa}_{c2}$ ) from Eq. (10), where  $i = 2$ , and the signal template ( $s$ ):

$$\mathbf{w}_{\text{CHO}} = \mathbf{\kappa}_{c2}^{-1} [\mathbf{U}^t s]. \quad (11)$$

This method for  $\mathbf{w}_{\text{CHO}}$  estimation is thoroughly used for the CHO, where the different acquired templates were investigated. This covariance calculation of the signal absent images for CHO experiments is found in the work of Diaz et al.<sup>20</sup> and Racine et al.<sup>21</sup> In this study, we compared the observer template given by Eqs. (8) and (11). For Eq. (11), different formulations of  $s$  were used as explained in more detail in Sec. 2.3.

The performance of the CHO was calculated in two steps. In the training step,  $\mathbf{w}_{\text{CHO}}$  is estimated. In the testing step, the decision variables are estimated. Ideally, for both the training and the testing stage, two independent sets of images of an infinite size

should be used. In practice, this is not feasible due to the limited number of exposures. In general, two different training–testing methods can be used when evaluating a limited set of images:<sup>22</sup> resubstitution, which means that the set of training images equals the set of test images, or the hold-out method, where the set of available images is divided into an independent training- and test-set. The results of both approaches will converge when a sufficient number of images is used.

## 2.2 Image Acquisition

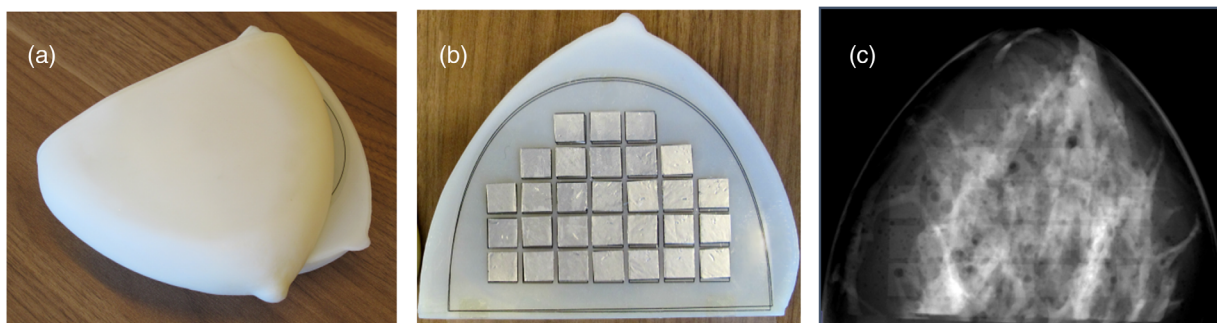
Images from an anthropomorphic breast phantom containing a sheet with gold disks of 0.25 mm in diameter (Fig. 1), as previously described,<sup>1</sup> were used in this study. Briefly, the anthropomorphic breast phantom is 3-D printed from a patient image acquired with a dedicated breast CT system (Koning Corp., Rochester, New York) that has undergone tissue classification and simulated mechanical compression. For the use as targets, in total, 29 gold disks with a thickness of 0.50  $\mu\text{m}$  were deposited on 0.05-mm-thick aluminum squares of approximately 10  $\times$  10 mm size. These aluminum squares were positioned on a transparent sheet. Images of this phantom with the sheet were acquired on a Selenia DM system (Hologic Inc., Bedford, Massachusetts) using a 28 kV W/Rh x-ray spectrum and a range of tube current-exposure time product values resulting in an incident air kerma at the phantom of approximately 0.36, 0.54, 0.81, and 1.26 mGy. To acquire images without signal, a similar sheet with aluminum squares without gold disks was used. Images with the disk-absent sheet were acquired in the same way as those with the disks present. From the acquired images, regions of interest (ROIs) with and without signal were extracted such that the center of the gold disk was approximately in the center of each ROI. For this purpose, an automatic routine was developed to select the aluminum squares and find the highest intensity within the center of this square to obtain the position of the gold disks.<sup>1</sup> Subsequently, the ROIs were cropped such that the gold disk is in the center of an ROI 89  $\times$  89 pixels in size. In total, 200 ROIs with and without signal were extracted from processed and unprocessed images for each dose level.

## 2.3 Template Generation from Acquired Images

To obtain a high-quality template from acquired images, a four-step procedure<sup>13</sup> is proposed:

1. Acquire 100 images of the transparent sheet with gold disks in air at a relatively high-incident air kerma (4.2 mGy). The sheet should be positioned such that the magnification is similar compared to that when the sheet is positioned inside the anthropomorphic breast phantom. During acquisition of these 100 images, the sheet is moved and repositioned four times from its original location (R1) after every 20 acquisitions. Each reposition involved a slight movement of the sheet in the left, right, anterior, and posterior directions, always making sure that the gold discs remained within the phantom. The reposition was done to take into account the partial filling effect in the template generation since the gold disk size was close to the system's pixel size. This means that 20 exposures were made at each of five positions (R1, R2, R3, R4, and R5).
2. Register all 100 (20  $\times$  5) images rigidly across repositions using the location of the gold discs as fiducials, which was feasible because they had the highest intensity. Images were registered on a whole-pixel level and not on a subpixel level thus without interpolation. Average over the new image ensemble and extract squares with the signal at their center.
3. Suppress the signal for each square inside a patch, which is  $\sim 2$  pixels bigger than the expected signal size using an inward interpolation algorithm with a built-in MATLAB function (MATLAB 2015a, Mathworks Inc.). For a 0.25-mm disk imaged on a system with an effective pixel size of 0.07 mm, the patch size for the inward interpolation was 5  $\times$  5 pixels.
4. Subtract the results of step 3 from the average image obtained in step 2 to get the signal template.

It has been demonstrated that after inclusion of the template shifting to address any location uncertainty of the signal center, the proposed procedure resulted in similar AUC measures compared to a synthetic signal for the NPWE.<sup>13</sup> This synthetic template was generated from a high-resolution binary disc. The binary disk was then rescaled and downsampled to match the size of the 0.25-mm diameter disc-shaped signal as being imaged by a system of pixel size 0.07 mm (without taking the modulation transfer function into account). Consequently,



**Fig. 1** Photograph of (a) the anthropomorphic breast phantom, (b) the transparent sheet with aluminum squares with gold disks, and (c) a DM image of the phantom with the embedded aluminum squares containing the gold dots.



the synthetic signal was of the same size as in the acquired images of the 0.25-mm gold discs.

However, the procedure described in steps 1 to 4 is very time-consuming, making it unsuitable for QC procedures and requiring optimization. For this optimization, we started from two different perspectives. Template generation from separately acquired images of (1) a homogenous background, as explained above, and (2) the anthropomorphic breast phantom with the sheet containing the signals.

For each acquired template, the NPWE and the CHO MO-based AUCs were determined for images acquired with an incident air kerma of  $\sim 0.36$  mGy. This dose level was used to achieve a SNR that would result in a MO performance of  $\sim 85\%$  correct. For the NPWE, the results were then compared to the synthetic template<sup>13</sup> and to the original proposal for construction of the acquired template with 100 acquisitions. In the case of the CHO, the AUC was obtained after training on both signal present and signal absent ROIs, referred to as SP-trained, and the results were then compared between the SP-trained and the original method to construct the acquired template.

### 2.3.1 Approach 1: Template from homogeneous backgrounds

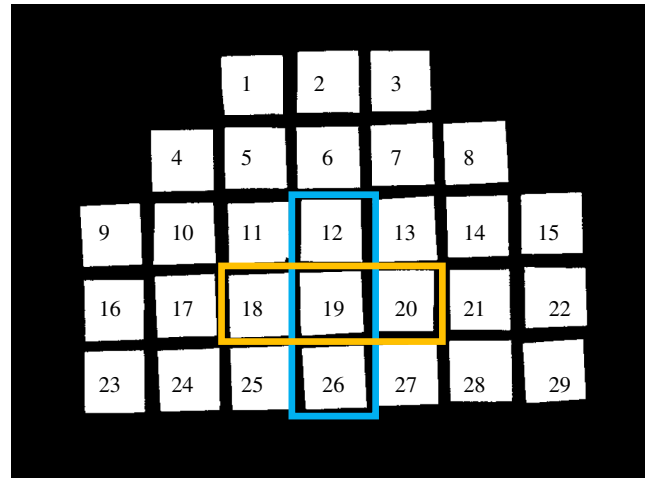
For the template generated from the separately acquired images, we evaluated eight different acquisition strategies. The AUC values obtained using the different acquired templates were compared to the AUC value obtained when using a synthetic signal. In Table 1, an overview of the eight acquisition strategies evaluated can be found. The synthetic signal used was produced by creating an artificial high-resolution disk, which was then down-sampled to match the 0.25-mm disk.

Using the eight acquisition strategies, the following five experiments were conducted:

1. Matching the relative position of the 29 gold disks (see Fig. 2) in the signal present image and the templates.
2. Obtaining an average template by averaging the signal over nine disk positions (average from disk positions:

**Table 1** The eight acquisition strategies to generate the signal template.

	Strategy	Number of exposures	Number of (re)positions	Total number of acquisitions
I	R1 to R5-acq.1 to 20	20	5	100 (original proposal)
II	R1-acq.1 to 20	20	None	20
III	R1-acq.1 to 10	10	None	10
IV	R1-acq.1 to 5	5	None	5
V	R1-acq.1	1	None	1
VI	R1 to R5-acq.1 to 10	10	5	50
VII	R1 to R5-acq.1 to 5	5	5	25
VIII	R1 to R5-acq.1	1	5	5



**Fig. 2** Layout of the transparent sheet with aluminum squares with 29 golden disks. The average 9/6/3/1 templates were obtained from disk-position “11, 12, 13, 18, 19, 20, 25, 26, 27,” “18, 19, 20, 25, 26, 27,” “18, 19, 20,” “12, 19, 26,” and 19, respectively. In this diagram, the chest wall side of the phantom and of the detector is below the 23–29 row of squares.

11, 12, 13, 18, 19, 20, 25, 26, 27, see Fig. 2), referred to as “average 9 template.”

3. Same as 2 but by averaging the signal over six disk positions (18, 19, 20, 25, 26, 27, see Fig. 2), referred to as “average 6 template.”
4. Averaging the signal over two different sets of three disk positions (“average 3 template\_a”: 18, 19, 20, see Fig. 2 orange rectangle; and “average 3 template\_b”: 12, 19, 26, see Fig. 2 blue rectangle).
5. Averaging the signal over one disk position (19, see Fig. 2), referred to as “average 1 template.”

### 2.3.2 Approach 2: Template from the anthropomorphic breast phantom

For the second approach, we explored template extraction directly from the acquired anthropomorphic breast phantom images. The benefit of direct extraction from acquired anthropomorphic breast phantom images is that image processing can be applied before the templates are constructed. Within the anthropomorphic breast phantom image, we have defined an area without structures, simulating a fatty area. The sheet with the gold disk was subsequently positioned in the anthropomorphic breast phantom such that one gold disk was positioned inside this area. Subsequently, five images were acquired at each of two phantom incident air kermas: 1.26 and 4.2 mGy. Subsequently, the sheet was moved slightly such that the disk remains inside the predefined “fatty area” of the phantom and again five exposures at each air kerma level were made. This procedure was repeated such that we acquired five exposures for each of the five (re)positions, for each dose level. From the acquired images, a template was generated in a similar way as described in Approach 1 (template from homogeneous backgrounds), and we compared the results of the MO using a template obtained after (1) a single acquisition, (2) five acquisitions,

and (3) five acquisitions with five different repositions for both dose levels.

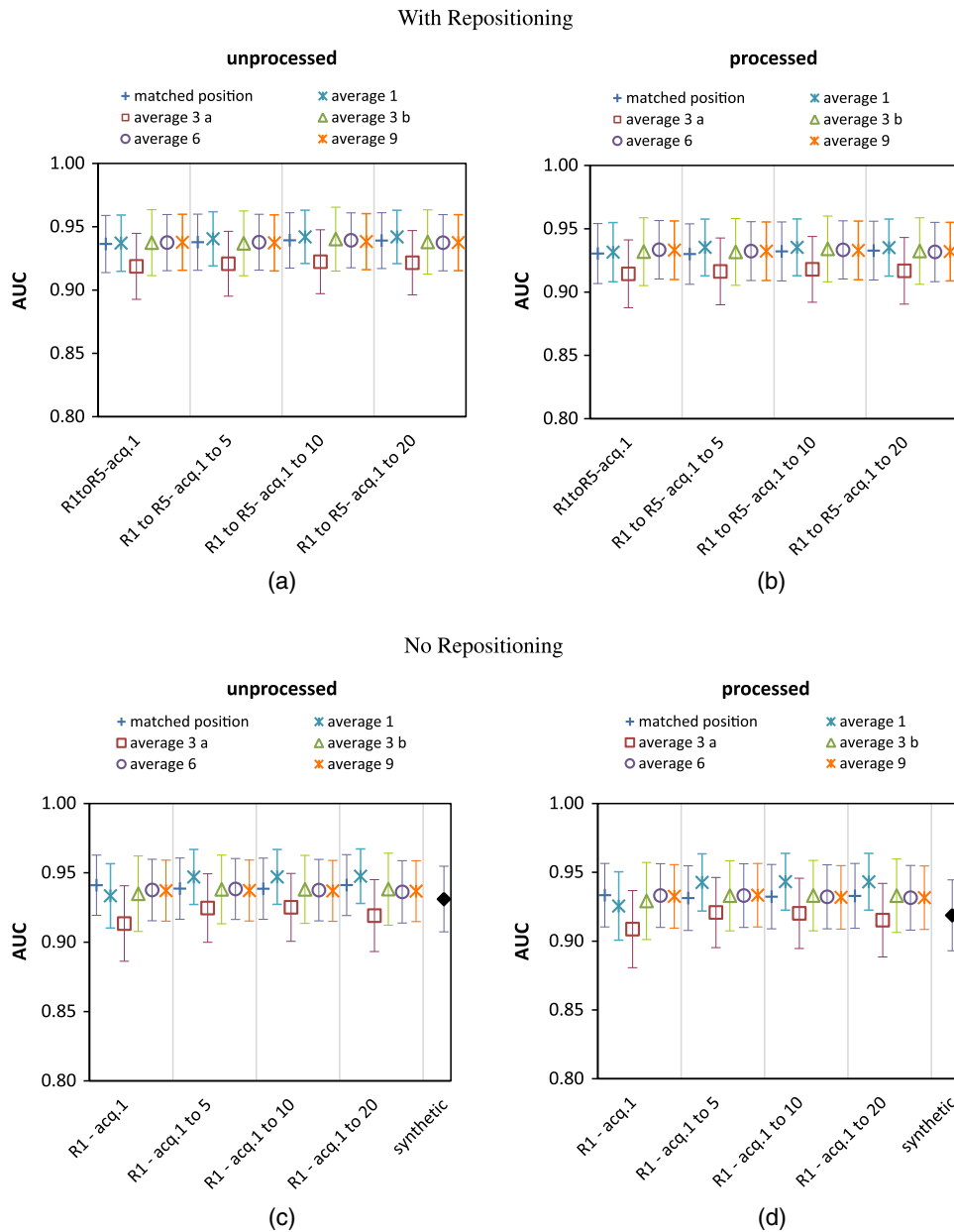
### 3 Results

#### 3.1 Acquired Template Strategies

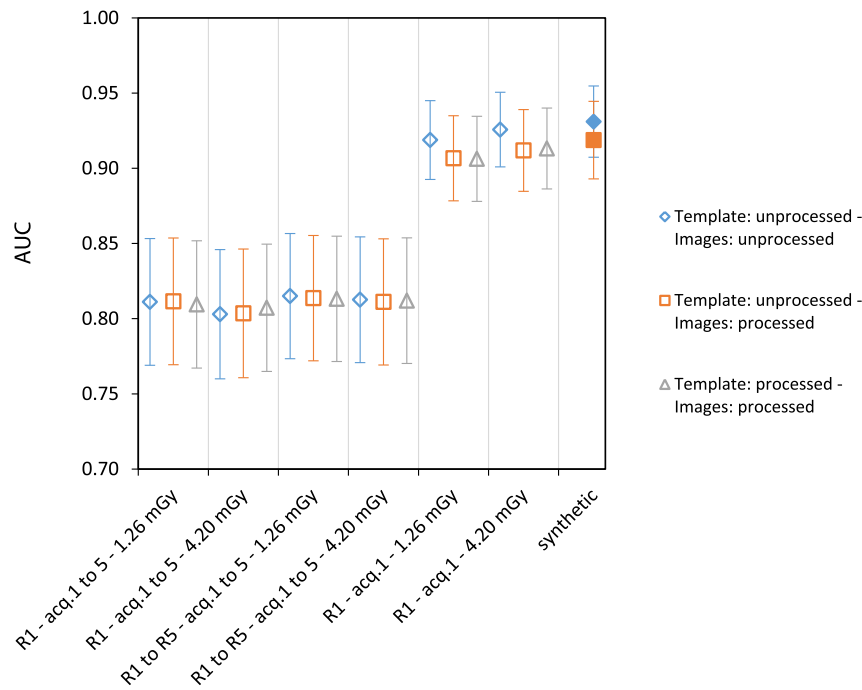
Figures 3 and 4 show the AUC estimated using the NPWE for the images acquired at 0.36 mGy with the templates generated using approaches 1 and 2, respectively. These figures show that for the acquired templates, the resulting AUC values are comparable to the AUC values obtained with the synthetic template. For approach 1 (template using a homogenous background, Fig. 3), it was found that averaging over multiple acquisitions

and multiple repositions did not have a significant impact on the AUC or its standard deviation ( $p > 0.05$ ). However, in case of averaging over three different disks (average 3 a and average 3 b), a dependence on the RoIs used is visually noted. This suggests that the disks involved in the averaging process were not identical, which could be caused by one of the gold disks being defective. Furthermore, matching the relative disk position of the template and the images compared to averaging over multiple disk positions did not affect the AUC.

For approach 2 (template using the anthropomorphic breast phantom images, Fig. 4), however, it was found that averaging over multiple acquisitions and multiple repositions decreased performance compared to using the synthetic template. By contrast, using a single image as the template resulted in a similar



**Fig. 3** AUC as a function of the different templates obtained from acquired images in a homogenous background (approach 1) using the NPWE in unprocessed (a) and (c) and in processed images (b) and (d). The AUC of the synthetic template is given as a reference in panels (c) and (d). The error bars represent 95% confidence intervals.



**Fig. 4** NPWE AUC as a function of the different templates obtained from the anthropomorphic breast phantom images (approach 2). The legend gives information about the processing status of the images that were used to construct the template and the analyzed images. The AUC of the synthetic template for both processed and unprocessed images is given as reference. The error bars represent 95% confidence intervals.

AUC value as with the synthetic template. Since the AUC was not dependent on the dose level used to generate the template, the same dose level as that used for the test images is preferred for practical reasons.

Figures 5 and 6 give the AUC for the CHO, showing that the observed AUC values obtained when training based on signal present and signal absent images (SP-trained) are comparable to those of training on signal absent images and the acquired template. For the CHO, the impact of the different approaches to obtain the template on the AUC was very similar to that found using the NPWE. The only noticeable difference when comparing Figs. 4 and 6 is that averaging the phantom template images seems to have a smaller effect on the AUC for the CHO than for the NPWE.

Since it was found for both the NPWE and the CHO that the number of images and the number of repositions did not influence the estimated AUC or its standard deviation for approach 1, choices regarding template construction can be based solely on practical aspects. Thus, to study the effect of dose on AUC, we chose to use the template constructed from a single disk position and a single image for approach 1. For approach 2, averaging multiple images was found to trigger only a small decrease in AUC values. Therefore, for approach 2, we also selected the template from a single image acquired with an incident air kerma of  $\sim 1.26$  mGy.

### 3.2 Effect of Dose

Figure 7 shows the five different templates used to evaluate the impact of dose on AUC. This figure shows that the template proposed originally [Fig. 7(b)] results in a somewhat blurred representation of the disk compared to the synthetic disk [Fig. 7(a)]. However, a disk-like signal can still be visually

identified. For both approaches 1 and 2, although identifying a disk is more difficult for the human observer, the results discussed above show that this had no or a limited effect on the reported AUC values.

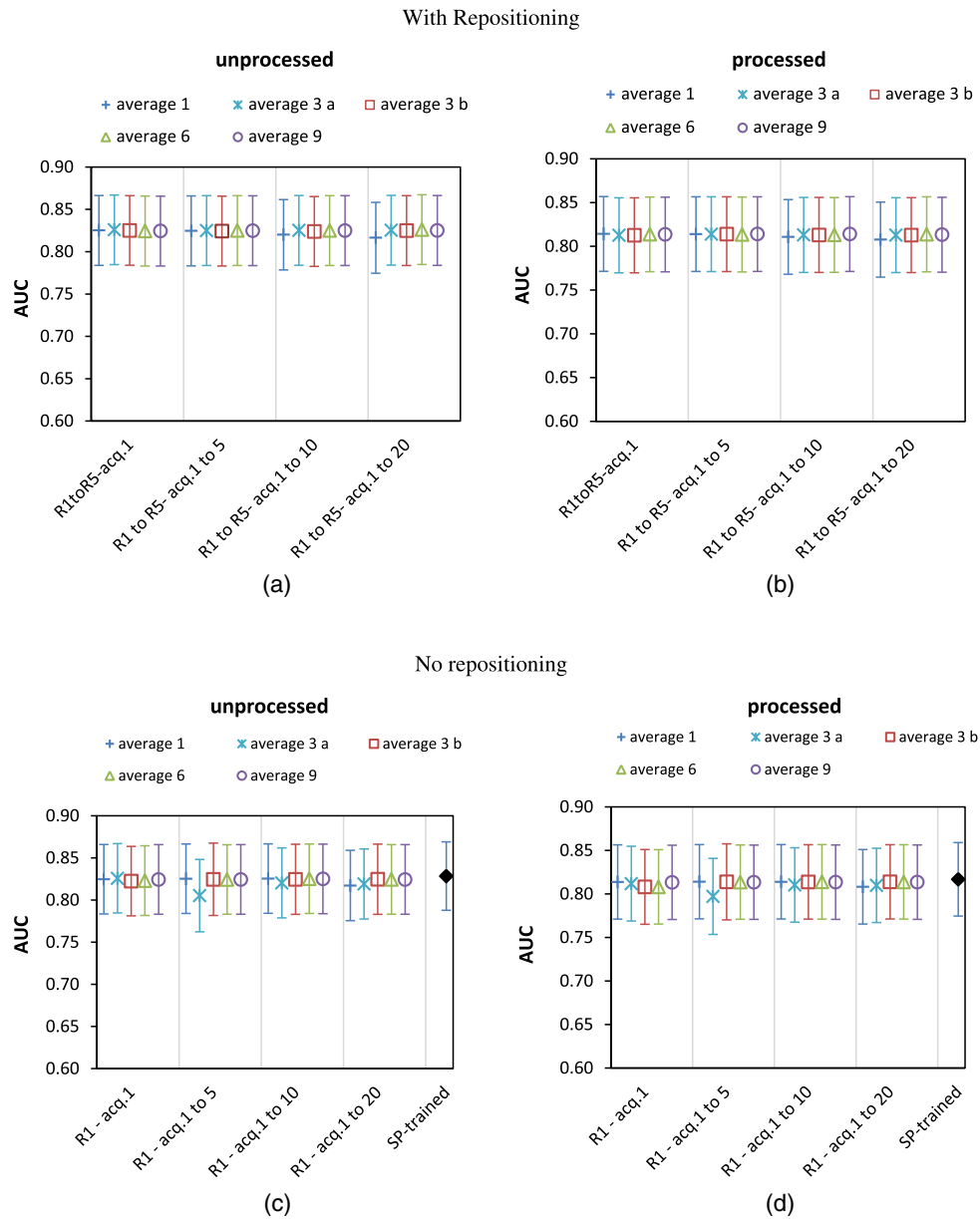
Figures 8 and 9 show the AUC as a function of dose for the NPWE MO and CHO, respectively, using the templates given in Fig. 7. These AUC values were compared with the synthetic template (for the NPWE) and the SP-trained condition (for the CHO). Both figures show that the trend with dose for the different acquired templates is as expected and comparable with the trend for the synthetic template or SP-trained condition, respectively. Comparing the NPWE and CHO graphs, it is noted that, in general, the NPWE results in substantially higher AUC values than the CHO. Comparison of the AUC values from different acquired templates shows that there are no substantial differences among them. However, minor differences are observed, especially at the lowest dose level, where approach 2 (templates from anthropomorphic breast phantom images) seems to result in slightly lower AUC values.

No difference in AUC values between unprocessed and processed images was noted for the different templates using the NPWE (Fig. 8).

Table 2 shows the error bars corresponding to Fig. 8. One can observe that the mean AUC at 0.54 mGy is lower than mean AUC when dose is doubled (i.e., at 1.26 mGy). However, the error bars show small partial overlap between the confidence interval of AUCs on images acquired at these two different dose level AUCs.

Moreover, the AUC of processed and unprocessed templates was almost identical. On the other hand, for the CHO (Fig. 9 and Table 3), minor differences were observed for the processed and unprocessed templates.





**Fig. 5** AUC as a function of the different templates obtained from acquired images in homogeneous background (approach 1) using the CHO in unprocessed (a) and (c) and in processed images (b) and (d). The AUC of SP-trained is given as a reference in panels (c) and (d). The error bars represent 95% confidence intervals.

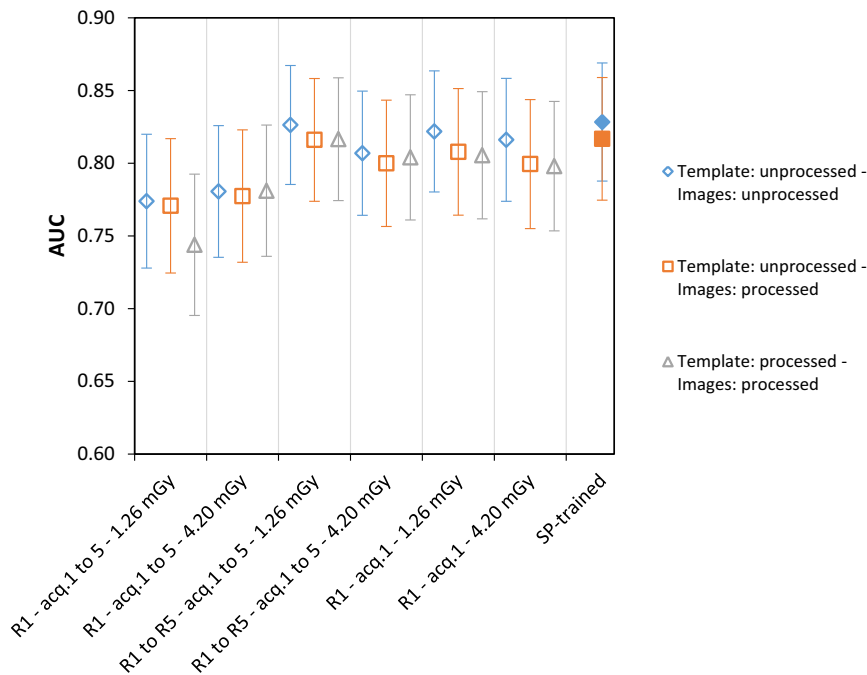
### 4 Discussion

In the current study, we optimized the procedure to generate signal templates from acquired images such that a minimum number of images can be used to construct the signal template without introducing a bias or an increase in the uncertainty in the measured AUC values. The most straightforward procedure to obtain the signal template was found to be derived from a single disk from a single acquired image of the anthropomorphic phantom. This approach is preferred since this method allows for the construction of templates from the processed image, resulting in a signal template that is expected to better resemble the signal in a clinical diagnostic task.

In some cases, AUCs were high (e.g., Fig. 3). If a more difficult detection task was used, the method to generate templates,

as described in this paper, would probably remain unaffected for an even lower dose level or lower signal attenuation (which both are methods to make the task more difficult). This is in fact shown in Fig. 4, where one can notice the robustness of the method even at lower AUC values. However, at lower AUCs, the error bars are larger. Therefore, at considerably lower AUC values, it would be expected that a higher number of images would be needed.

The images of the gold disks embedded in anthropomorphic backgrounds were acquired at low dose levels, in order to avoid MO performance of 100% correct. However, we found that the task is more difficult for the humans. We have actually conducted a human observer study with these images for the detection of the 0.25-mm disks and found AUC of ~0.85 at the dose



**Fig. 6** CHO AUC as a function of the different templates obtained from the anthropomorphic breast phantom images (approach 2). The legend gives information on the processing status of the images that were used to construct the template and whether the test images were unprocessed or processed. The AUC for SP-trained using both processed and unprocessed images are given as reference. The error bars represent 95% confidence intervals.

level of 0.36 mGy. This is published by Balta et al.<sup>1</sup> In QC procedures, we envision that the human observer performance is predicted rather than matched with the MO. We would like to predict the human out of the MO performance using a regression curve, for instance, and not matching the exact values of AUC of the MO and AUC of the humans.

Although the results for using acquired templates from a single object in a single exposure look promising, it is noted that the signal does not seem to resemble a disk due to blurring and/or noise. This is even more apparent when the disk is obtained from images of the anthropomorphic phantom (approach 2). It is expected that this difference in visual appearance arose because of the low signal difference to noise ratio of the object. This means that we might need to improve the inward interpolation used to cancel out the noise. Noise reduction can also be obtained by averaging multiple exposures together. However, this study shows that averaging over multiple acquisitions results in more noise and/or blurring of the signal (Fig. 7). It is expected that these issues could be reduced by applying a sub-pixel-shift approach when summing the individual images to construct the template (super-resolution<sup>23</sup>). Nevertheless, it could be argued that the benefit of further optimization would be limited since the AUC value of the acquired template is already approaching the AUC obtained using a synthetic template, which is, by definition, noise-less and sharp.

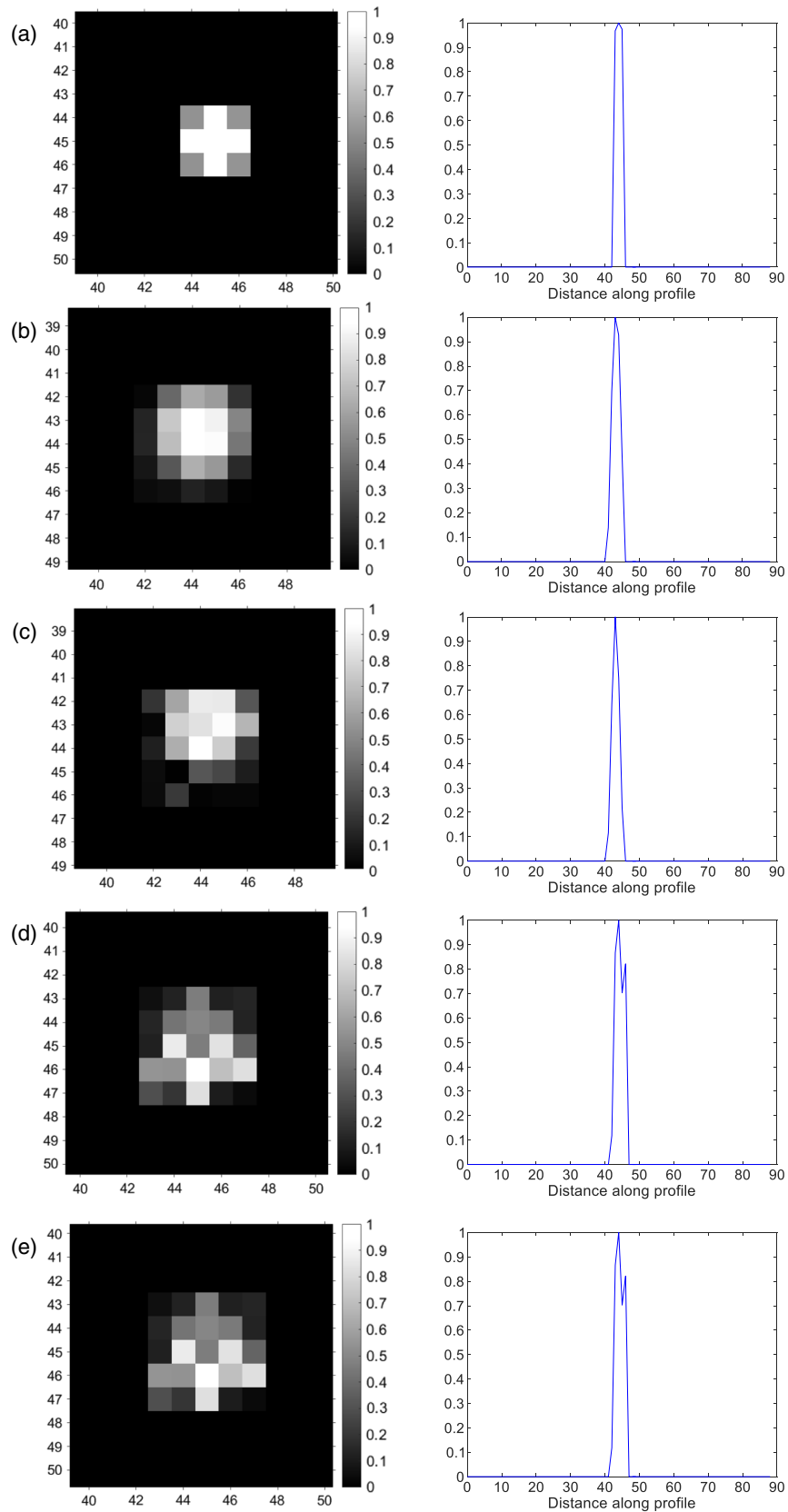
Image processing is expected to have an impact on the detection performance since it influences the visualization of the images. However, in this study, we found that using either the processed or unprocessed template did not have an effect on the AUC using the NPWE, and only a minor effect when using the CHO. This suggests that the impact of image processing on this system is small for the detection of calcification-like signals. Analyzing these images by human observers,<sup>1</sup> as done

previously, showed that for this system and this type of processing, the impact on detectability of the 0.25 mm object was small.

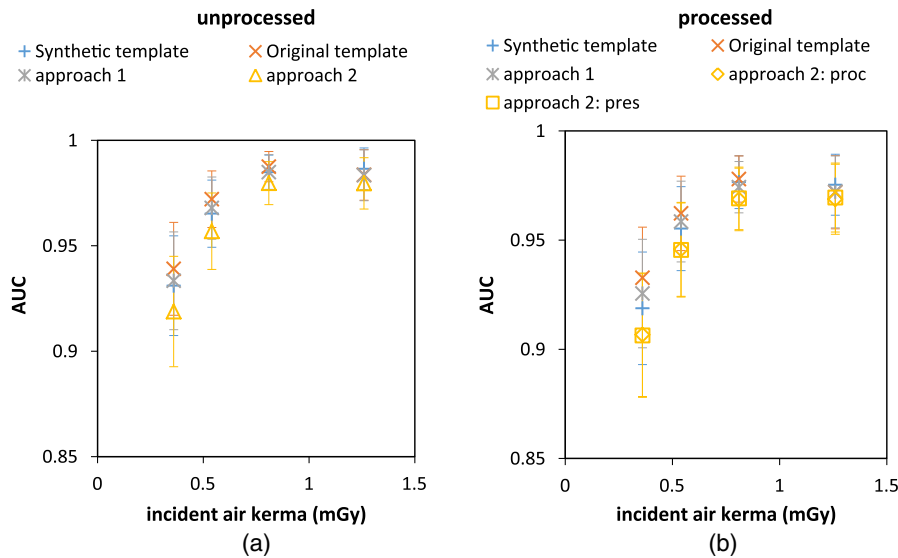
This study showed that the choices made regarding the template formulation have a minimal impact on the results of MOs. Therefore, the template construction can follow clear instructions that can be used for QC procedures. Based on these and our previous results,<sup>1</sup> we could design the new QC phantom consisting of an anthropomorphic breast phantom and a thin sheet positioned inside. This thin sheet should be designed with multiple gold disks, such that a few of them could be positioned in an area of adipose-like tissue so that they can be used to obtain the expected signal from a single exposure as shown in this work.

Regarding the number of RoIs needed for the MO experiments, the phantom choice also plays an important role. The phantom area needs to be sufficiently large to provide the MO with the number of RoIs required for training and testing. In this study, and in previous MO studies, 200 pairs of RoIs were used for each condition.<sup>1,2</sup> Based on the study of Yu et al.,<sup>19</sup> the average breast in craniocaudal view is 154 cm<sup>2</sup>. Therefore, a physical anthropomorphic phantom to be used for QC tests should have these dimensions. With such a phantom, we would be able to fit a number of RoIs per exposure, and then reposition the aluminum sheet with the gold discs and re-expose. To make the procedure suitable for QC, this number of exposures and repositions can be as low as possible by using a sufficient number of gold disks within the phantom.

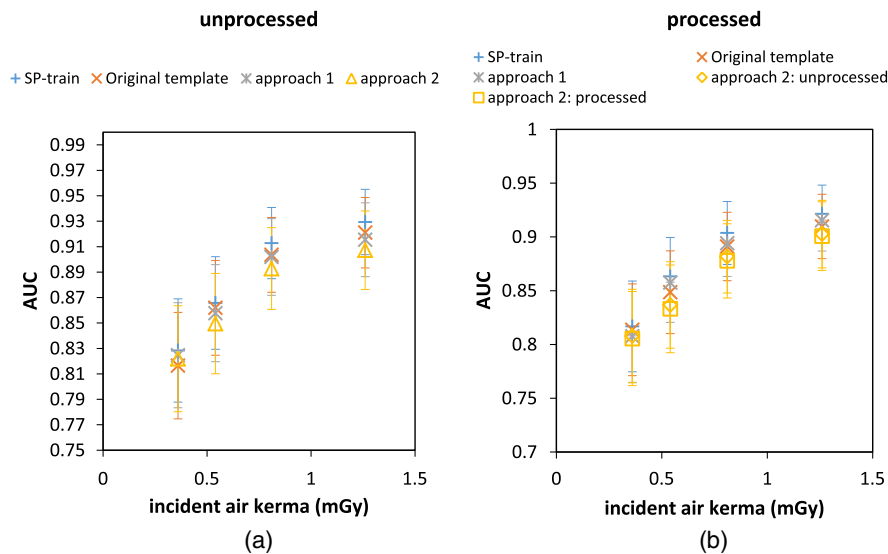
One limitation of this study is that calcification detection is just one part of the task of the radiologist in mammography. Further research should therefore also include other tasks like, for example, the detection of masses. The proposed approach to determine the acquired template needs to be re-evaluated for different signal sizes and shapes. The interpolation



**Fig. 7** Template images using (a) the synthetic template, (b) the originally proposed template, and (c)–(e) the templates resulting from approaches 1 and 2. The templates are cropped around the center and they are displayed in min–max window width. The images shown are zoomed such that each pixel can be individually evaluated.



**Fig. 8** NPWE AUC as a function of incident air kerma at the surface of the anthropomorphic breast phantom for the different template formulation for (a) unprocessed images and (b) processed images. The error bars represent 95% confidence intervals.



**Fig. 9** CHO AUC as a function of incident air kerma at the surface of the anthropomorphic breast phantom for the different template formulation for (a) unprocessed images and (b) processed images. The error bars represent 95% confidence intervals.

patch will probably need to be adjusted accordingly. For example, the circular signals used in this study were small enough to be approximated by a squared-shaped interpolation patch. However, if larger or irregular-shaped signals are going to be used, the patch shape and size should follow the geometric characteristics of the signal, taken into account the blurring of the mammography system and minimize the amount of noise found within the final template.

Another limitation is that this method was implemented only on one type of mammography system. Should another system with worse spatial resolution be tested, the size of the RoIs and the size of the template signal would be larger. This also means that the signal size in the testing image size would appear larger. For acquired templates, by definition, the template is influenced

by system characteristics (like blur). Therefore, if another system would have been used, this would affect the results but the general method described in this study would still be applicable.

Also, in the final QC method, a difference in dose should be discernable. The sensitivity of the future QC method should be taken into account in the development of the final QC procedure: number of phantom images, number of signal RoIs in a phantom, number of background RoIs, and the choice of parameters of the MO. This will be investigated in the future.

Using the next generation of phantoms, we can further evaluate the detection of calcifications with different DM systems including image processing. Eventually, this should result in a framework to assess image quality, which could form the basis for the next generation of QC guidelines.

**Table 2** The mean AUC and error bars for the different template formulations investigated using the NPWE. “Synth. template” stands for the synthetic template, “orig. template” stands for the original template and “IAK” for the incident air kerma in mGy.

	Unprocessed Images				Processed Images							
	IAK (mGy)	AUC	Min-max		IAK (mGy)	AUC	Min-max					
Synth. template	<b>0.36</b>	<b>0.939</b>	<b>0.916–0.960</b>	Synth. template	0.36	0.918	0.891–0.944					
	0.54	0.972	0.959–0.984		0.54	0.955	0.934–0.975					
	0.81	0.987	0.980–0.994		0.81	0.976	0.963–0.989					
	1.26	0.983	0.970–0.995		1.26	0.975	0.961–0.989					
Orig. template	0.36	0.939	0.917–0.962	Orig. template	0.36	0.932	0.909–0.957					
	0.54	0.972	0.959–0.985		0.54	0.962	0.945–0.980					
	0.81	0.987	0.979–0.996		0.81	0.977	0.968–0.988					
	1.26	0.983	0.971–0.997		1.26	0.972	0.954–0.988					
Approach 1	0.36	0.993	0.909–0.957	Approach 1	0.36	0.925	0.901–0.951					
	0.54	0.967	0.952–0.981		0.54	0.958	0.939–0.977					
	0.81	0.984	0.978–0.992		0.81	0.974	0.963–0.985					
	1.26	0.983	0.970–0.994		1.26	0.972	0.958–0.987					
Approach 2	0.36	0.918	0.890–0.945	Approach 2: proc	0.36	0.906	0.878–0.935					
				0.54	0.956	0.938–0.974	0.54	0.945	0.922–0.969			
							0.81	0.979	0.969–0.990	0.81	0.968	0.956–0.982
										1.26	0.979	0.968–0.991
	Approach 2: pres	0.36	0.906									
	0.54	0.945	0.923–0.967									
	0.81	0.969	0.953–0.982									
	1.26	0.969	0.952–0.983									

**Table 3** The mean AUC and error bars for the different template formulations investigated using the CHO. “Orig. template” stands for the original template and “IAK” for the incident air kerma in mGy.

	Unprocessed images				Processed images		
	IAK (mGy)	AUC	min-max		IAK (mGy)	AUC	min-max
SP-train	0.36	0.828	0.787–0.869	SP-train	0.36	0.816	0.772–0.860
	0.54	0.865	0.831–0.903		0.54	0.857	0.822–0.892
	0.81	0.912	0.882–0.942		0.81	0.904	0.872–0.936
	1.26	0.929	0.903–0.958		1.26	0.921	0.892–0.950



Table 3 (Continued).

	Unprocessed images				Processed images		
	IAK (mGy)	AUC	min-max		IAK (mGy)	AUC	min-max
Orig. template	0.36	0.817	0.772–0.859	Orig. template	0.36	0.813	0.769–0.857
	0.54	0.862	0.821–0.901		0.54	0.848	0.809–0.887
	0.81	0.904	0.872–0.932		0.81	0.891	0.858–0.924
	1.26	0.921	0.896–0.941		1.26	0.909	0.878–0.940
Approach 1	0.36	0.824	0.781–0.868	Approach 1	0.36	0.807	0.762–0.852
	0.54	0.857	0.821–0.898		0.54	0.857	0.819–0.895
	0.81	0.902	0.873–0.932		0.81	0.894	0.861–0.927
	1.26	0.915	0.889–0.947		1.26	0.915	0.888–0.942
Approach 2	0.36	0.821	0.779–0.861	Approach 2: proc	0.36	0.807	0.763–0.851
	0.54	0.849	0.810–0.890		0.54	0.836	0.792–0.880
	0.81	0.892	0.859–0.922	0.81	0.882	0.849–0.915	
	1.26	0.907	0.874–0.934	1.26	0.903	0.868–0.938	
				Approach 2: pres	0.36	0.806	0.761–0.851
					0.54	0.833	0.789–0.877
					0.81	0.878	0.842–0.914
					1.26	0.900	0.869–0.931

## 5 Conclusion

We optimized a previously proposed procedure to derive a signal template from acquired images for the NPWE MO and the CHO. We found that a straightforward procedure based on a single exposure of a single disk was sufficient to generate an appropriate signal template. Based on the findings of this study, the next step will be to design an anthropomorphic phantom with a reference disk positioned in a fully adipose-like region so that the expected signal template can be derived directly from the acquired images. Using this phantom, further studies should be performed to propose procedures for the actual image quality evaluation with MOs using an anthropomorphic phantom.

### Disclosures

This research manuscript is an extension and revision of the work presented at the SPIE 2017 Medical Imaging conference.<sup>13</sup> The authors have no relevant conflicts of interest to declare.

### Acknowledgments

The authors would like to acknowledge Mr. Stephan Schopphoven, Prof. Dr. Martin Fiebich, and Mr. Ulf Maeder of the Institute of Medical Physics and Radiation Protection from the Technische Hochschule Mittelhessen for allowing us to use the prototype anthropomorphic breast phantom and Artinis Medical systems to design the sheet with gold disks.

Both were developed as part of the CLUES project, which was funded by the Dutch Technology Foundation, STW (Grant No. 13592).

### References

1. C. Balta et al., "A model observer study using acquired mammographic images of an anthropomorphic breast phantom," *Med. Phys.* **45**(2), 655–665 (2018).
2. R. W. Bouwman et al., "Toward image quality assessment in mammography using model observers: detection of a calcification-like object," *Med. Phys.* **44**(11), 5726–5739 (2017).
3. I. Hernandez-Giron et al., "Automated assessment of low contrast sensitivity for CT systems using a model observer," *Med. Phys.* **38**(S1), S25–S35 (2011).
4. L. C. Ikejimba et al., "Assessing task performance in FFD, DBT, and synthetic mammography using uniform and anthropomorphic physical phantoms," *Med. Phys.* **43**(10), 5593–5602 (2016).
5. M. Reginatto, M. Anton, and C. Elster, "Assessment of CT image quality using a Bayesian approach," *Metrologia* **54**(4), S74–S82 (2017).
6. F. R. Verdun et al., "Image quality in CT: from physical measurements to model observers," *Phys. Med.* **31**(8), 823–843 (2015).
7. D. Georgian-Smith, *Breast Imaging and Pathologic Correlations: A Pattern-Based Approach*, Wolters Kluwer Health, Philadelphia, Pennsylvania (2015).
8. L. M. Warren et al., "Effect of image quality on calcification detection in digital mammography," *Med. Phys.* **39**(6), 3202–3213 (2012).
9. R. Visser et al., "Increase in perceived case suspiciousness due to local contrast optimisation in digital screening mammography," *Eur. Radiol.* **22**(4), 908–914 (2012).

10. F. Zanca et al., "Evaluation of clinical image processing algorithms used in digital mammography," *Med. Phys.* **36**(3), 765–775 (2009).
11. L. C. Ikejimba et al., "A novel physical anthropomorphic breast phantom for 2D and 3D x-ray imaging," *Med. Phys.* **44**, 407–416 (2016).
12. L. Cockmartin et al., "Comparison of digital breast tomosynthesis and 2D digital mammography using a hybrid performance test," *Phys. Med. Biol.* **60**(10), 3939–3958 (2015).
13. C. Balta et al., "Signal template generation from acquired mammographic images for the non-prewhitening model observer with eye-filter," *Proc. SPIE* **10136**, 101360M (2017).
14. B. Gallas, "One-shot estimate of MRMC variance: AUC," *Acad. Radiol.* **13**(3), 353–362 (2006).
15. A. E. Burgess, "Statistically defined backgrounds: performance of a modified nonprewhitening observer model," *J. Opt. Soc. Am. A* **11**(4), 1237 (1994).
16. R. W. Bouwman et al., "Can the non-pre-whitening model observer, including aspects of the human visual system, predict human observer performance in mammography?" *Phys. Med.* **32**(12), 1559–1569 (2016).
17. C. K. Abbey and H. H. Barrett, "Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability," *J. Opt. Soc. Am. A* **18**(3), 473–488 (2001).
18. L. Platiša et al., "Channelized Hotelling observers for the assessment of volumetric imaging data sets," *J. Opt. Soc. Am. A. Opt. Image Sci. Vision* **28**(6), 1145–1163 (2011).
19. L. Yu et al., "Prediction of human observer performance in a 2-alternative forced choice low-contrast detection task using channelized Hotelling observer: impact of radiation dose and reconstruction algorithms," *Med. Phys.* **40**(4), 41908 (2013).
20. I. Diaz et al., "Derivation of an observer model adapted to irregular signals based on convolution channels," *IEEE Trans. Med. Imaging* **34**, 1428–1435 (2015).
21. D. Racine et al., "Objective assessment of low contrast detectability in computed tomography with Channelized Hotelling Observer," *Phys. Med.* **32**, 76–83 (2016).
22. R. M. Gagne, B. D. Gallas, and K. J. Myers, "Toward objective and quantitative evaluation of imaging systems using images of phantoms," *Med. Phys.* **33**, 83–95 (2006).
23. R. J. Acciavatti and A. D. A. Maidment, "Investigating the potential for super-resolution in digital breast tomosynthesis," *Proc. SPIE* **7961**, 79615K (2011).

**Christiana Balta** has background in physics and biomedical engineering and she is a PhD student at the Dutch Expert Centre for Screening (LRCB) and at the Department of Radiology and Nuclear Medicine of the Radboud University Medical Center in Nijmegen, The Netherlands.

**Ramona W. Bouwman** is a physicist with 12 years of experience in quality control in mammography, senior researcher at the Dutch Expert Centre for Screening (LRCB) and radiation protection expert.

**Wouter J. H. Veldkamp** is a medical physicist and associate professor at Leiden University Medical Centre in Leiden, The Netherlands. He is the project leader of the CLUES project, which aims to build the gaps between physicotchnical measurements and image quality in mammography and computed tomography.

**Mireille J. M. Broeders** is an epidemiologist and associate professor at the Department of Health Evidence at the Radboud University of Nijmegen. She is also the scientific director of the Dutch Expert Centre for Screening (LRCB).

**Ioannis Sechopoulos** is a medical physicist and associate professor at the Department of Radiology and Nuclear Medicine of the Radboud University Medical Center in Nijmegen. He is the director of the Advance X-ray Tomography Imaging (AXTI) lab. And he is a researcher at the Dutch Expert Centre for Screening (LRCB).

**Ruben E. van Engen** is a physicist with 20 years of experience in quality control in mammography, senior researcher at the Dutch Expert Centre for Screening (LRCB) and radiation protection expert. He is the first author of the European Guidelines for quality control in screening mammography.