Physica Medica 58 (2019) 8–20

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Systematic approach to a channelized Hotelling model observer implementation for a physical phantom containing mass-like lesions: Application to digital breast tomosynthesis

Dimitar Petrov[a,*], Nicholas W. Marshall[a,b], Kenneth C. Young[c], Hilde Bosmans[a,b]

[a] Dept. of Medical Physics and Quality Assessment, KU Leuven, Leuven, Belgium
[b] Dept. of Radiology, UZ Leuven, Belgium
[c] NCCPM, Royal Surrey County Hospital, Guildford, UK

ARTICLE INFO

Keywords:
Channelized Hotelling observer
Digital breast tomosynthesis
Physical phantom
Human observer
Mass lesions

ABSTRACT

Purpose: to develop a channelized model observer (CHO) that matches human reader (HR) scoring of a physical phantom containing breast simulating structure and mass lesion-like targets for use in quality control of digital breast tomosynthesis (DBT) imaging systems.

Methods: A total of 108 DBT scans of the phantom was acquired using a Siemens Inspiration DBT system. The detectability of mass-like targets was evaluated by human readers using a 4-alternative forced choice (4-AFC) method. The percentage correct (PC) values were then used as the benchmark for CHO tuning, again using a 4-AFC method. Three different channel functions were considered: Gabor, Laguerre-Gauss and Difference of Gaussian. With regard to the observer template, various methods for generating the expected signal were studied along with the influence of the number of training images used to form the covariance matrix for the observer template. Impact of bias in the training process on the observer template was evaluated next, as well as HR and CHO reproducibility.

Results: HR performance was most closely matched by 8 Gabor channels with tuned phase, orientation and frequency, using an observer template generated from training image data. Just 24 DBT image stacks were required to give robust CHO performance with 0% bias, although a bias of up to 33% in the training images also gave acceptable performance. CHO and HR reproducibility were similar (on average 3.2 PC versus 3.4 PC).

Conclusions: The CHO algorithm developed matches human reader performance and is therefore a promising candidate for automated readout of phantom studies.

1. Introduction

Breast screening employing conventional 2D full field digital mammography (FFDM) plays a key role in early breast cancer detection and, when implemented carefully, is known to reduce breast-cancer mortality [1,2]. A limitation associated with FFDM is the projection of all breast structures into a single image. This can obscure lesions and therefore reduce sensitivity or increase the number of false alarms [3,4]. Digital breast tomosynthesis (DBT) systems acquire a series of projection images using an x-ray source that moves over a limited angle. The resulting projection data are used to reconstruct a set of planes parallel to the detector [5–7], generating a 3D dataset of the breast anatomy. Recent studies have shown that the addition of DBT to digital mammography or even the stand-alone use of DBT can decrease recall rate and increase lesion detection compared to standard FFDM [8,9]. Digital breast tomosynthesis is therefore a modality that shows great potential [10,11].

Physical phantoms are an established means of quantifying a (technical) measure of image quality and ensure that image quality meets some sufficient level while remaining within accepted dose levels [12]. It is likely that the practice of physical phantom evaluation, initially scored by human observers, will be carried over to the assessment and optimization of DBT devices. One such candidate is the 3D structured phantom developed by Cockmartin et al. [13], which contains clinically relevant targets and was developed for comparative studies of FFDM and DBT. Future applications may include the use in routine QC and in technical optimization of DBT systems. Currently, phantom images are read by human readers (HR). This is a time consuming approach, and furthermore the same observers may not always be available to score the images of successive investigations.

* Corresponding author.
E-mail address: dimitarbp@gmail.com (D. Petrov).

https://doi.org/10.1016/j.ejmp.2018.12.033
Received 21 June 2018; Received in revised form 5 December 2018; Accepted 25 December 2018
Available online 17 January 2019
1120-1797/ © 2018 Published by Elsevier Ltd on behalf of Associazione Italiana di Fisica Medica.

Computerized evaluations of physical phantoms may boost the development and application of QC protocols for DBT systems, in line with the European approach [14].

While evaluation of technical parameters that influence imaging performance yields useful information [12,15–17], some of these methods involve the use of Fourier techniques, which require system linearity and therefore must be applied with care to reconstructed image datasets. As an alternative, many studies have described the potential of spatial domain model observers (MO) [18–20] to perform these detection tasks for a variety of medical imaging modalities, including CT [21], FFDM [22] and fluoroscopy [23]. More specifically, the channelized Hotelling observer (CHO) has been proven to be a useful classifier in DBT images [24–30]. For example Young et al. [25] used a CHO to quantify different DBT setups with varying scan angles and number of projections. Zeng et al. [27] use three types of CHO to compare two DBT reconstruction methods in different angular span and number of views, which results were then verified by a human observer study. Wen et al. [28] compared CHO model observers for identifying the optimal DBT acquisition geometries. Most of these studies focus on optimization via simulation with mathematical phantoms [14]. Turning to physical phantoms, Park et al. [30] applied a Hotelling observer to assess the detectability of spheres while Michielsen et al. [29] used a CHO in the evaluation of iterative algorithms to reconstruct microcalcifications in a physical DBT phantom.

CHO methods typically involve three distinct steps: design and tuning of a channel set relevant to the targets/backgrounds, training of the observer template and finally application. This last step generates a scalar test statistic, characterizing the response of the CHO to a given image. Evaluating this test statistic at various thresholds generates a performance metric such as the area under the curve (AUC). Channel selection influences CHO performance and the degree to which HR performance is approximated. Castella et al. [31], and Bouwman et al. [32] showed that some channels provide better anthropomorphic matching than others, while Zeng et al. [27] showed that channel parameter tuning can improve agreement with HR scores. There is currently little consensus on the selection or development of CHOs for physical test objects. The goal of this study was therefore to describe the systematic development and validation of a CHO algorithm for use with the physical phantom described by Cockmartin [13], with the focus on channel design, the generation of signal templates and the number of images required for a robust estimate of the covariance matrix.

The channel design incorporates studies on the channel types and investigating appropriate channel parameters used to approximate human observer results. The CHO signal template sets the target to be detected and a proper estimation is crucial for the performance. The study will compare how different estimations of the signal template influence the CHO scores. For practical considerations the number of DBT phantom acquisitions needed for assessment of a CHO reading is crucial for the implementation in daily practice. The CHO will be trained with different number of DBT scans for training, and the minimal amount of training DBT acquisitions will be investigated. In the same context, the amount of training acquisitions can be lowered, if some of the images for reading are used also for training. This will clearly add observer bias, which could influence the results. The CHO will be trained with datasets inducing different amounts of bias, to test whether a small percent of bias can be allowed, when the scanning time is a limit. Finally the developed model observer will be tested for reproducibility against human results and at three dose levels.

2. Materials and methods

2.1. Image acquisition

This study is based around the DBT test object described by Cockmartin et al. [13] (Fig. 1), a 3D physical phantom that uses spheres to generate the backgrounds for the detection study. Readers are referred to that text [13] for a detailed discussion of the advantages and limitations of using spheres for this purpose. Briefly, the phantom is made from a poly(methyl) methacrylate (PMMA) semi-circular container filled with PMMA spheres of six different diameters (15.88, 12.70, 9.52, 6.35, 3.18, and 1.58 mm), with water filling the remaining volume (Fig. 1). The spheres are free to move and thus shaking the phantom produces another background realization, but with similar power spectra characteristics. This study evaluated the non-spiculated masses, with average diameters equal to 1.5, 2.1, 3.0, 4.3 and 5.7 mm, with a fixed position within the phantom volume. Each phantom scan generates a single set of signal present data, while the same scan yields 15 signal absent datasets.

The phantom was scanned 108 times in total on a Siemens Inspiration Tomosynthesis system (Siemens-Healthineers, Erlangen, Germany). Acquisition parameters were related to settings under automatic exposure control (AEC), namely 30 kV, W/Rh anode/filter combination and 204 mAs. Sixty acquisitions were taken at the AEC dose level with manually selected factors (30kVp and 200mAs) and 24 acquisitions were then taken at a 'Low' dose level, close to half the standard dose (30kVp and 112mAs) and 24 scans at a 'High' dose level (30kVp and 400mAs). DBT volumes were reconstructed using the "Enhanced Multiple Parameter Iterative Reconstruction (EMPIRE)" algorithm. The reconstruction algorithm is a filtered back projection (FBP) with additional iterative processing for artefact reduction, noise reduction and higher resolution [33,34]. Between each scan, the phantom was shaken for 10 s to generate a different background realization.

Once the scans were acquired then the VOIs containing the lesions were located as follows. Sets of small high contrast, microcalcification simulating lesions are also present within the phantom and these were used as the localization landmark. DBT scans were made of the phantom with the lesion models, but without the background spheres or water, and distances to the five non-spiculated masses were measured from the calcification reference point. All the lesions are located at fixed positions on a 2 mm thick PMMA sheet within the phantom and therefore do not move during the shaking. Knowing the location of the microcalcifications enables the position of the non-spiculated masses to be calculated and VOIs extracted at the appropriate locations (lesions are at the centre of the green squares in Fig. 2). As a check, the extraction method was applied to a dataset of the empty phantom (no spheres) and a visual check made that the extracted masses were at the centre of the relevant VOI. The signal present VOIs were formed from a cropped volume of size $20 \times 20 \times 30\ mm^3$ centred around the target mass. The physical height of the phantom is 48 mm including the phantom walls [13], generating ~ 40 DBT slices between the top and bottom PMMA plates. Five slices at the top and bottom were not used due to the influence of the PMMA plates on the background, giving VOIs with 30 planes that have similar background statistics. This also meant that the signal absent VOIs had to be extracted at the same z-height as the signal present VOIs, from lesion free areas of the phantom (Fig. 2.a).

2.2. Human observer study

Six medical physicists participated in a four alternative forced choice (4-AFC) image reading study [29]. For detection tasks (as considered here), Jäkel and Wichmann [29] suggest that 4-AFC and 8-AFC methods are more time efficient and have lower uncertainty (after some given number of trials) than a 2-AFC method. However, an 8-AFC experiment requires more signal absent images than a 4-AFC study and therefore we consider the 4-AFC study used here to be a good compromise. Three signal absent VOIs and one signal present VOI were shown, with the task of indicating the signal present image. A software tool developed in-house ('Foursquares' [35]) (Fig. 2.b)) was used to conduct the 4AFC study.

Reading was performed on a diagnostic monitor (5MP Barco MDNG-

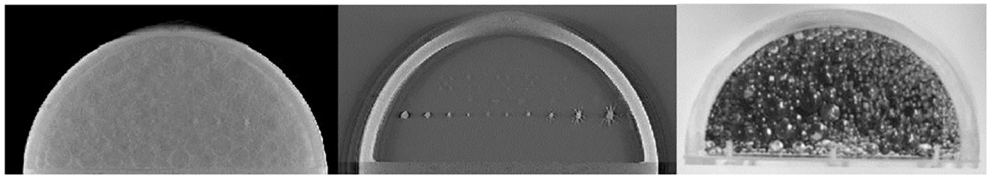

Fig. 1. Images of the phantom, from left to right: DBT reconstructed plane at the level of the mass models, mammographic image without the background spheres and a photograph.

6121) at an ambient light level of 3 lx and viewing distance of 40–50 cm. No magnification and window level adjustments were allowed, and only scrolling through the VOI was permitted. No time constraints were imposed for the reading sessions. While it is common for observers to read a training set before participating in this type of study, this was not done in this study. All the observers had read many DBT images of this phantom, acquired under different conditions for a different study, and thus a training set was not required. The readers were therefore experienced in this type of perception experiment. One reading session (of 12 images/trials) typically took 5 min. During reading, the observers were shown an additional ROI (similar to that in Fig. 5.d.) containing the specific target to be detected.

The 60 DBT acquisitions taken at AEC dose level were split into 5 groups of 12 DBT stacks. A further 2 groups were formed for the high and low dose level acquisitions, each with 12 DBT stacks. Volumes of interest were extracted and subsequently sorted into 5 reading sessions, corresponding to the 5 lesion sizes for each reading group. This way a reading session consisted of twelve 4AFC trials, which gave one percentage correct (PC) result for each human reader. Given the 3 dose levels and the 5 lesion sizes, each observer read 35 sessions in total. The overall PC for a given lesion size and dose level was found by averaging over PC for all 6 readers, with uncertainty taken as the standard error on the mean.

2.3. Channelized Hotelling observer and general work flow

The Hotelling observer computes a test statistic t for each image g, by applying an observer template [36]:

$$t(g) = w^T g = \Delta\bar{g}^T S_g^{-1} g, \tag{1}$$

where $\{\}^T$ is the vector transpose operator and w is the observer template, formed from the mean difference between the signal present and signal absent data ($\Delta\bar{g}$) and the inverse of the interclass covariance matrix of the signal present and signal absent data (S_g^{-1}).

In practice, due to the large dimensions of the covariance matrix, inversion is often impossible. To overcome this dimensionality problem, a channel mechanism was introduced in the definition of the template [37], resulting in the channelized Hoteling observer (CHO):

$$t_{CHO}(v) = w_{CHO}^T v = (\bar{v}_{sp} - \bar{v}_{sa})^T S_v^{-1} v, \tag{2}$$

where $v = U^T g$ is the channel output vector, formed by the product of the channel set U and image g. The work flow of application of the CHOs is split into three interlinked phases: 'channel tuning', 'training' and 'reading' (Fig. 3). In the tuning phase, the channels are generated and associated parameters adjusted using a set of training images to find a CHO whose performance matches the human observer performance. During training, the observer template is usually estimated from a set of training images [36]. In the last phase, called 'reading', the observer template is applied to the images that have to be evaluated.

The 4-AFC method used to estimate human observer performance was also implemented for the CHO. Decision variables for the 4 images in a given 4-AFC trial were calculated and if the decision variable for the signal present image had the highest value, the algorithm counted a 'hit' (Fig. 3). This action was repeated for all signal present images in a reading session, where each time the algorithm picks 3 random signal absent decision variables with replacement. The PC is then calculated from the total number of 'hits'. This was repeated six times for the same reading session, matching the number of human readers and the average PC was taken as a final result for the session. The uncertainty of it was estimated via bootstrapping. The resampling process was repeated 30 times, and the standard deviation was taken as an uncertainty measure. This work implemented a single slice CHO (ssCHO) in the terminology of Platisa et al. [24], who showed that ssCHO and multi-slice CHO had similar performance for clustered lumpy background (CLB) images. As Bochud et al. [38] and Castella et al. [39] have shown, CLB images have similar appearance to real mammographic backgrounds. The CHO implementation requires signal known exactly conditions and given that perfect segmentation cannot be guaranteed, the CHO was applied to 25 positions around the expected signal position. This action was also performed over 5 planes in the through plane (z)-direction, which resulted in a 125 voxel scanning area around the initial (expected) position of the lesion (i.e. or $4.25 \times 4.25 \times 5\,\text{mm}^3$ real space volume). After calculating the CHO performance for all points, the maximum of this data was taken as the final result

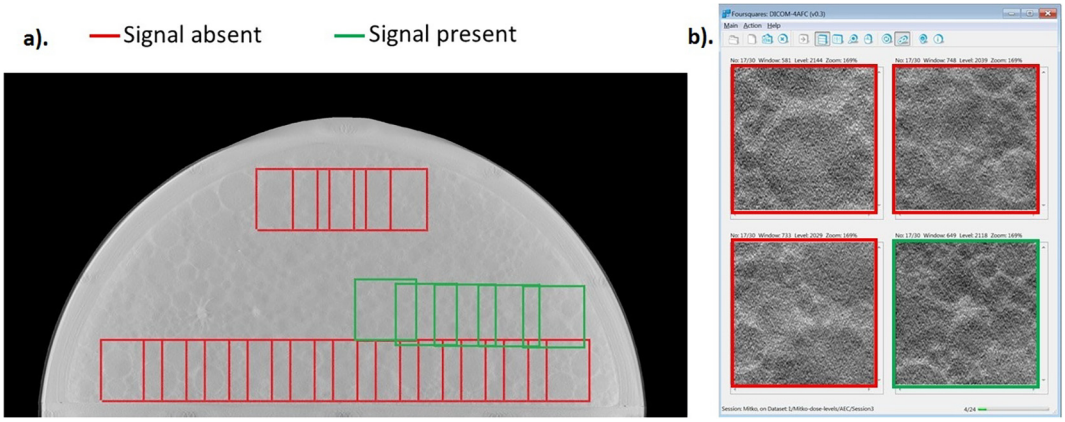

Fig. 2. a). Extraction positions for signal present and signal absent volumes of interest. b). Screenshot of the in-house developed 'Foursquares' 4AFC software [35]. For illustration purposes the signal present image is marked in green and the signal absent images – in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

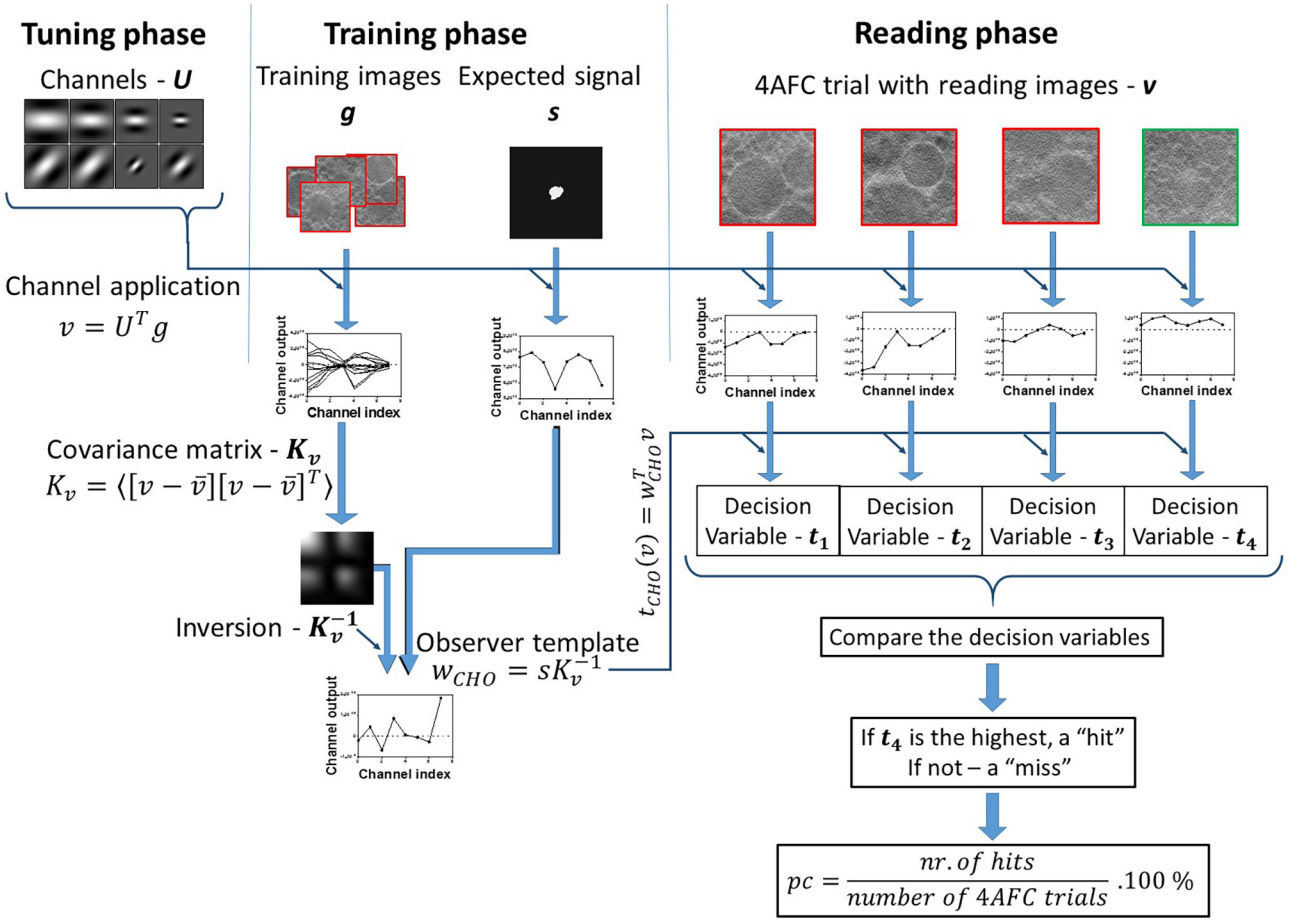

Fig. 3. Flow chart of channelized Hotelling model observer performance assessment in 4AFC test.

2.4. Channel selection and tuning

The first step is selection of the type of channel for the CHO as this obviously has a large influence on MO performance. The channels can be viewed as a type of filter applied to the images that extract features relevant to task (e.g. detection) performed using of the targets known to be present in the images. These channels have to be adjusted (tuned) so that they match the HR performance obtained using the phantom. Two types of channels are commonly used [40]: anthropomorphic and efficient. Anthropomorphic channels are used to incorporate some characteristics of the human visual system within the model observer, while efficient channels are used to approximate the ideal observer performance. Three of the most common types of channels were studied in this work: Gabor, Difference of Gaussians (DOG) and Laguerre-Gauss (LG), where the first two are considered anthropomorphic and the third is considered efficient. Castella et al. [31] observed that using higher numbers of channels may lead to an increased overall observer performance, but not necessarily to a better matching of human results. For our study, the channel number was set to 8. This was based on some earlier tests (data available, but not shown) in which it was found that 8 channels represented stable results and a workable compromise in terms of image acquisition work load and required number of training images.

The Gabor function is defined in the spatial domain by multiplying a sinusoidal wave by a Gaussian function (Fig. 4.1) [41]:

$$C_{i,j,k}(x, y) = e^{\frac{-4\ln(2)}{w_i^2}(x^2+y^2)} \cos[2\pi f(x\cos\theta_j + y\sin\theta_j) + \phi_k], \quad (3)$$

where $w_i = \frac{t_w}{(e^i+2)pxs}$, $f = \frac{t_f}{W_i}$, $\theta_j = \frac{\pi j}{t_t}$ and $\phi_k = 45k$.

The Gaussian function determines the width (w_i) of the channels and the sinusoidal function – their frequency (f), orientation (θ) and phase (ϕ). Here, the frequency was set as a function of the standard deviation of the Gaussian function, guaranteeing that only one maximum of the sinusoidal wave has an impact and therefore forcing the sensitive region of the CHO to the center of the image. Three parameters are required to generate a channel set: the number of frequencies (I), orientations (J) and phases (K). Thus, in order to generate 8 channels, one phase, two orientations and 4 frequencies were selected. To set the properties of each channel within the channel set, three tuning factors were implemented: t_w, t_f and t_t. These set the Gaussian standard deviation, the sine wave frequency and the orientation. During tuning, the parameters were varied as follows: t_w and t_f ranged from 5 to 100 and t_t was varied such that θ_1 varied from 10° to 180°.

DOG channels are radially symmetric channel sensitivity functions, formed by subtracting two Gaussian functions with different standard deviations (Fig. 4.2). They are defined as a function of radial frequency (pixels^{-1}) [20]:

$$C_j(\rho) = exp\left[-\frac{1}{2}\left(\frac{\rho}{Q\sigma_j}\right)^2\right] - exp\left[-\frac{1}{2}\left(\frac{\rho}{\sigma_j}\right)^2\right] \quad (4)$$

with $\sigma_j = \sigma_0\alpha^j$, the standard deviation of each channel. After generation in frequency space, an inverse Fourier transform is applied, yielding a real space set of channels. The first 8 DOG channels are generated by varying j from 0 to 7 to produce the channel set. There are three channel parameters: Q defines channel bandwidth, σ_0 is standard deviation of the first channel and α determines the difference in standard deviation between the channels. During tuning, σ_0 ranged from 0.002 to 0.02 while α and Q were varied from 1.1 to 3.0.

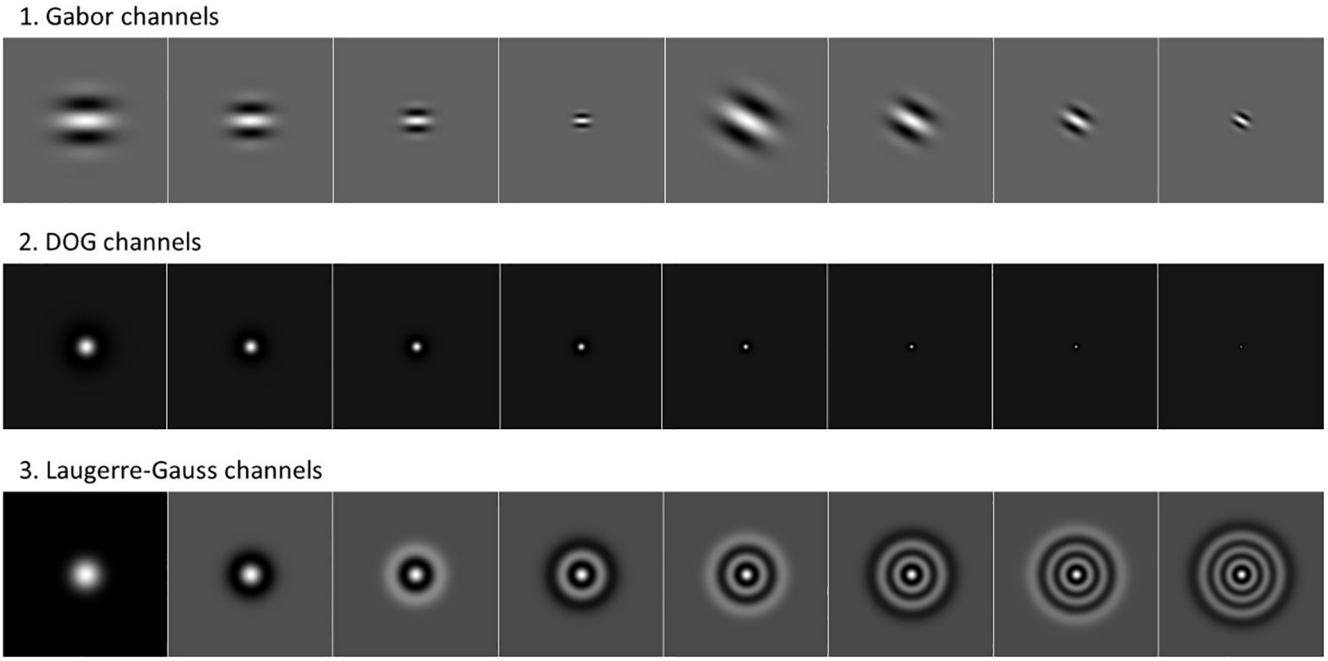

Fig. 4. Images of the studied channel types.

As with DOG channels, LG channels are rotationally symmetric, and are formed in the spatial domain as the product of Laguerre polynomials and Gaussian functions [42] (Fig. 4.3):

$$C_j(r) = \frac{\sqrt{2}}{a_u}\exp\left(\frac{-\pi r^2}{a_u^2}\right)L_j\left(\frac{2\pi r^2}{a_u^2}\right) \quad (5)$$

where $a_u = \sqrt{2\pi}\sigma_u$ i.e. the standard deviation of the Gaussian. L_j is the Laguerre polynomial, defined as:

$$L_j(x) = \sum_{k=0}^{j}(-1)^k\binom{j}{k}\frac{x^k}{k!} \quad (6)$$

For this channel type, only the standard deviation factor σ_u was tuned for the purpose of matching observer performance, with values ranging from 3 to 100.

First, rigorous tuning was performed against the human observer results at the AEC dose level images. For the three channel types, each tuning parameter was varied while the rest were held constant. The PC for the CHO was calculated for all five mass lesion models and compared against those for the averaged human reader data by applying the evaluation criteria in Table 1. The three evaluation results generated by each set of channel parameters (for a given channel type) were tabulated and the best performing parameter sets were initially selected by requiring the absolute ME to be lower than 5 PC, a linear slope (a) between 0.9 and 1.1 and a correlation (r) greater than 0.9. Out of this initial selection, the channel parameter set that had the best overall evaluation score for all three test scores were then selected for that channel type and used for the further studies. In so doing, we systematically covered a wide range of channel settings, however it is recognized that not all combinations in the parameter space could be tested.

2.5. CHO training

Training the CHO requires the definition of the (expected) signal in the form of a template that is applied to the images and this section examines how different signal templates influence the CHO scores. The training phase also builds the covariance matrix and therefore we also examine the number of DBT phantom acquisitions needed for the implementation of a robust and practically achievable CHO. In this context, the number of training acquisitions can be reduced if some of the images used in the reading stage are used also for CHO training. This adds bias to CHO which could influence the results and therefore this section also examines CHO training with datasets that have a range of bias levels.

The CHO template is given by the following formula [42]:

$$w_{CHO} = (\bar{v}_{sp} - \bar{v}_{sa})^T K_v^{-1} = s^T K_v^{-1} \quad (8)$$

This can be split into two parts: (1) the subtraction of mean signal present and mean signal absent channel output vectors and (2) the covariance matrix of the ensembles. Both parts are determined in the training phase.

The standard way of including the expected signal is from image

Table 1
Criteria used to assess CHO performance.

Criterion	Description
Mean error (ME)	Measure of the distance between the MO scores and HR scores. If the ME value is positive the MO underestimates the human results, if negative, the MO overestimates the human observer scores. A value closer to 0 is desired. $ME = \frac{\sum_{i=1}^{n} y_i - x_i}{n}$ (7) where y_i and $x_i$are the HR and MO scores respectively and n is lesion size
Linear regression slope (a)	Object size dependency of the MO versus that of the human observer. If lower than 1, the effect of size is less pronounced for the MO than for human observers and vice versa. A value closer to 1 for the linear regression slope is desired for the MO.
Pearson correlation coefficient (r)	Linear relation between the two observers, where the better observer has an r value closer to 1.

Table 2
Methods used for template formation.

Template method	Figure
Signal estimated from training images. Expected signal for a given lesion diameter estimated from 144 signal present VOIs, using 3 central (adjacent) planes.	5a
Central slice of the mass model, taken from the binary 3D printing stereo lithography (STL) file of the mass models, scaled and rotated to match the physical mass model in the phantom	5b
Maximum intensity projection (MIP) of the 3D mass model (STL file) in z-direction.	5c
DBT scan of just the physical mass models (i.e. no spheres or water): template is the central plane of the reconstructed mass model	5d
Gaussian blob with FWHM set to match the average mass diameter measured in images of the phantom with no spheres, i.e. background free DBT acquisitions.	5e
Landolt C [43]. Outer diameter set to insert diameter for a given lesion.	5f
Landolt C [43]. Inner diameter set to insert diameter for a given lesion.	5g
Rectangle with height equal to lesion diameter.	5h

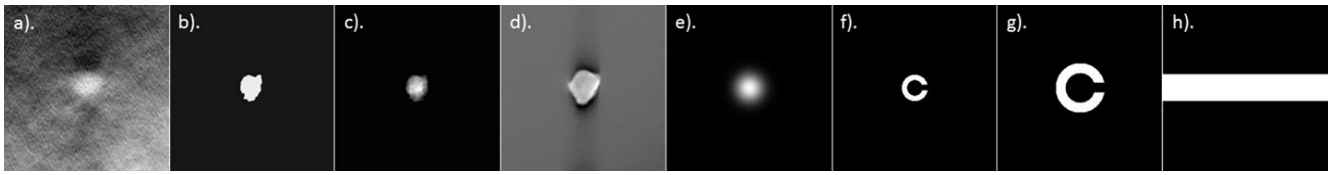

Fig. 5. Images of the studied 'expected signal' models: a). signal estimated from training images; b). Central slice of the mass model; c). Maximum Intensity Projection of the mass model in z-direction; d). DBT scan of the physical mass model; e). Gaussian blob; f). Small Landolt C; g). Large Landolt C; h). Rectangle.

sets with known truth, often acquired in low noise (high dose) conditions [36]. Nevertheless the signal template can also be provided in other ways [29,30] and therefore the influence of signal template on CHO performance was explored for seven alternative templates, listed in Table 2 and illustrated in Fig. 5. Templates f), g) and h) are only weakly related to the expected signal content but were included to assess the sensitivity of the CHO to the template choice (variation in PC and scale values) and potential template mismatches. Note that all are 2D templates [24]. Again, the criteria in Table 1 were applied and the expected signal estimation method with the highest score were used in further studies.

The covariance matrix in the template characterizes correlations between the channel output values, with the diagonal giving the variances of the channel output values. If the channel output vector v is multiplied by $K_v^{-1/2}$, where K_v has already been estimated from vectors with the same statistics as v, the covariance matrix of their product is the identity matrix, i.e. $cov(y) = cov\left(K_v^{-\frac{1}{2}} v\right) = \mathrm{I}$. This means that the fluctuations within *y* have only white noise properties. In the case of the CHO, the pre-whitening is needed for both expected signal *s* and the channel output vector *v*, as both their backgrounds are strongly correlated. Using a CHO in correlated backgrounds requires two different estimates of the covariance matrix – one for signal present and one for signal absent [36]. For this study, this was not feasible due to the limited number of signal present images that can be derived from a single DBT acquisition and therefore the covariance matrix was only estimated from the signal absent training images.

The number of elements in the CHO covariance matrix is equal to the number of channels squared. This is crucial for the inversion, as the inverse does not exist unless the number of training channel output vectors exceeds the number of vector elements, i.e. channels. For CHO implementations, this is easily achievable but stable results require more than just the bare minimum and therefore the number for a robust estimate was investigated.

To form the observer template, signal present training images were used to estimate the expected signal present channel output vector, while signal absent training images were used to form both the expected signal absent channel output vector and the covariance matrix. In order to estimate the minimum number of training images needed to derive CHO observer results similar to the HR ground truth, the observer performance was studied in different conditions of training, with images extracted from 2 to 48 DBT acquisitions. From each acquisition, 3 signal present training images for each mass model diameter (the central 3 slices from a signal present VOI) and 75 signal (the central 5 slices at 15 positions) absent training images were extracted. Training with the maximum available number of images was defined as the ground truth (48 DBT acquisitions giving 144 signal present and 2160 signal absent images).

The lower limit for training images was defined as the minimum number of DBT scans needed to achieve CHO performance within the 95% confidence interval (CI) of the ground truth PC results. The number of DBT acquisitions required was then averaged over the 5 lesion sizes and this number of training images was used to form the observer template for the following studies.

Given the limited number of images available for training and reading, these two image datasets are often mixed. Using the same dataset for both training and reading introduces bias into the results – ideally a different image dataset should be used for reading. In the literature there are two methods to train an observer [44] – the holdout method and the re-substitution method. In order to study how bias influences the observer results and what constitutes an acceptable level of bias, 24 DBT acquisitions taken at the standard dose were divided into two equal datasets for training and reading [36]. Bias was varied from 0% to 100% in steps of 1/12, with initially 0% bias meaning that the two sets consisted of unique scans of the phantom – this is the so-called 'holdout' method. Then, in successive (1/12) steps, training DBT acquisitions were substituted by reading DBT acquisitions until the CHO only used reading images for both training and reading. This is termed the 're-substitution' method and has 100% bias. The mean error for each biased MO result from the respective non-biased result was calculated, and the highest bias percentage that had a performance within the 95% CI of the non-biased results was defined as the 'permitted' bias percentage. In order to explore the influence of dose (noise) level on this bias estimate, this process was then repeated for the low and high dose acquisition datasets.

If sound conclusions are to be drawn regarding image quality assessment of an imaging system in a QC setting, then CHO reproducibility is crucial. This was evaluated using 60 DBT acquisitions, split into 5 groups, each group then read by the finalized CHO. The standard deviation from the 5 readings for each lesion was taken as an estimate for the CHO reproducibility and compared to HR reproducibility for the same 5 image groups. Finally, the potential use of the MO was tested on a first practical application: the CHO was applied to the three sets of DBT acquisitions made at low, AEC and high dose levels, to examine whether the finalized CHO tracked HR performance as dose changed. The criteria in table 1 were used to compare the results of the CHO to

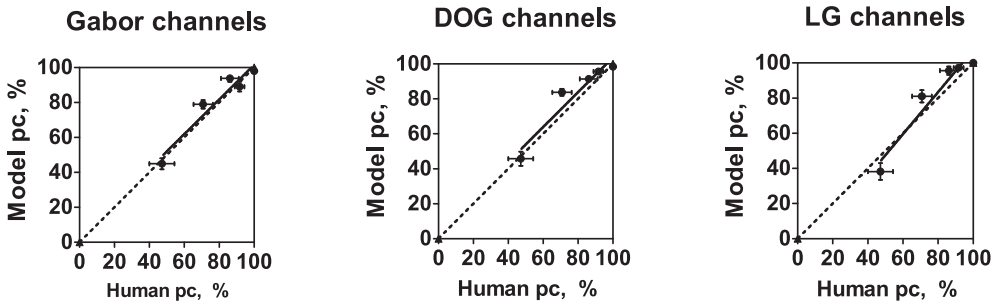

Fig. 6. Graphs of CHO performance plotted against human observer performance. The three graphs represent the results of the best tuning parameter set of the three channel types, listed in Table 3.

the human readings.

3. Results

3.1. Channel comparison

After channel tuning for the best combination of correlation (r), slope and ME against human observer, the MO performance of each channel type with the best tuning parameters is shown in Fig. 6. The corresponding tuning parameter values are listed in Table 3, with the Gabor channels giving the best overall score for the lesion targets in this phantom.

3.2. Channel tuning parameters

Fig. 7 shows CHO performance as channel parameters are systematically varied for the three different channel types and separately for all lesion sizes. In this figure, one parameter is varied while the remaining parameters are fixed at the values in Table 7. For Gabor and LG channels, large tuning ability is observed, whereas DOG performance is only influenced by the parameter σ_0 and not by α or Q. Vertical lines indicate the parameters that give the closest match between CHO and human observer performance (Fig. 6 and Table 7).

3.3. Expected signal

The influence of signal template on PC for the Gabor CHO performance is plotted in Fig. 8 for the different expected signal templates studied. The first set of results represents the HR scores (gold standard). Table 4 presents the evaluation results for each CHO; best match is given by estimating the expected signal from a training set of images i.e. what could be considered the standard method [36]. This is followed by using a Gaussian 'blob' with FWHM equal to the mass diameter. The observer templates used for the upcoming tests were thus produced from the training images.

3.4. Training images

Fig. 9 shows the result of changing the number DBT scans used to train the CHO template from 2 to 48. The dashed vertical lines in this figure show the points at which the PC moves outside the 95% CI of the ground truth, as the number of DBT training scans is reduced from 48. The overall number of DBT scans or acquisitions was found by averaging over data for the different lesions (Table 5). As a result of this analysis, further studies with this CHO/phantom combination will be performed with 12 training DBT acquisitions.

3.5. Bias

Fig. 10 shows the influence of changing bias, in 13 steps from 0% to 100%, on the PC for the Gabor channel CHO, with the values ordered from 0% to 100% bias. Table 6 gives the bias threshold ratio, defined as the fraction of biased training images for which the MO results remain within the 95% CI of completely non-biased training. Table 6 also shows the ME difference between 100% biased and non-biased MO results.

As a result of this tuning and training procedure, the CHO with parameters summarized in Table 7 was found as the finalized CHO for the structured phantom used with the Siemens Inspiration DBT system.

3.6. Reproducibility

Fig. 11 shows the mean and the standard deviation of the PC data from the reproducibility study, indicating comparable reproducibility for both observers. This figure shows slightly lower PC for CHO for the 4.3 mm diameter mass, relative to the HR data. It is possible that this is a result of the final tuning parameters used for the Gabor channels. Table 8 lists the standard deviation of the PC results from the five observations (for both model observer and human reading) and for all lesion sizes (units of percentage correct).

3.7. Influence of dose level

The results of applying the finalized CHO (Table 7) to the DBT scans acquired at low, AEC and high dose levels are compared to HR data in Fig. 12.a). Average PC scores for human observer and CHO are plotted versus dose level in Fig. 12.b). for the 2.1 mm diameter lesion. The evaluation criteria applied on these data sets are listed in Table 9.

4. Discussion

The impetus behind this study was the design, tuning, validation and documentation of a CHO for use with a 3D structured test object in DBT. Although this test object has been shown to produce reliable

Table 3
The optimal CHO channel parameters arrived at following the tuning process. The final mean absolute error, linear regression slope (a) and Pearson correlation (r) between human observer and CHO are given for the final parameter set. The 95% CIs for the slope and correlation are shown in brackets.

Channel type	*Channel parameters*	*ME*	*a*	*r*
Gabor	$t_t = 1.2$, $t_f = 65$, $t_w = 15$	−1.75	0.99 (0.50–1.48)	0.967 (0.558–0.998)
DOG	$\sigma_0 = 0.016$, $\alpha = 1.4$, $Q = 2.4$	−3.87	0.99 (0.47–1.52)	0.961 (0.517–0.998)
LG	$\sigma_u = 14$	−3.27	1.20 (0.59–1.80)	0.964 (0.548–0.998)

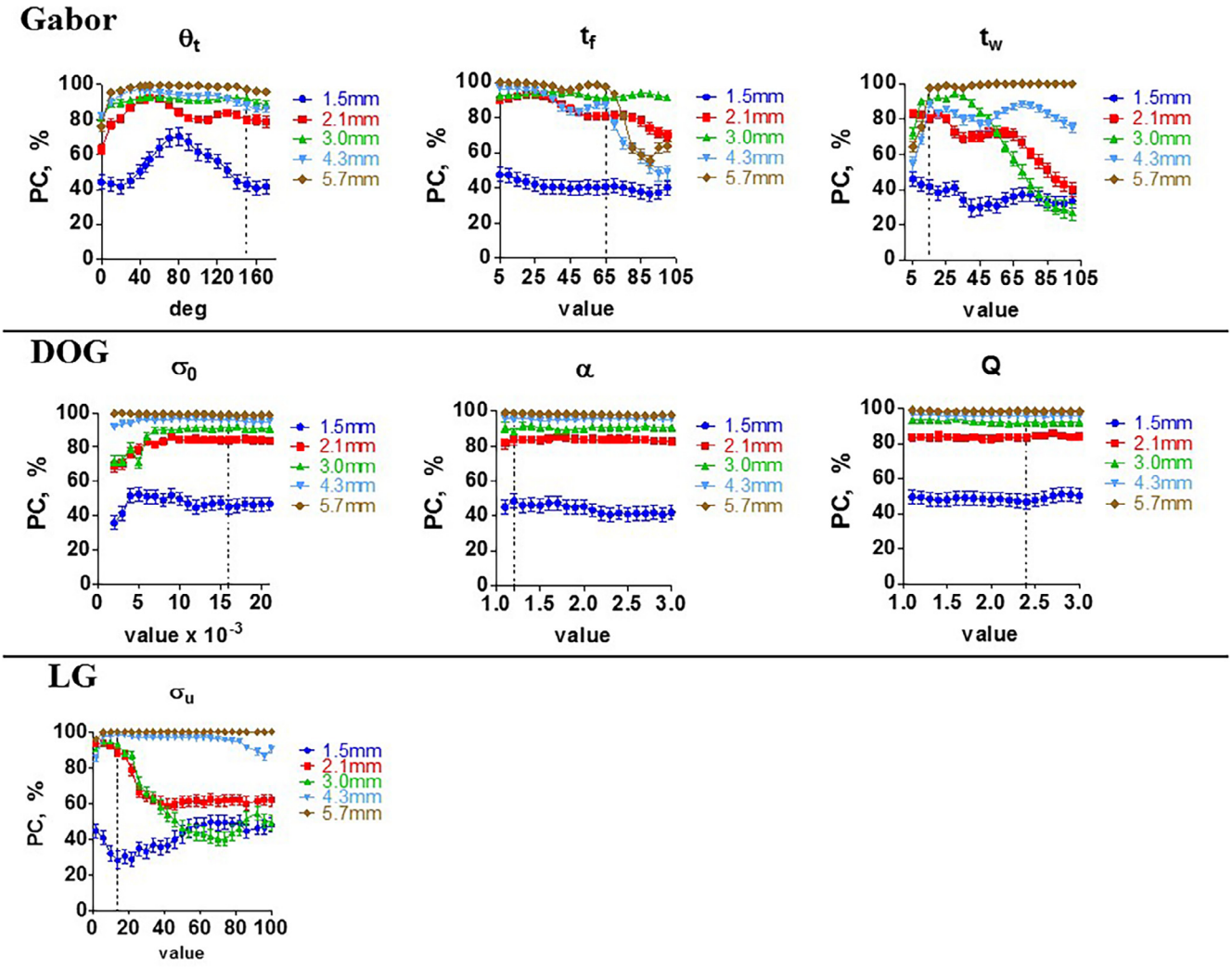

Fig. 7. CHO performance versus channel tuning parameter value. The parameter varied is indicated above each graph (Eqs. (3)–(5)), where the other tuning parameters were kept constant at the values indicated by the vertical dashed line. The combination of tuning parameters specified by the dashed lines, form the channels described in Table 7 and produce observer results shown in Fig. 6.

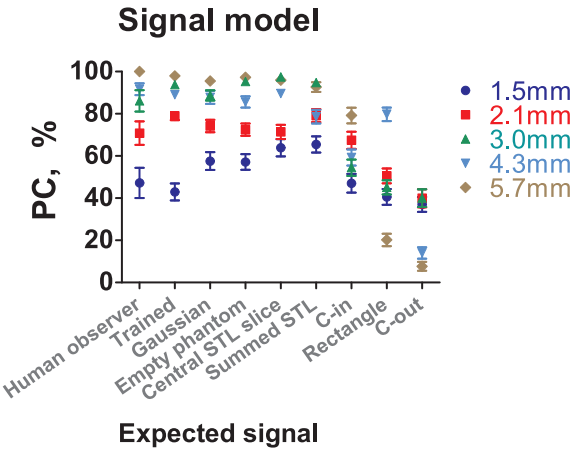

Fig. 8. Human observer performance along with CHO performances using different estimations of the expected signal, ordered from highest to lowest correlation with HR results.

results when read with human readers [13], the use of the phantom for routine QC purposes is expected to benefit from more objective methods such as MO methods. This has been the case for the CDMAM test object, where the availability of automatic readout via CDCOM has greatly facilitated the used of this phantom in routine of FFDM systems [45].

Table 4
Influence of signal template method on CHO performance.

Template	*ME*	*a*	*r*
Trained CHO	−1.35	1.03 (0.50–1.56)	0.963 (0.537–0.998)
Gaussian blob	−1.61	0.72 (0.56–0.88)	0.993 (0.892–0.999)
Empty phantom	−2.40	0.77 (0.31–1.23)	0.951 (0.425–0.997)
Central slice	−4.50	0.67 (0.15–1.19)	0.922 (0.210–0.995)
Summed	−2.97	0.48 (−0.12–1.07)	0.827 (−0.205–0.989)
C-small	17.70	1.02 (−0.39–1.20)	0.685 (−0.499–0.977)
Rectangle	31.94	0.01 (−1.89–1.91)	0.008 (−0.881–0.884)
C-large	51.44	−0.52 (−1.52–0.47)	−0.695 (−0.978–0.485)

Three main aspects of CHO design were studied: channel type and associated parameters, the expected signal template and the covariance matrix. This CHO design phase was then completed with a study on the required number of test images and the amount of acceptable bias. As a start, channel selection examined three channel types, namely Gabor, DOG and LG, and assessed the influence of channel tuning on CHO performance. After tuning, all three channel types tracked human observer results with good accuracy. The number of channels for each

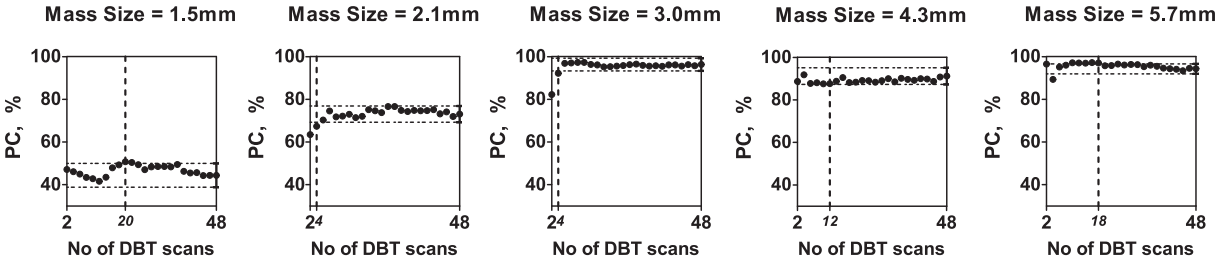

Fig. 9. Gabor channel CHO performance for the 5 masses as a function of the number of training DBT acquisitions. The dotted horizontal lines represent the 95% CI for the gold standard CHO trained with 48 DBT scans. The dashed vertical line points to the highest number of acquisitions producing the CHO performance outside the 95% CI of the gold standard.

Table 5
The number of training DBT stacks and the number of signal present and signal absent ROIs cropped from them, forming a CHO template, which produces observer results inside the 95% CI of the gold standard CHO results.

Lesion diameter [mm]	1.5	2.1	3.0	4.3	5.7	Overall
DBT scans	20	4	4	12	18	11.6
Absent ROIs	1500	300	300	900	1350	870
Present ROIs	60	12	12	36	54	35

channel type was fixed to 8, as preliminary results showed that using higher number of channels generally requires more training images. In addition higher number of channels increase the CHO performance, as investigated by Castella et al. [31], which could lead to poorer estimation, when the results are compared to human observer scores. Thus, 8 channels were chosen as a good compromise between observer performance and requirement of training images. The CHO with Gabor channels was selected for further investigation, as this gave the closest correspondence to our human reader scores. DOG channels gave similar

Table 6
Bias threshold and ME results between 0% and 100% biased model observer results.

Dose level	Lesion diameter (mm)	Bias threshold	ME_{0-100}
Low	1.5	5/12	10.1
	2.1	12/12	2.57
	3.0	10/12	−1.7
	4.3	11/12	5.4
	5.7	12/12	0.6
AEC	1.5	4/12	21.6
	2.1	12/12	0.9
	3.0	12/12	0.6
	4.3	5/12	−0.8
	5.7	12/12	0.6
High	1.5	7/12	14.7
	2.1	12/12	−1.0
	3.0	12/12	−2.71
	4.3	12/12	0.5
	5.7	12/12	0.4

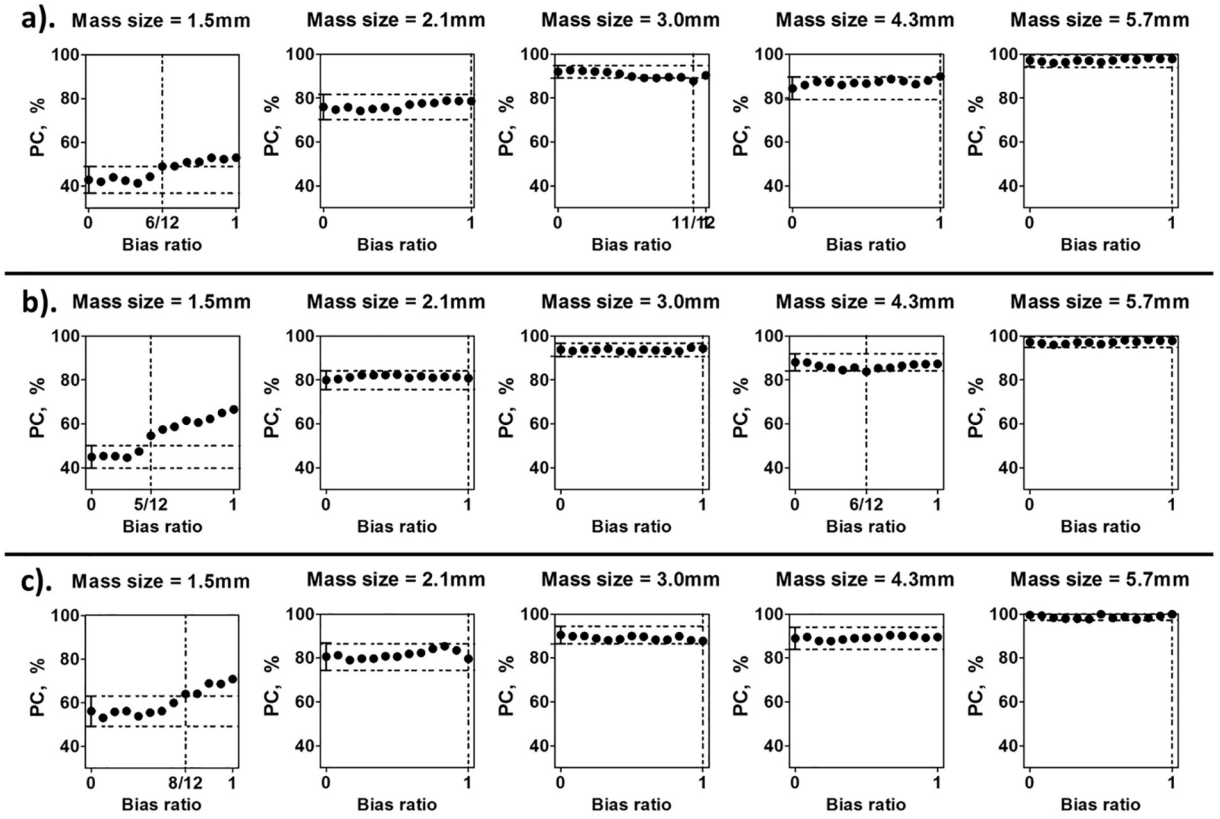

Fig. 10. a). Low dose level; b). AEC dose level; c). High dose level. Bias ratio of Gabor channel CHO readings with result outside the 95% CI of the non-biased data, evaluated for different mass sizes. Horizontal dotted lines represent 95% CI of the non-biased data and the dashed vertical line – the limit bias ratio.

Table 7
CHO formation method with best overall performance against human observer scores from structured phantom scanned on Siemens.

CHO component	Channels	Expected signal	Training images	Bias
Trained Gabor channels	$t_t = 1.2$, $t_f = 65$, $t_w = 15$	From acquired (training) images	Cropped from 12 DBT acquisitions	Separate DBT acquisitions for training and reading. Some bias not detrimental

slope and Pearson correlation, however the mean error was lower for the Gabor set. With regard to channel tuning, the DOG channels showed a rather low tuning potential: only the initial channel width had an impact on the model observer performance. This lack of tuning ability would mean that where CHO and human reading results differed, one would probably be forced to implement some form of internal noise method, rather than using tuning to match performance [46].

For Gabor channels, where orientation, width and frequency can be tuned, certain orientations (controlled by the parameter θ) gave notably higher detectability. This may be related to the particular shape of the masses in the phantom, especially for the 'trained' template, formed from many images. The main aim of the lesion model was to represent a real mass [47] and there was therefore no attempt to select a more isotropic mass model. Only one shape of mass-like lesion was used (with different dimensions) and the models were carefully glued to have a similar orientation. While a change in the mass-like lesions used in the phantom is not anticipated, the use of Gabor channels means that the CHO could probably be adapted to new mass lesion types if required. At higher channel frequencies (t_f), there was a fall in the detectability for larger lesions, as the channels start to exclude parts of the target. A drop in detectability was also seen for smaller diameter masses as width (t_w) was increased, due to the larger background area around the targets, included in the computation of the decision variable. The response (i.e. in terms of PC) of Laguerre Gauss channels were also found to be sensitive to changes in the channel parameters. Increasing the initial channel width reduced the detectability score for the masses with diameter 3.0 mm, 2.1 mm and 1.5 mm, as expected.

Of equal importance is the training phase. Expected signal approximations ranging from visually good estimates, expected to give strong performance, to objects which were clearly not related to the targets, were examined. The template built from signal present/signal absent training images gave the closest match to the HR data (Table 4), supporting the standard approach to CHO template formation [36]. This was followed by the Gaussian blob signal, which somewhat surprisingly outperformed the a-priori signal methods such as using a slice through the STL file or a summed projection of the STL file. This brings us the question of which template should be used for (comparative) performance testing. This could be something close to a physical version of the signal (i.e. the input), or the signal as rendered by the imaging system (i.e. the output. As expected, signal template choice has strong impact on CHO performance. Significant correlation (p less than 0.026)

Table 8
Standard deviation of the CHO and human observer performance in the reproducibility study.

Lesion diameter	Standard deviation (PC)				
	1.5 mm	2.1 mm	3.0 mm	4.3 mm	5.7 mm
Model observer	5.3	3.2	2.5	2.9	2.3
Human reader	7.3	5.1	1.9	2.3	0.6

was only found for the template trained from signal present/signal absent images, the Gaussian blob, the empty phantom and the central STL file slice signal estimation methods. The remaining signal templates (the MIP, the rectangle and Landolt C) gave some idea of the CHO sensitivity to mismatched templates. Poorest performance occurred for the large Landolt C (C-large) ($r = -0.69$), whose diameter excluded the majority of expected image signal, but was sensitive to a region around the target. The results suggest that if limited amount of images are acquired, due to a practical infeasibility, the use of a Gaussian blob might be a good candidate for the observer template.

The number of images used for CHO training depends (theoretically) on the minimum number required and (practically) on the number of feasible acquisitions for phantom study such as this. This is in contrast to CHO studies using simulated images, where computation resources are generally the limiting factors [26,48]. As discussed, training images may also be needed for signal template generation, in which case this aspect must also be considered during image acquisition. The training for the covariance matrix is crucial for inversion and for the task of pre-whitening, where insufficient images give an unreliable covariance estimate or even a singular (non-invertible) covariance matrix. For the 3D structured phantom background in this work, the smallest number of training images needed to achieve observer performance with a CHO using 8 Gabor channels within 95% CI of the ground truth PC results was 12 DBT acquisitions. This gave 900 signal absent VOIs with in plane dimensions of 236 × 236 pixels and 5 adjacent planes ($20 \times 20 \times 5\,mm^3$) and 36 signal present VOIs with in plane dimensions of 236 × 236 pixels and 3 adjacent planes ($20 \times 20 \times 3\,mm^3$). Meaning that for a full reading of the phantom 24 DBT acquisitions (12 for training and 12 for reading) are enough to produce observer results not significantly different than an observer trained with images cropped from 48 DBT acquisitions and applied on a

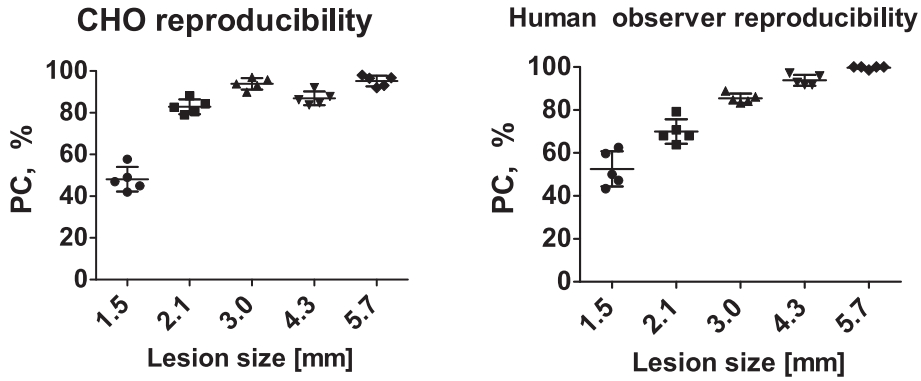

Fig. 11. Reproducibility of the Gabor channel MO (left) and human observers (right). The longer line for each mass size in both graphs represent the mean PC results and the two smaller lines – the upper and lower limit of the standard deviation.

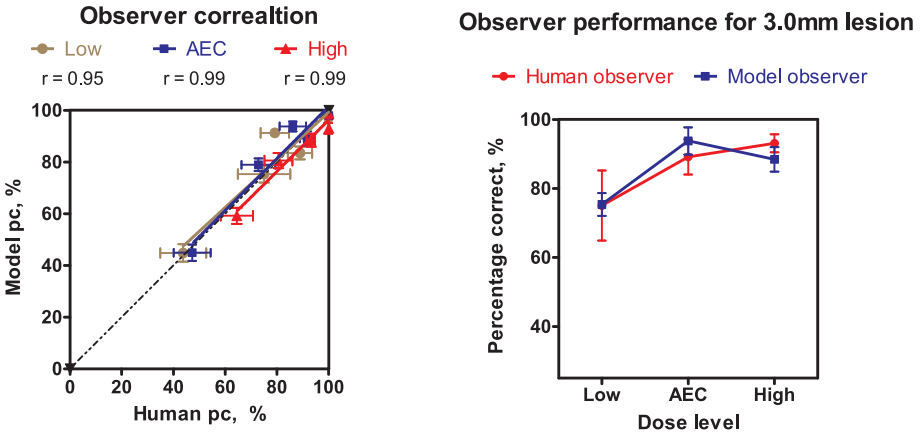

Fig. 12. Observer performance comparison for the three dose levels. a) presents the CHO scores against Human observer readings, with the correlation coefficients shown in the legend along with the dose level indication. b) PC scores for the 3.0 mm mass model for both observers versus dose level.

Table 9
CHO evaluation criteria for the dose levels study.

Dose level	*ME*	*a*	*r*
Low	−1.07	0.92 (0.36–1.49)	0.948 (0.407–0.997)
AEC	−1.75	0.99 (0.50–1.48)	0.967 (0.558–0.998)
High	3.68	0.99 (0.65–1.34)	0.983 (0.752–0.999)

separate set of 12 DBT acquisitions for reading. The results in Fig. 9 show the somewhat surprising trend that the number of images used for training has only a small influence on the PC value for most mass diameters, although there was some influence on PC for the 2.1 mm and 3.0 mm diameter masses. To examine this further, the training study was repeated using the "central slice" expected signal (Fig. 5b) instead of the trained expected signal (acquired from many images) (Fig. 5a). It was thought this would give some insight as to whether it was the signal template definition or the covariance estimation that was responsible for the trend seen. The results were closer to the expected behaviour (i.e. there was an increased influence of number of training images), but showed that a greater number of training images would be required. This suggests that using the expected signal may provide more information on the task than the central slice signal template to the final observer template, thus requiring fewer training images. This will be investigated in future work. Another limitation in the training study is the method of extraction of the signal absent images. As seen in Fig. 2.a. the signal absent ROIs overlap by 35% resulting in cropped ROIs that are not completely independent from one another and this can introduce a degree of bias in the training process. Extracting ROIs with 0% overlap would increase the number of acquisitions required to train the model observer, which could substantially prolong the scanning time. Acquisition of 24 scans typically takes between 30 and 45 min on current DBT systems. This precludes the use of such a test as part of daily QC, but is certainly practically feasible for physicists performing acceptance and yearly QC tests. The number of VOIs generated per DBT acquisition could be increased without a time penalty if a larger phantom were constructed and/or more signals were present in the phantom.

Turning to the bias study, the largest impact was seen on the smallest mass (1.5 mm), where bias for the AEC dose level gave a ME equal to 21.6 PC. However, for this mass, PC remained within the 95% CI until the bias ratio reached 5/12, while for the other 4 mass diameters, ME is less than 1% PC at all bias levels. Similar trend is observed for the Low and High dose levels, where the highest bias impact is seen for the smallest lesion, meaning that the dataset bias is not influenced by the dose level. Ideally, CHOs should use two different image sets for training and reading, but for cases where this is not possible, these data show that partly biased CHOs give acceptable results for the 3D structured phantom in this work. The extent to which this holds for other targets and backgrounds remains to be seen.

One potential application of the phantom used in this study is the routine assessment of DBT system image quality. Once the image quality score is established for some system baseline, the reproducibility determines the smallest deviations that can be tracked. The current reproducibility study showed that, when the same CHO is applied on another set of training and reading images from the same modality, the observer performance results might be slightly different. Nevertheless the differences were very similar to what was seen for human observers reading the same test sets. Neither CHO nor human readers could reliably detect the smallest mass (1.5 mm diameter), where PC was close to 50%. For the next smallest lesions (2.1 mm and 3.0 mm), the CHO had higher reproducibility while spread on the human reader results was smallest for the largest mass (5.0 mm). It is unclear as to why the CHO PC performance is slightly lower for the 4.3 mm diameter mass. Obviously, the CHO enables automated reading of phantom images, and suggests a great utility and benefit to physics QC services with limited staffing resources.

These steps resulted in a CHO that gave the closest correspondence with HR data for the phantom, yet using a practical number of channels and an acceptable minimal number of training images. The features of this CHO are summarized in Table 7. Application of this CHO on the 3D structured phantom data acquired at Low, AEC and High dose levels showed good agreement with human observer results. Linear correlation coefficient was higher than 0.95, slope larger than 0.92 and ME less than 2% PC. As for the impact of dose on detectability of masses, the CHO and human observers found the same overall result: the dose level has a minimal impact on detectability of these mass like lesions, a finding well known for FFDM and which also appears to apply to DBT images of the targets and background in the 3D structured phantom [49,50].

The present study has a number of limitations. First, the choice was made to use a ssCHO [24], applied over 5 adjacent planes, rather than a full 3D (volumetric) CHO. One could argue that a fully 3D CHO would offer a more robust MO implementation, with access to the volumetric dataset and possibly improved detection performance. Instead, three different channels function were used with the ssCHO and we were able to successfully match HR performance using all three methods. Given that the aim of the study was to implement a CHO for the evaluation of the 3D structured phantom for QA/QC purposes, we consider the ssCHO a simple and efficient approach. Further work may examine the use of multi-slice or volumetric CHOs for the phantom evaluation.

Second, the CHO algorithm was tuned and trained using DBT acquisitions made on a single device (Siemens Inspiration, using the EMPIRE reconstruction algorithm). Future work will investigate the

applicability of this CHO to other DBT systems and reconstruction algorithms. Differences in DBT scanning parameters (angular range, anti-scatter grid use, reconstruction algorithms [6]), lead to strong differences in the appearance of DBT images between the various systems. Whether this translates into differences in task performance evaluated using a CHO is unclear, yet preliminary data [51] show that a CHO with Gabor channels generated in real space can be a starting point in this investigation.

In the present analysis, medical physicists performed the human reading tests. Compared to standard contrast-detail phantoms with homogenous backgrounds [45], the structured test object has more realistic targets and background, closer to a radiologist task. While radiologists would likely perform differently from medical physicists, we expect that this would give a systematic offset, with minimal impact on scoring for QC purposes [52]. Differences in scoring between different groups of medical physicists can be expected too. We are using a 4AFC method for observer performance evaluation. This was considered a good compromise between the number of tests and the required statistical stability, as seen in similar studies in the literature [53].

One could question whether a mass-like object is required when evaluating imaging performance when some studies have shown a link between radiologist scores using simulated realistic (calcification) lesions and technical test object scores using sharp edged gold discs [54]. Future developments in DBT, for example the synthetic mammogram calculated from the reconstructed stack, may use a-priori information on breast anatomy and search for lesion-like objects in the volume, enhancing their appearance in the final image. Purely technical (perhaps structure-less or lack of breast-like anatomy) test objects may fail to provide an accurate assessment of system or algorithm performance in this case. If this is pursued, then some decision as to the shape and type of the mass-model lesions to be included must be made – clearly this needs to cover the range of lesions relevant to breast imaging. Alternatively, the results from the different template study suggest we may not expect large differences using cancer-like targets with slightly different shapes, as the Gaussian blob gave promising results when used as signal template. This could be seen as a kind of averaging over a range of non-spiculated lesion orientations. Nevertheless, if the models are too far–removed from reality, then there may be differences in absolute performance for both human (and correspondingly the developed CHOs) performance, as found in the study by Elangovan et al. [55].

Finally, the CHO was only tested on one type of background i.e. that of the structured phantom. Thus, the tuning parameters and number of channels are only relevant to the background and lesion properties of this particular phantom. In future, we expect to extend the use of the CHO to applications in virtual clinical trials [56]. This will require further evaluation in real and simulated anthropomorphic backgrounds covering a range of breast glandularities and lesion types. Whether the Gabor based CHO derived here for the phantom background would prove successful is not known, however the systematic CHO design process laid out here would help in the design of a robust model observer.

5. Conclusions

With reconstructed images that may use non-linear methods of image generation and processing, CHOs are a promising means of image evaluation, but only if carefully tuned and trained. This study has shown that a CHO can be tuned, trained and applied with acceptable reproducibility to DBT acquisitions of a physical phantom. The CHO could also be applied to sets of images acquired at different dose levels. The systematic procedure outlined here should help in CHO development for other tomosynthesis systems and for new applications.

Acknowledgements

This work is part of the OPTIMAM2 project funded by Cancer Research UK (grant number: C30682/A17321). We wish to thank our (human) readers for their time and help.

References

[1] Glasziou P, Houssami N. The evidence base for breast cancer screening. Prev. Med. (Baltim) 2011;53(3):100–2.
[2] Donzelli Alberto. The benefits and harms of breast cancer screening: an independent review. Lancet Nov. 2012;380(9855):1778–86.
[3] Rafferty EA. Digital Mammography: novel Applications. Radiol Clin North Am 2007;45(5):831–43.
[4] Poplack SP, Tosteson TD, Kogel CA, Nagy HM. Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography. Am J Roentgenol 2007;189(3):616–23.
[5] Sechopoulos I. A review of breast tomosynthesis. Part I. The image acquisition process. Med Phys 2013;40(1):1–12.
[6] Sechopoulos I. A review of breast tomosynthesis. Part II. Image reconstruction, processing and analysis, and advanced applications. Med Phys 2013;40(1):1–17.
[7] Maldera A, De Marco P, Colombo PE, Origgi D, Torresin A. Digital breast tomosynthesis: dose and image quality assessment. Phys. Medica Jan. 2017;33:56–67.
[8] Svahn TM, Chakraborty DP, Ikeda D, Zackrisson S, Do Y, Mattsson S, Andersson I. Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy. Br J Radiol 2012;85(1019).
[9] Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, et al. Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration. Eur Radiol 2013;23(8):2061–71.
[10] Niklason LT, Christian BT, Niklason LE, Kopans DB, Castleberry DE, Opsahl-Ong BH, et al. Digital tomosynthesis in breast imaging. Radiology 1997;205(2):399–406.
[11] Ferreira P, Baptista M, Di Maria S, Vaz P. Cancer risk estimation in Digital Breast Tomosynthesis using GEANT4 Monte Carlo simulations and voxel phantoms. Phys. Medica May 2016;32(5):717–23.
[12] van Engen RE, Bosmans H, Bouwman RW, Dance DR, Heid P, Lazzari B, et al. A European Protocol for Technical Quality Control of Breast Tomosynthesis Systems. Cham: Springer; 2014. p. 452–9.
[13] Cockmartin L, Marshall NW, Zhang G, Lemmens K, Shaheen E, Van Ongeval C, Fredenberg E. Design and application of a structured phantom for detection performance comparison between breast tomosynthesis and digital mammography. Phys. Med. Biol. Vol. 62, Number 3 2017;15.
[14] Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis., vol. 19, no. 4. 2006.
[15] Zhao B, Zhou J, Hu Y-H, Mertelmeier T, Ludwig J, Zhao W. Experimental validation of a three-dimensional linear system model for breast tomosynthesis. Med Phys 2008;36(1):240–51.
[16] Marshall NW, Bosmans H. Measurements of system sharpness for two digital breast tomosynthesis systems. Phys Med Biol 2012;57(22):7629–50.
[17] Rodríguez-Ruiz A, Castillo M, Garayoa J, Chevalier M. Evaluation of the technical performance of three different commercial digital breast tomosynthesis systems in the clinical environment. Phys. Medica 2016;32(6):767–77.
[18] Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. Proc Natl Acad Sci USA 1993;90(21):9758–65.
[19] Burgess AE, Jacobson FL, Judy PF. Human observer detection experiments with mammograms and power-law noise. Med Phys 2001;28(4):419–37.
[20] Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. J Opt Soc Am A 2001;18(3):473–88.
[21] Racine D, Ba AH, Ott JG, Bochud FO, Verdun FR. Objective assessment of low contrast detectability in computed tomography with Channelized Hotelling Observer. Phys Medica 2016;32(1):76–83.
[22] Castella C, Eckstein MP, Abbey CK, Kinkel K, Verdun FR, Saunders RS, et al. Mass detection on mammograms: influence of signal shape uncertainty on human and model observers. J Opt Soc Am A 2009;26(2):425–36.
[23] Fetterly KA, Favazza CP. Direct estimation and correction of bias from temporally variable non-stationary noise in a channelized Hotelling model observer. Phys Med Biol 2016;61(15):5606–20.
[24] Platiša L, Goossens B, Vansteenkiste E, Park S, Gallas BD, Badano A, et al. Channelized Hotelling observers for the assessment of volumetric imaging data sets. J Opt Soc Am A Opt Image Sci Vis 2011;28(6):1145–63.
[25] Young S, Bakic PR, Myers KJ, Jennings RJ, Park S. A virtual trial framework for quantifying the detectability of masses in breast tomosynthesis projection data. Med Phys May 2013;40(5):051914.
[26] Park S, Zhang G, Myers KJ. Comparison of channel methods and observer models for the task-based assessment of multi-projection imaging in the presence of structured anatomical noise. IEEE Trans Med Imaging 2016;35(6):1431–42.
[27] Zeng R, Badano A, Myers KJ. Optimization of digital breast tomosynthesis (DBT) acquisition parameters for human observers: effect of reconstruction algorithms. Phys Med Biol 2017.
[28] Wen G, Markey MK, Haygood TM, Park S. Model observer for assessing digital breast tomosynthesis for multi-lesion detection in the presence of anatomical noise.

Phys. Med. Biol. 2018;63(4).
[29] Michielsen K, Nuyts J, Cockmartin L, Marshall NW, Bosmans H. Design of a model observer to evaluate calcification detectability in breast tomosynthesis and application to smoothing prior optimization. Med Phys 2016;43(12):6577–87.
[30] Park S, Jennings R, Liu H, Badano A, Myers K. A statistical, task-based evaluation method for three-dimensional x-ray breast imaging systems using variable-background phantoms. Med Phys 2010;37:6253–70.
[31] Castella C, Abbey CK, Eckstein MP, Verdun FR, Kinkel K, Bochud FO, et al. linear template with mammographic backgrounds estimated with a genetic algorithm. J Opt Soc Am A Opt Image Sci Vis 2007;24(12):B1–12.
[32] Bouwman RW, Goffi M, van Engen RE, Broeders MJM, Dance DR, Young KC, et al. Can the channelized Hotelling observer including aspects of the human visual system predict human observer performance in mammography? Phys. Medica 2017;33:95–105.
[33] Abdurahman S, Dennerlein F, Jerebko A, Fieselmann A, Mertelmeier T. Optimizing high resolution reconstruction in digital breast tomosynthesis using filtered back projection. Lect Notes Comput Sci (including Subser Lect. Notes Artif Intell Lect Notes Bioinformatics) 2014;8539 LNCS:520–7.
[34] Abdurahman S, Jerebko A, Mertelmeier T, Lasser T, Navab N. Out-of-plane artifact reduction in tomosynthesis based on regression modeling and outlier detection. Berlin, Heidelberg: Springer; 2012. p. 729–36.
[35] Zhang G, Cockmartin L, Bosmans H. A four-alternative forced choice (4AFC) software for observer performance evaluation in radiology, 2016, p. 97871E.
[36] Barrett HH, Myers KJ, Rathee S. Foundations of Image. Science 2004.
[37] Myers KJ, Barrett HH. Addition of a channel mechanism to the ideal-observer model. J Opt Soc Am A 1987;4(12):2447–57.
[38] Bochud F, Abbey C, Eckstein M. Statistical texture synthesis of mammographic images with super-blob lumpy backgrounds. Opt Express 1999;4(1):33–42.
[39] Castella C, Kinkel K, Descombes F, Eckstein MP, Sottas P-E, Verdun FR, et al. Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm. Opt Express 2008;16(11):7595.
[40] Barrett HH, Myers KJ, Hoeschen C, Kupinski MA, Little MP. Task-based measures of image quality and their relation to radiation dose and patient risk. Phys Med Biol 2015;60(2):R1–75.
[41] Watson AB. Detection and recognition of simple spatial forms. Berlin SpringerVerlag 1983:100–14.
[42] Gallas BD, Barrett HH. Validating the use of channels to estimate the ideal linear observer. J Opt Soc Am A Opt Image Sci Vis 2003;20(9):1725–38.
[43] Schrauf M, Stern C. The visual resolution of Landolt-C optotypes in human subjects depends on their orientation: the 'gap-down' effect. Neurosci Lett Feb. 2001;299(3):185–8.
[44] Gallas BD, Variance of the channelized-hotelling observer from a finite number of trainers and testers, 2003, p. 100.
[45] Karssemeijer N, Thijssen MAO. Determination of contrast-detail curves of mammography systems by automated image analysis. Elsevier 1996:115–60.
[46] Brankov JG. Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection. Phys Med Biol 2013;58:7159–82.
[47] Shaheen E, De Keyzer F, Bosmans H, Dance DR, Young KC, Van Ongeval C. The simulation of 3D mass models in 2D digital mammography and breast tomosynthesis. Med. Phys. Phys 2014;41(36):81913–4920.
[48] Young S, Park S, Anderson SK, Badano A, Myers KJ, Bakic PR. Estimating breast tomosynthesis performance in detection tasks with variable-background phantoms. SPIE Med Imaging Phys Med Imaging 2009;7258(301):72580O.
[49] Saunders RS, Baker JA, Delong DM, Johnson JP, Samei E. Does image quality matter? Impact of resolution and noise on mammographic task performance. Med Phys 2007;34(10):3971–81.
[50] Timberg P, Båth M, Andersson I, Svahn T, Ruschin M, Hemdal B, et al. Impact of dose on observer performance in breast tomosynthesis using breast specimens. SPIE Med Imaging Phys Med Imaging 2008;6913:69134J.
[51] Petrov D, Cockmartin L, Marshall N, Vancoillie L, Young K, Bosmans H. Real space channelization for generic DBT system image quality evaluation with channelized Hotelling observer. 2017, vol. 10136, p. 101360N.
[52] Elangovan P, Mackenzie A, Dance DR, Young KC, Wells K. Using non-specialist observers in 4AFC human observer studies. 2017, p. 1013256.
[53] Jäkel F, Wichmann FA. Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. J Vis 2006;6(11):1307–22.
[54] Warren LM, Mackenzie A, Cooke J, Given-Wilson RM, Wallis MG, Chakraborty DP, et al. Effect of image quality on calcification detection in digital mammography. Med Phys 2012;39(6):3202–13.
[55] Elangovan P, Mackenzie A, Dance DR, Young KC, Wells K. Lesion detectability in 2D-mammography and digital breast tomosynthesis using different targets and observers. Phys Med Biol 2018;63(9):1–15.
[56] Maidment ADA. Virtual clinical trials for the assessment of novel breast screening modalities. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2014;8539 LNCS:1–8.