Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

A comparison of resampling schemes for estimating model observer performance with small ensembles

To cite this article: Fatma E A Elshahaby *et al* 2017 *Phys. Med. Biol.* **62** 7300

View the article online for updates and enhancements.

myQA is trusted by >1,000 RT centers!
myQA®
iba
INTEGRATED QUALITY ASSURANCE
myQA provides one central platform and database integrating all your QA applications.
All-in-One. All Connected. All Secure.
iba-dosimetry.com

IOP Publishing | Institute of Physics and Engineering in Medicine Physics in Medicine & Biology

Phys. Med. Biol. **62** (2017) 7300–7320 https://doi.org/10.1088/1361-6560/aa807a

A comparison of resampling schemes for estimating model observer performance with small ensembles

Fatma E A Elshahaby[1,2,3], **Abhinav K Jha**[2], **Michael Ghaly**[2] **and Eric C Frey**[1,2]

[1] Department of Electrical and Computer Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21218, United States of America
[2] The Russell H Morgan Department of Radiology and Radiological Science, School of Medicine, Johns Hopkins University, Baltimore, MD 21287, United States of America
[3] Department of Computers and Systems, Electronics Research Institute, Cairo, Egypt

E-mail: felshahaby@gmail.com

Received 28 February 2017, revised 8 July 2017
Accepted for publication 18 July 2017
Published 22 August 2017

Abstract

In objective assessment of image quality, an ensemble of images is used to compute the 1st and 2nd order statistics of the data. Often, only a finite number of images is available, leading to the issue of statistical variability in numerical observer performance. Resampling-based strategies can help overcome this issue. In this paper, we compared different combinations of resampling schemes (the leave-one-out (LOO) and the half-train/half-test (HT/HT)) and model observers (the conventional channelized Hotelling observer (CHO), channelized linear discriminant (CLD) and channelized quadratic discriminant). Observer performance was quantified by the area under the ROC curve (AUC). For a binary classification task and for each observer, the AUC value for an ensemble size of 2000 samples per class served as a gold standard for that observer. Results indicated that each observer yielded a different performance depending on the ensemble size and the resampling scheme. For a small ensemble size, the combination [CHO, HT/HT] had more accurate rankings than the combination [CHO, LOO]. Using the LOO scheme, the CLD and CHO had similar performance for large ensembles. However, the CLD outperformed the CHO and gave more accurate rankings for smaller ensembles. As the ensemble size decreased, the performance of the [CHO, LOO] combination seriously deteriorated as opposed to the [CLD, LOO] combination. Thus, it might be desirable to use the CLD with the LOO scheme when smaller ensemble size is available.

1361-6560/17/187300+21$33.00

Keywords: channelized model observers, leave-one-out resampling scheme, rank correlation coefficient, Hotelling observer

(Some figures may appear in colour only in the online journal)

1. Introduction

Model observers have been widely used in medical imaging to objectively assess image quality (Barrett *et al* 1993, Barrett and Myers 2004). Model observers are especially important in applications such as instrumentation or imaging method optimization where task performance needs to be measured for a large number of configurations. Both ideal and anthropomorphic observers have been used for a variety of applications (Barrett and Myers 2004). One commonly used model observer is the channelized Hotelling observer (CHO), which consists of the Hotelling observer (HO) applied to the outputs of a channel model. A channel model consists of a set of templates that are applied to an image and produce a feature vector. The length of the feature vector is equal to the number of templates. The HO is a linear classifier that uses the 1st and 2nd order statistics of the data; for the CHO the data are the feature vectors. A typical channel model is comprised of a set of band-pass filters. With an appropriate channel model, the CHO can effectively model the performance of the human observer for signal-known exactly and background-known exactly or statistically (SKE/BKE and SKE/BKS) tasks (Myers and Barrett 1987, Wollenweber *et al* 1999, Gifford *et al* 2000, Park *et al* 2005a).

For realistic medical images, analytical expressions for the distribution of input data (or feature vectors) are often not available. Thus, the 1st and 2nd order statistics of the data are frequently estimated using ensemble techniques. A large number of images with known truth is needed to provide reliable estimates of these quantities (Fukunaga and Hayes 1989a, Kupinski *et al* 2007, Ge *et al* 2014). Furthermore, in large optimization and evaluation studies, task performance is assessed for many different combinations of system parameters and methods, such as different collimator designs (Yihuan *et al* 2014, Ghaly *et al* 2016), reconstruction methods and parameters (Frey *et al* 2002, Gilland *et al* 2006, He *et al* 2006), and post-reconstruction filters and processing techniques (Frey *et al* 2002, Sankaran *et al* 2002). Thus, ensemble techniques require an enormous number of images to be obtained and stored, which often limits the number of parameters that can be explored. Smaller ensemble sizes can be used, but result in imprecise estimates of observer performance (Fukunaga and Hayes 1989a, 1989b, Kupinski *et al* 2007). Methods for reducing the ensemble sizes have been proposed in Wunderlich and Noo (2009) and Tseng *et al* (2016).

One commonly used method to maximize the statistical power (i.e. the ability to correctly rank the systems) from a small ensemble of images is to use resampling schemes such as the leave-one-out (LOO) scheme (Fukunaga 1990). A comparison of different resampling schemes as a function of ensemble size can be found in Chan *et al* (1999) and Sahiner *et al* (2008). Also, the performances of the linear and quadratic discriminants have been investigated in previous studies (Fukunaga and Hayes 1989b, Chan *et al* 1999, Sahiner *et al* 2008). However, to best of our knowledge, the performance of the CHO was not compared to other observers, such as the channelized linear discriminant (CLD) and channelized quadratic discriminant (CQD) (Fukunaga 1990), for different resampling schemes.

Since the performance of model observers is affected by the reduction of the ensemble size, the goal of this study was to develop a strategy that is able to handle small ensembles. This was done by evaluating the performance of different combinations of model observers and resampling schemes and by exploring the trade-off between reliability of performance measures and ensemble size. Selecting the combination that has the best performance could help in providing more statistical power for a given number of images. This could help in studying a larger number of parameters. The evaluation was performed in the context of optimizing the post-reconstruction filter cut-off frequencies for myocardial perfusion SPECT defect detection (Frey *et al* 2002, He *et al* 2004, 2006). In particular, we compared the CHO, CLD and CQD (Fukunaga 1990). The two resampling schemes used were the half train/half test (HT/HT) and the LOO (Fukunaga 1990) schemes. The metric for performance of each observer was the area under the receiver operating characteristic curve (AUC). Since systems ranking is as important as the ability to predict the absolute task performance (Park *et al* 2005b), we also computed the Spearman rank-correlation coefficient (Daniel 1990) for AUC values from different cut-off frequencies as a function of ensemble size. A preliminary version of this work was reported in Elshahaby *et al* (2015b).

2. Model observers

For a binary classification task, we denote the defect-absent and the defect-present hypotheses by H_1 and H_2, respectively. Information about the classification decision is communicated via a scalar decision variable called the test statistic. In this work, we studied three model observers: the HO, the quadratic discriminant (QD) and the linear discriminant (LD).

2.1. Hotelling observer (HO)

A thorough explanation of the HO is given in Barrett and Myers (2004). The test statistic $\widehat{t}_{\mathbf{HO}}\,(\mathbf{g})$ for the HO that can be computed from the available data is given by:

$$\widehat{t}_{\mathrm{HO}}\,(\mathbf{g}) = \left[(\mathbf{S}_{\mathbf{g}}^{-1}(\overline{\mathbf{g}}_{|\mathrm{H}_2} - \overline{\mathbf{g}}_{|\mathrm{H}_1}))^{\mathrm{T}}\mathbf{g}\right], \tag{1}$$

where $\mathbf{g} \in \mathbb{R}^{M\times 1}$ is the measurement vector to be classified and $M \in \mathbb{Z}^{+}$, $\overline{\mathbf{g}}_{|\mathrm{H}_i} \in \mathbb{R}^{M\times 1}$ is the sample mean vector of the measurements from the ith class, and $\mathbf{S}_{\mathbf{g}}$ is the $M \times M$ intra-class scatter matrix, defined as the average of the sample covariance matrices from both classes and given mathematically by:

$$\mathbf{S}_{\mathbf{g}} = \Pr\,(H_1)\,\mathbf{S}_{\mathbf{g}|\mathrm{H}_1} + \Pr\,(H_2)\,\mathbf{S}_{\mathbf{g}|\mathrm{H}_2}\,, \tag{2}$$

where $\Pr(H_i)$ is the probability of occurrence of the ith class and $\mathbf{S}_{\mathbf{g}|\mathrm{H}_i} \in \mathbb{R}^{M\times M}$ is the sample covariance matrix of the ith class.

2.2. Quadratic discriminant (QD)

The test statistic of the QD is a quadratic function of $\mathbf{g}$ and it can be computed from the available measurements as below (Fukunaga 1990):

$$\begin{aligned}\widehat{t}_{\mathrm{QD}}\,(\mathbf{g}) = {} & \left[-\tfrac{1}{2}\mathbf{g}^{\mathrm{T}}\left(\mathbf{S}_{\mathbf{g}|\mathrm{H}_2}^{-1} - \mathbf{S}_{\mathbf{g}|\mathrm{H}_1}^{-1}\right)\mathbf{g}\right] \\ & + \left[(\mathbf{S}_{\mathbf{g}|\mathrm{H}_2}^{-1}\,\overline{\mathbf{g}}_{|\mathrm{H}_2} - \mathbf{S}_{\mathbf{g}|\mathrm{H}_1}^{-1}\,\overline{\mathbf{g}}_{|\mathrm{H}_1})^{\mathrm{T}}\mathbf{g}\right] \\ & - \tfrac{1}{2}\left[\overline{\mathbf{g}}_{|\mathrm{H}_2}^{\mathrm{T}}\,\mathbf{S}_{\mathbf{g}|\mathrm{H}_2}^{-1}\,\overline{\mathbf{g}}_{|\mathrm{H}_2} - \overline{\mathbf{g}}_{|\mathrm{H}_1}^{\mathrm{T}}\,\mathbf{S}_{\mathbf{g}|\mathrm{H}_1}^{-1}\,\overline{\mathbf{g}}_{|\mathrm{H}_1}\right] - \tfrac{1}{2}\log\frac{\left|\mathbf{S}_{\mathbf{g}|\mathrm{H}_2}\right|}{\left|\mathbf{S}_{\mathbf{g}|\mathrm{H}_1}\right|},\end{aligned} \tag{3}$$

where $\left|\mathbf{S}_{\mathbf{g}|\mathrm{H}_i}\right|$ is the determinant of the matrix $\mathbf{S}_{\mathbf{g}|\mathrm{H}_i}$. The QD will have optimal performance (i.e. the same as the IO performance) if the data follow multivariate normal (MVN) distribution with unequal covariance matrices under both hypotheses (Fukunaga 1990, Chan *et al* 1999).

2.3. Linear discriminant (LD)

The test statistic of the LD is a linear function of $\mathbf{g}$ and is defined as:

$$\begin{aligned}\widehat{t}_{\mathrm{LD}}\left(\mathbf{g}\right) = &\left[\left(\mathbf{S}_{\mathbf{g}}^{-1}\left(\overline{\mathbf{g}}_{|\mathrm{H}_2} - \overline{\mathbf{g}}_{|\mathrm{H}_1}\right)\right)^{\mathrm{T}}\mathbf{g}\right] \\ &-\tfrac{1}{2}\left[\overline{\mathbf{g}}_{|\mathrm{H}_2}^{\mathrm{T}}\mathbf{S}_{\mathbf{g}}^{-1}\overline{\mathbf{g}}_{|\mathrm{H}_2} - \overline{\mathbf{g}}_{|\mathrm{H}_1}^{\mathrm{T}}\mathbf{S}_{\mathbf{g}}^{-1}\overline{\mathbf{g}}_{|\mathrm{H}_1}\right].\end{aligned} \tag{4}$$

In this case, $\widehat{t}_{\mathbf{LD}}\left(\mathbf{g}\right)$ consists of two main terms. The first term is a linear function of $\mathbf{g}$, and it is the same as $\widehat{t}_{\mathbf{HO}}\left(\mathbf{g}\right)$. The second term is independent of the unknown measurement $\mathbf{g}$, but dependent on the estimated means and covariance matrices. Thus, equation (4) can be rewritten as follows:

$$\widehat{t}_{\mathrm{LD}}\left(\mathbf{g}\right) = \widehat{t}_{\mathrm{HO}}\left(\mathbf{g}\right) + \Delta, \tag{5}$$

where the extra term Δ is given by:

$$\Delta = -\frac{1}{2}\left[\overline{\mathbf{g}}_{|\mathrm{H}_2}^{T}\mathbf{S}_{\mathbf{g}}^{-1}\,\overline{\mathbf{g}}_{|\mathrm{H}_2} - \overline{\mathbf{g}}_{|\mathrm{H}_1}^{T}\,\mathbf{S}_{\mathbf{g}}^{-1}\overline{\mathbf{g}}_{|\mathrm{H}_1}\right]. \tag{6}$$

The performance of LD is optimal if the data have MVN distribution with equal covariance matrices under both hypotheses (Fukunaga 1990, Chan *et al* 1999). This observer also has additional optimal properties as described previously in Li *et al* (2017).

3. Methods

3.1. Projection data generation

This study was performed in the context of myocardial perfusion SPECT (MPS) imaging using 10 mCi of Tc-99m labeled tracer. We used projection data similar to that used previously in Ghaly *et al* (2014) and Elshahaby *et al* (2016). We used the male phantom with small body size, small heart size and small subcutaneous adipose tissue thickness described in Ghaly *et al* (2014). The simulated perfusion defect was a mid-ventricular placed in the anterolateral wall of the myocardium. The defect severity and extent were 10% and 25%, respectively, where the defect severity is defined as the percentage reduction in tracer uptake in the defect relative to the normal myocardium, and the defect extent is defined as the percentage of myocardial volume occupied by the perfusion defect. We generated 2000 pairs of noise-free defect-absent and defect-present projection images; uptake variability in organs was modeled (Ghaly *et al* 2014, Elshahaby *et al* 2016). Noise was simulated using a Poisson distributed random number generator.

3.2. Image reconstruction and post-reconstruction processing

The simulated noisy projection data were reconstructed using filtered back-projection (FBP) and a ramp filter with cut-off at the Nyquist frequency. The reconstructed voxels were cubic with a side length of 0.442 cm. We reconstructed a 48-transaxial-slice region centered on the heart, resulting in a 128 × 128 × 48 reconstructed image matrix. The reconstructed images

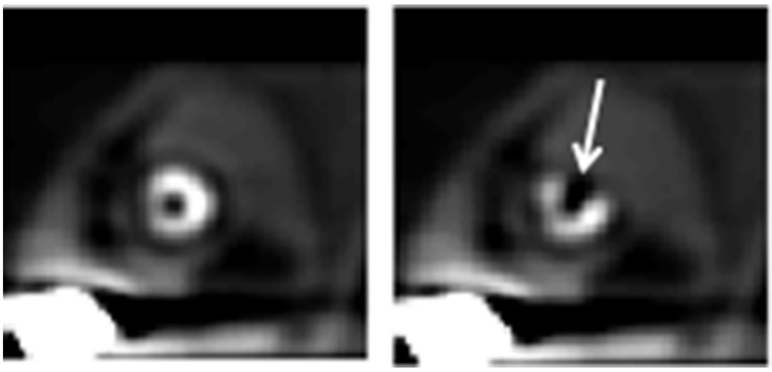

Figure 1. Noise-free short-axis images where the image on the left represents the defect-absent case and on the right represents the defect-present case. The arrow points to the defect; the defect shown has severity of 100% for visualization purposes.

were filtered with a 3D Butterworth filter of order 8 at cut-off frequencies 0.08, 0.1, 0.12, 0.14, 0.16, 0.2 and 0.24 cycles per pixel. The filtered images were reoriented into a standard short-axis orientation (Frey *et al* 2002, Ghaly *et al* 2015, Elshahaby *et al* 2015a) and a 64 × 64 image centered on the position of the defect for the defect-present class or the corresponding defect location for the defect-absent class was extracted and windowed (Frey *et al* 2002, He *et al* 2004, 2006, 2010, Gilland *et al* 2006, Ghaly *et al* 2015). In the windowing step, negative values were set to zero, values that were larger than or equal to the maximum value in the heart were set to 255, and the remaining values were mapped to the range [0, 255]. Lastly, the resulting floating-point values were rounded to integers. Figure 1 shows the resulting noise-free short-axis defect-absent and defect-present images.

3.3. Application of the frequency-selective channel model

A set of 6 rotationally symmetric channels (Myers and Barrett 1987) were used as shown in figure 2. The rotationally symmetric channel model has been widely used in similar tasks involving the assessment and optimization of different nuclear medicine systems using myocardial perfusion images (Wollenweber *et al* 1999, Frey *et al* 2002). In particular, for the MPS images, the rankings of the systems using the CHO were in good agreement with the rankings of human observers (Wollenweber *et al* 1999). In the frequency domain, these channels were non-overlapping passbands having square profiles with cut-offs of $\left[\frac{1}{128}, \frac{1}{64}\right]$, $\left[\frac{1}{64}, \frac{1}{32}\right]$, $\left[\frac{1}{32}, \frac{1}{16}\right]$, $\left[\frac{1}{16}, \frac{1}{8}\right]$, $\left[\frac{1}{8}, \frac{1}{4}\right]$, and $\left[\frac{1}{4}, \frac{1}{2}\right]$ cycles per pixel. The 2D frequency domain channels were transformed analytically to the spatial domain and sampled at the image voxel size. The DC component for each channel was explicitly removed by subtracting the mean value of the spatial domain template (Frey *et al* 2002, Elshahaby *et al* 2016). The dot product of the post-processed image with each of the spatial domain templates produced a 6 × 1 feature vector for each image. This process resulted in 2000 pairs of feature vectors (channel output vectors) for the defect-absent and defect-present classes.

3.4. Feature vector ensembles

Above we described the ensemble of feature vectors generated from realistic MPS data. We refer to the underlying distribution of these feature vectors as F-MPS. As observed in Elshahaby *et al* (2016) and Li *et al* (2017), the probability distribution of theses feature

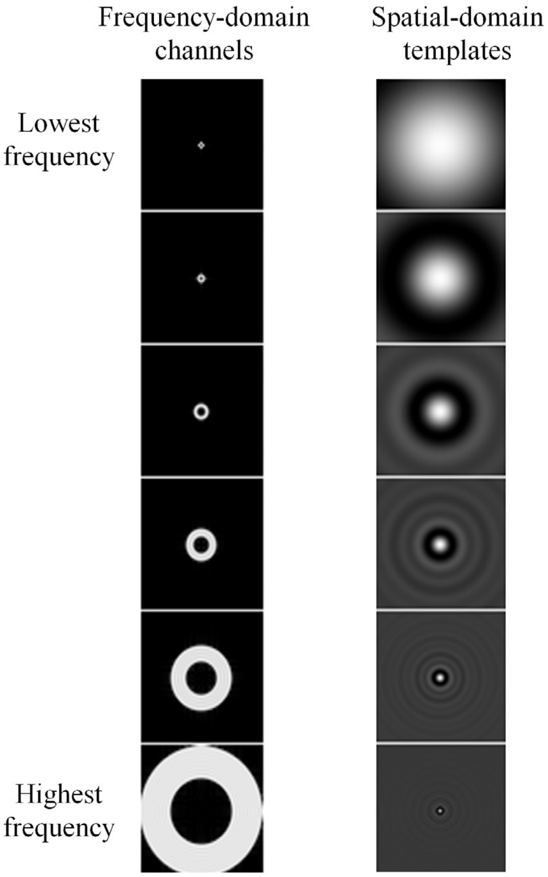

Figure 2. Images of the six rotationally symmetric frequency-domain channels (left) and the corresponding spatial-domain templates (right) are shown.

vectors is not necessarily consistent with an MVN distribution. The HO and linear discriminant (LD) have performance equal to the ideal observer (IO) when applied to data that are MVN distributed with equal covariance matrices under both hypotheses; the QD has the same performance as the IO when the data are MVN distributed and the covariance matrices are not necessarily equal (Fukunaga 1990). Since the MVN assumption was most likely violated with the considered dataset, the performance of all 3 observers applied to the feature vector data is expected to be sub-optimal.

Based on the above discussion and for completeness, we also generated data from 2 additional ensembles whose members were drawn from synthetic distributions (MVN with equal and unequal covariance matrices) to test the various observers and resampling schemes.

Both synthetic distributions modeled the mean and covariance of the feature vectors in F-MPS. We estimated these by calculating the sample mean vectors $\tilde{\mathbf{g}}_{\mathrm{H}_i}^{2000}$ and the covariance matrices $\tilde{\mathbf{S}}_{\mathbf{g}|\mathrm{H}_i}^{2000}$, where i denotes the class ($i \in \{1,2\}$), from the available 2000 feature vectors from each class. The first synthetic ensemble, F-MVNUNEQ, was created by generating 2000 feature vectors for each class using an MVN distributed random-number generator, where the parameters of the MVN distribution were $\tilde{\mathbf{g}}_{\mathrm{H}_i}^{2000}$ and $\tilde{\mathbf{S}}_{\mathbf{g}|\mathrm{H}_i}^{2000}$ for the ith class. The second synthetic ensemble, F-MVNEQ, was modeled by generating 2000 feature vectors for each class using an MVN distributed random number generator, where the parameters of the MVN distribution were $\tilde{\mathbf{g}}^{2000}_{\mathbf{H}_i}$ and $\frac{1}{2}\left[\tilde{\mathbf{S}}^{2000}_{\mathbf{g}|\mathbf{H}_1} + \tilde{\mathbf{S}}^{2000}_{\mathbf{g}|\mathrm{H}_2}\right]$ for the ith class. These two ensembles represented the cases of MVN data with unequal and equal covariance matrices, respectively.

We generated multiple realizations of each of the above ensembles of feature vectors in order to provide estimates of the precision of the AUC or correlation coefficient, and thus allow computation of the MSE. For F-MPS, we generated 1000 bootstrap samples by drawing random samples of size 2000 (with replacement) from the full set of the available 2000 feature vectors of each class. For the F-MVNUNEQ and F-MVNEQ, we repeated the process of generating feature vectors to generate 1000 ensembles of 2000 feature vectors for each class.

A major focus of this work was to investigate the performance of the various methods when using small ensembles of feature vectors. For each of the three full ensembles described above, we created a number of smaller ensembles. To do this, we selected the first n samples from each of the 1000 repetitions, where the ensemble size (number of samples per class) was $n \in \{20, 30, 40, 50, 70, 100, 150, 200, 500, 1000, \text{and } 2000\}$.

3.5. Resampling schemes and model observers

In this work, we investigated whether the accuracy and precision of task performance for an observer operating on an ensemble of feature vectors depended on the resampling scheme used to generate the test statistics. In this context, a resampling scheme is the method for selecting feature vectors used to train (i.e. calculate the 1st and 2nd order statistics) and test (i.e. apply the observer to obtain a set of test statistics) the observer. In this work, we used the HT/HT and the LOO (Fukunaga 1990) schemes. These resampling schemes have been used in nuclear medicine, where the CHO was used with the HT/HT scheme in (He *et al* 2004, 2006, 2010) and with the LOO scheme in Sgouros *et al* (2011) and Ghaly *et al* (2015).

In the first resampling scheme used, the HT/HT, assume a dataset of size $2n$, where n is the number of feature vectors per class. In this case, half of the feature vectors from both classes were used to train the observer by computing the 1st and 2nd order statistics of the data, and the observer was tested using each feature vector in the other half. This resulted in $\frac{n}{2}$ test statistics per class.

In the second resampling scheme, the LOO scheme (Fukunaga 1990), $2n$ experiments were carried out. For each experiment, one feature vector was held out and the remaining $2n - 1$ feature vectors were used to estimate the 1st and 2nd order statistics of the data. Then, the held-out vector was used to compute the corresponding test statistic. By holding out a different vector each time, we obtained n test statistics per class.

Since the feature vectors were used by the observers to get the corresponding test statistics, we refer to the observers as the CHO, CLD, and CQD.

3.6. ROC analysis and comparison of observers

The obtained test statistics were analyzed by the ROC-kit software package to estimate the AUC values (Metz *et al* 1998). For each observer, the mean AUC value for an ensemble size of 2000 samples (i.e. feature vectors) per class served as a gold standard for that observer. Since we are interested in the performance of the observers as a function of ensemble size n, we computed the mean square error (MSE) of the AUCs defined as:

$$\mathrm{MSE}_n^J = \frac{1}{1000}\sum_{i=1}^{1000}\left[\overline{\mathrm{AUC}}^J_{2000} - \mathrm{AUC}^J_n(i)\right]^2, \tag{7}$$

where $\overline{\mathrm{AUC}}_{2000}^{J}$ was the mean AUC for the Jth observer over the 1000 bootstrap repetitions at 2000 samples per class (i.e. the gold standard), and $\mathrm{AUC}_{n}^{J}(i)$ was the AUC for the Jth observer at n samples per class from the ith bootstrap repetition.

We also compared the different combinations of observers and resampling schemes based on the mean Spearman's rank correlation coefficient R, defined as (Daniel 1990):

$$R_{n}^{J} = \frac{1}{1000}\sum_{i=1}^{1000} R\left\{\overline{\mathrm{AUC}}_{2000}^{J}, \mathrm{AUC}_{n}^{J}(i)\right\}, \tag{8}$$

where R_{n}^{J} was the rank correlation coefficient for the Jth observer at n samples per class and $R\left\{\overline{\mathrm{AUC}}_{2000}^{J}, \mathrm{AUC}_{n}^{J}(i)\right\}$ was the rank correlation coefficient between the gold standard and the AUC for the Jth observer at n samples per class and from the ith bootstrap repetition. A mean rank correlation coefficient closer to 1 implies a greater ability to correctly rank the performance for the various cut-offs on average. A smaller standard error (defined as the standard deviation/$\sqrt{1000}$) of the rank correlation coefficient implies that there is a good estimation of the mean of the rank correlation coefficient over the ensembles.

3.7. Comparison between the combinations [CHO, LOO] and [CLD, LOO]

To understand the behavior of the two combinations [CHO, LOO] and [CLD, LOO], we conducted the following study using the full set of the available 2000 feature vectors per class of the F-MPS ensemble.

Step 1: Calculate the test statistics. From the available 2000 feature vectors per class, we held out one feature vector from each class and randomly drew 1999 samples with replacement from the remaining 1999 feature vectors of each class. This random sampling was done 1000 times. From each of the 1000 repetitions, we selected the first $n-1$ samples from each class. Each observer was trained using the selected $n-1$ samples from one class and n samples (i.e. the selected $n-1$ samples and the held-out sample) from the other class and then tested with the remaining held-out sample. By holding out a different sample each time, we obtained 2000 test statistics for each class. We tried two cases: $n = 20$ and 2000. This process was repeated 1000 times.

Step 2: Evaluate the differences between observers. The root mean square difference (RMSD) in the test statistics gave an indication about the difference between the test statistics estimated using small ensemble size (i.e. 19 samples from one class and 20 samples from the other class were used for training) and the estimated test statistics using large ensemble size (i.e. 1999 samples from one class and 2000 samples from the other class were used for training). The RMSD in the test statistics $\mathrm{RMSD}(t_i)$ for an observer under the ith hypothesis was calculated from the combined data as the square root of the average of the squared difference between the estimated test statistics using the small ensemble size and the large ensemble size over all 2000 test statistics and 1000 repetitions. The RMSD is given mathematically by the following:

$$\mathrm{RMSD}(t_i) = \sqrt{\frac{1}{2000\,000}\sum_{j=1}^{1000}\sum_{k=1}^{2000}\left(t_i^{\mathrm{large}}(j,k) - t_i^{\mathrm{small}}(j,k)\right)^2}, \tag{9}$$

where $t_i^{\mathrm{large}}(j,k)$ and $t_i^{\mathrm{small}}(j,k)$ represent the kth test statistic from the jth bootstrap index, under the ith hypothesis, calculated from the large and small ensemble sizes, respectively. For the convenience of the reader, the list of abbreviations used and the corresponding descriptions are provided in table 1.

Table 1. A summary of the abbreviations and the corresponding description.

Abbreviation	Description
CHO	Channelized Hotelling observer
CLD	Channelized linear discriminant
CQD	Channelized quadratic discriminant
HT/HT	Half train/half test resampling scheme
LOO	Leave-one-out resampling scheme
MVN	Multivariate normal
MPS	Myocardial perfusion SPECT
F-MPS	Ensemble of feature vectors generated from realistic MPS data
F-MVNEQ	Ensemble of feature vectors generated from MVN distribution with the same covariance matrix under both hypotheses
F-MVNUNEQ	Ensemble of feature vectors generated from MVN distribution with unequal covariance matrices under both hypotheses
AUC	Area under the ROC curve
MSE	Mean square error of the estimated AUC values

4. Results

4.1. Validation of the number of samples used for gold standard

In this work, it was assumed that the mean AUC value for an ensemble size of 2000 samples per class as the gold standard. Thus, it was important to validate that using 2000 samples per class was sufficient. For this purpose, different ensemble sizes, ranging from 50 to 2000, were used to train and test each of the three observers described in section 2. This process was repeated for the 1000 bootstrap repetitions. Then, the means from the 1000 bootstrap AUC values were calculated for each obsever. The AUC values as a function of ensemble size from the F-MPS ensemble at the lowest, middle, and highest cut-off frequencies are shown in figure 3. For all six combinations of observers and resampling methods, the estimated AUC values appeared to converge as the ensemble size increased. This observation was true for the other two ensembles: F-MVNEQ and F-MVNUNEQ (not shown). To quantify convergence, we computed the percentage change in the mean AUC value from an ensemble size of 1000 (i.e. $\overline{\mathrm{AUC}}_{1000}^{J}$) compared to ensemble size of 2000 (i.e. $\overline{\mathrm{AUC}}_{1000}^{J}$) for observer J:

$$\text{Percentage change} = 100 \times \frac{(\overline{\mathrm{AUC}}_{2000}^{J} - \overline{\mathrm{AUC}}_{1000}^{J})}{\overline{\mathrm{AUC}}_{2000}^{J}}. \tag{10}$$

For F-MPS, the percentage change in the mean between AUCs at ensemble sizes 1000 and 2000 was less than 0.3% for all observers and resampling schemes.

In order to determine how variable the mean of the AUC was, we drew random resamples of size $n_{\text{resample}} = 1000$ with replacement from the available 1000 AUCs. This process was repeated 1000 times. For each of these repetitions, we calculated the mean of the AUC. The standard deviation of the estimated means was computed to measure the variability in the mean of the AUC. The standard deviation of the mean of the AUC was much smaller (less than or equal to ~4×10^{-3}) relative to the mean. This observation was true for all 7 cut-offs, 11 ensemble sizes, 3 observers, 2 resampling schemes, and 3 ensembles.

The combination of the small change in the mean AUC from 1000 to 2000 and the small standard deviations of the mean AUCs justified the use of the AUC value computed from 2000 samples from each class as a gold standard for judging the AUCs estimated from smaller ensemble sizes.

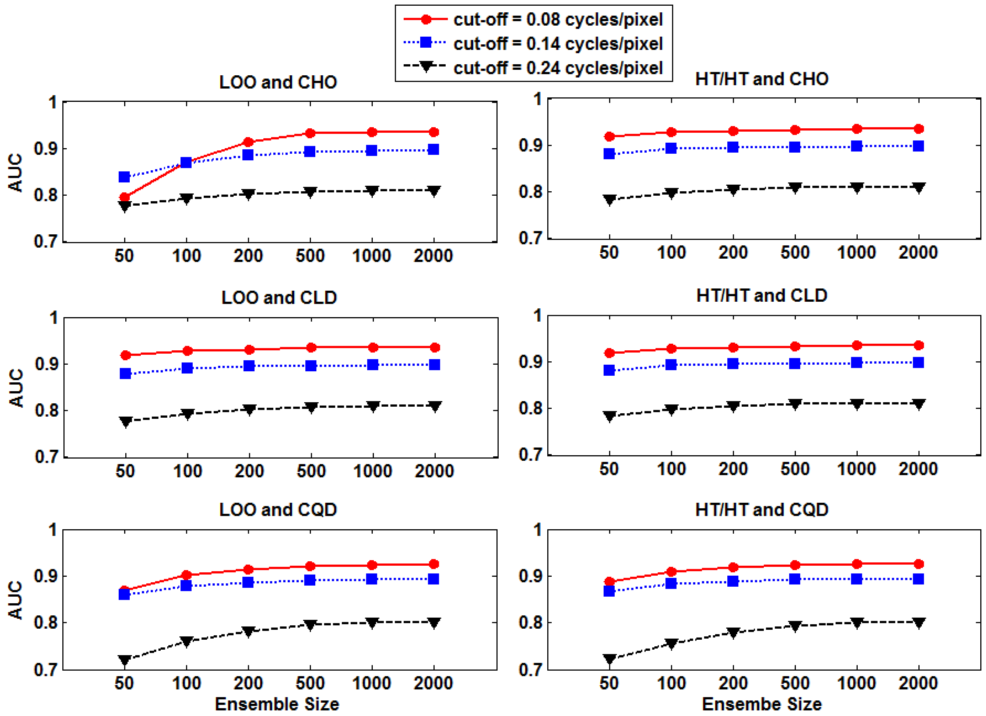

Figure 3. AUC values obtained for different combinations of observers and resampling schemes as functions of ensemble size (i.e. number of samples/class) are shown. The AUC plots represent the mean of 1000 bootstrap repetitions using the F-MPS ensemble.

4.2. Effect of ensemble size on observer performance

In this section, we studied the effect of ensemble size on the performance of the observers for each of the three ensembles: F-MPS, F-MVNEQ, and F-MVNUNEQ.

4.2.1. Effect of ensemble size on observer performance using the F-MPS ensemble. The estimated AUC values as functions of the cut-off frequency of the post-reconstruction filter using the F-MPS ensemble are shown in figure 4. The obtained AUCs for the three observers using both resampling techniques were similar for large ensemble sizes (i.e. 2000 samples/class). This was true for all the cut-off frequencies. However, the performances of the observers diverged as the ensemble size decreased. The smaller ensemble sizes resulted in negatively biased AUCs for all six combinations. Using the HT/HT scheme, the CHO and the CLD gave the same AUCs. However, the performance of CHO and CLD was different for small ensembles when the LOO scheme was used.

Figure 5 shows the MSE of the estimated AUCs as functions of the ensemble size at the middle cut-off frequency (i.e. 0.14 cycles/pixel). Using the LOO scheme at ensemble sizes smaller than 50, the CLD provided the smallest MSE, followed by the CQD, and the largest MSE was obtained when the CHO was used. Figure 6 shows the MSE as a function of the cut-off frequency for an ensemble size of 20 samples/class. The combinations that provided the smallest and almost constant MSEs for the different frequencies were the [CLD,LOO], the [CHO, HT/HT], and the [CLD,HT/HT]. Using the LOO scheme with the CHO for an

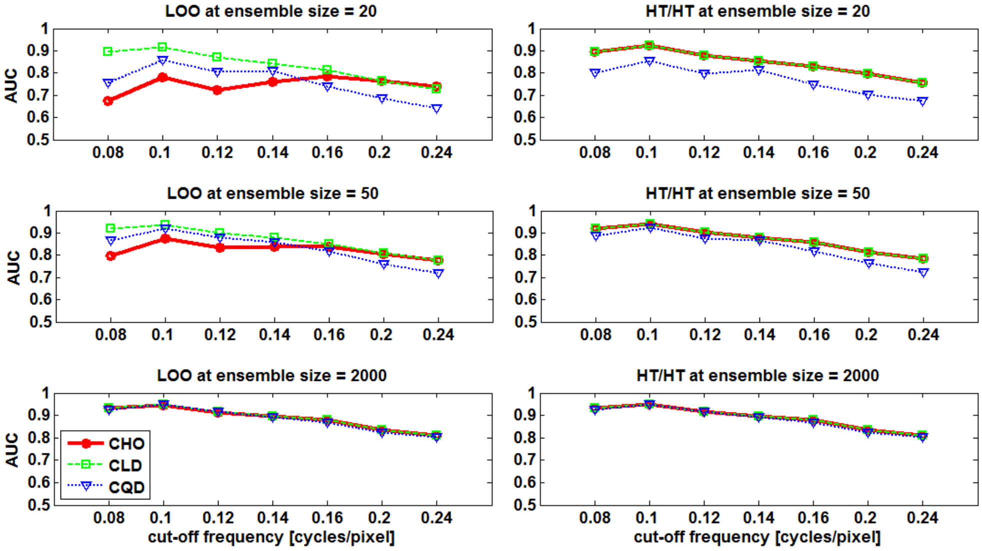

Figure 4. The estimated mean AUC values as functions of the cut-off frequency of the post-reconstruction filter using the F-MPS ensemble are shown. The plots are for six different combinations of observers and resampling schemes for various ensemble sizes.

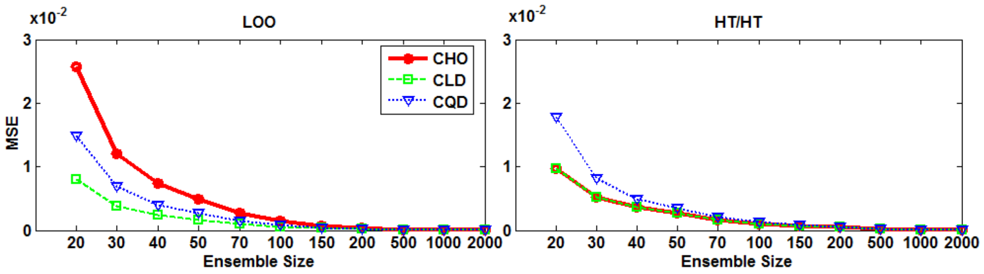

Figure 5. The MSEs of the estimated AUC values using the F-MPS ensemble as functions of the ensemble size for a cut-off of 0.14 cycles/pixel are shown.

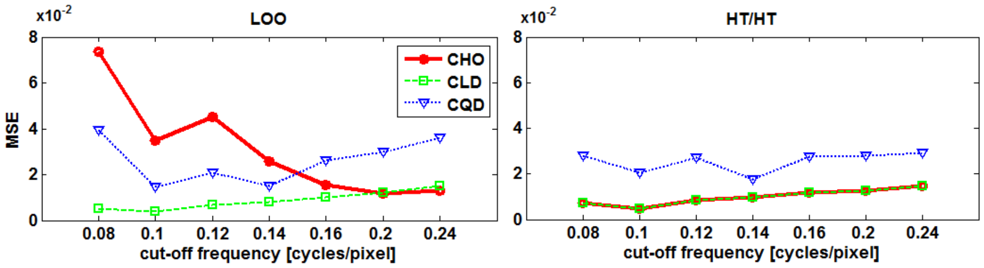

Figure 6. The MSEs of the estimated AUC values using the F-MPS ensemble as functions of the cut-off frequency for an ensemble size of 20 samples/class are shown.

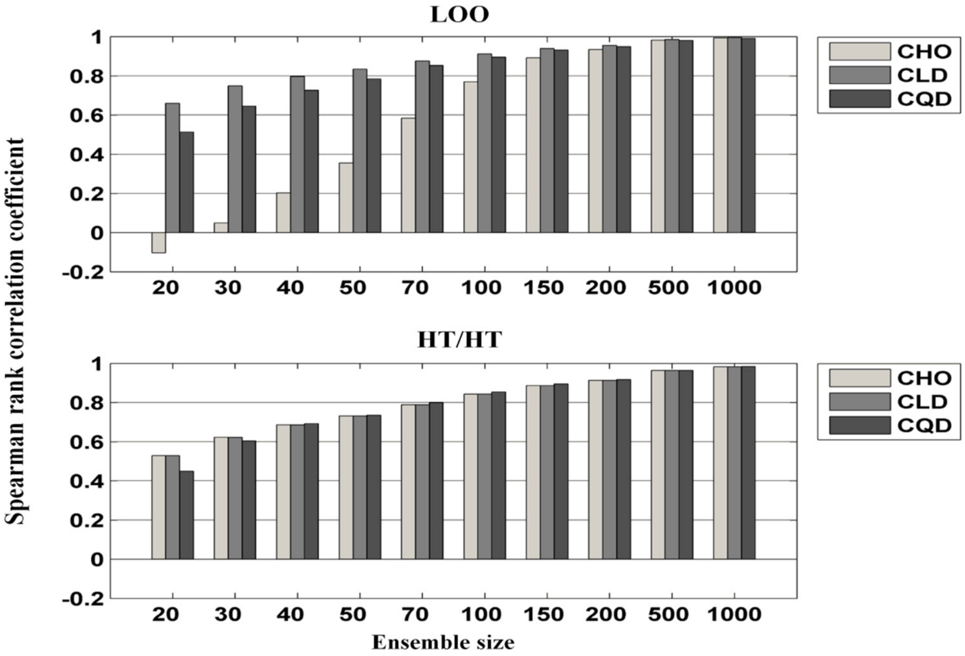

Figure 7. The Spearman's rank correlation coefficients of the AUCs as functions of the ensemble size using the F-MPS ensemble are shown. The plots represent the mean of the 1000 bootstrap repetitions. The standard error was approximately on the order of magnitude of 10^{-4} to 10^{-2} and is thus not displayed.

ensemble size of 20 gave MSE values that varied from ~1% to ~7%. When the MSE for different cut-offs at small ensembles was near constant, this implied that the ranking of the cut-offs was less affected by the small ensemble size. The performance rankings for the filter cut-offs, measured by the Spearman's rank correlation coefficient, R, are shown in figure 7. Using an ensemble size $\geqslant$200 resulted in R values close to one (i.e. larger than 0.91) for all observers and resampling schemes. For smaller ensemble sizes, the rankings of the cut-offs frequencies were presereved best by the combination [CLD, LOO], followed by [CQD, LOO].

4.2.2. Effect of ensemble size on observer performance using the F-MVNEQ ensemble. Figures 8–11 show the results for the F-MVNEQ ensemble. The observers had similar performances for both F-MPS and F-MVNEQ ensembles for most combinations and ensemble sizes.

4.2.3. Effect of ensemble size on observer performance using the F-MVNUNEQ ensemble. Figures 12–15 show the results for the F-MVNUNEQ ensemble. At ensemble size of 2000, the CQD outperformed the CHO and the CLD for some cut-off frequencies and for both resampling schemes as shown in figure 12. For smaller ensemble sizes, the observers had almost similar performances to those from the F-MPS and F-MVNEQ ensembles.

4.3. Comparison between the combinations [CHO, LOO] and [CLD, LOO]

The RMSD of the estimated test statistics for both the CHO and the CLD are shown in figure 16. The RMSD using the CHO was much larger than that using the CLD, especially for lower cut-off frequencies. This observation was true for both the defect-absent and

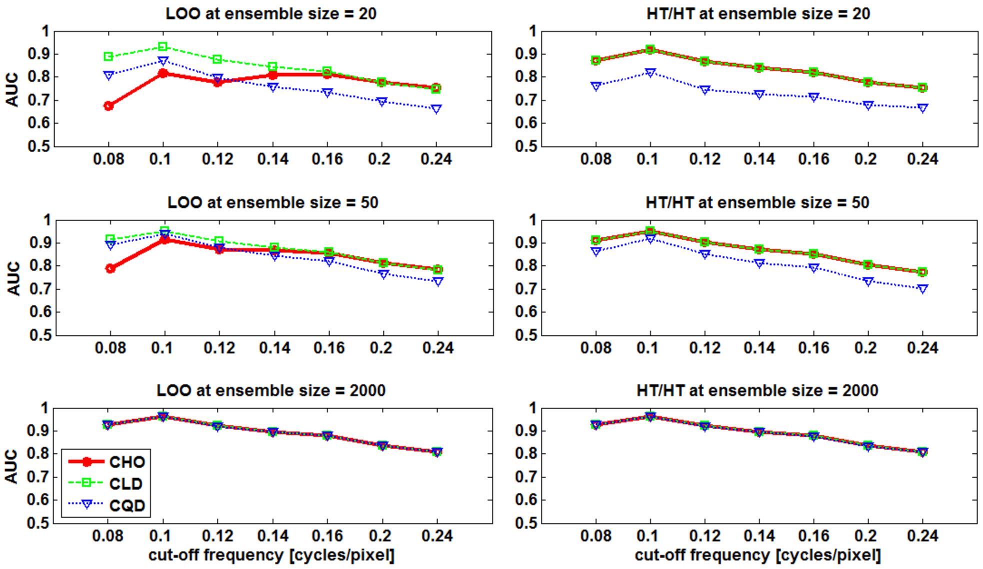

Figure 8. The estimated mean AUC values as functions of the cut-off frequency of the post-reconstruction filter using the F-MVNEQ ensemble are shown. The plots are for six different combinations of observers and resampling schemes for various ensemble sizes.

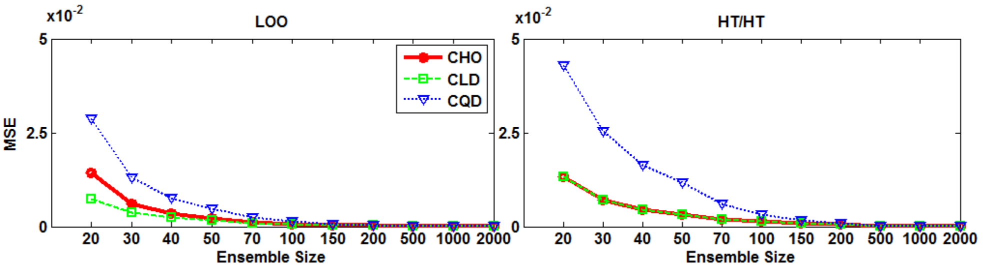

Figure 9. The MSEs of the estimated AUC values using the F-MVNEQ ensemble as functions of the ensemble size for a cut-off of 0.14 cycles/pixel are shown.

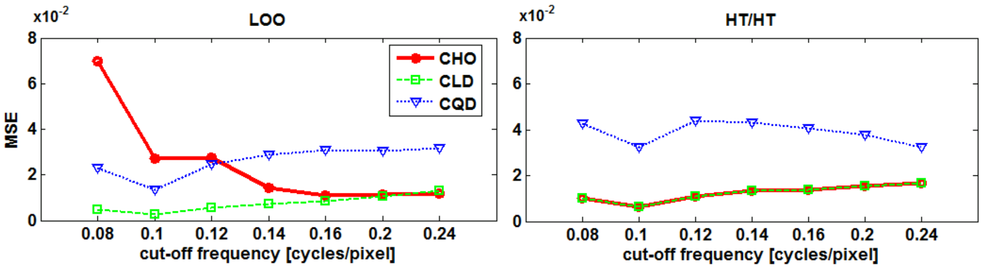

Figure 10. The MSEs of the estimated AUC values using the F-MVNEQ ensemble as functions of the cut-off frequency for an ensemble size of 20 samples/class are shown.

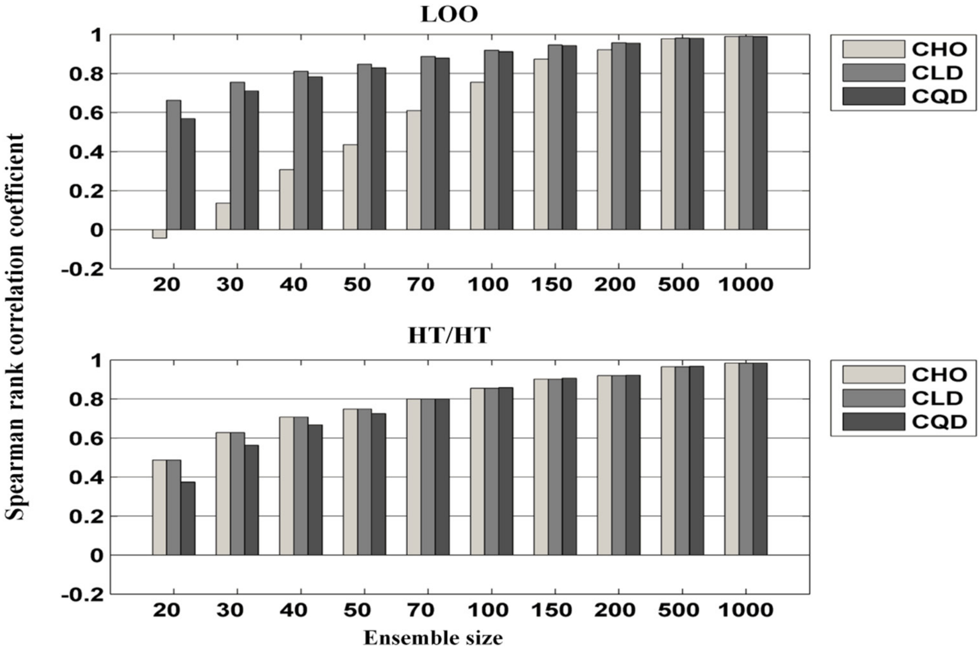

Figure 11. The Spearman's rank correlation coefficients of the AUCs as functions of the ensemble size using the F-MVNEQ ensemble are shown. The plots represent the mean of the 1000 bootstrap repetitions. The standard error was approximately on the order of magnitude of 10^{-4} to 10^{-2} and is thus not displayed.

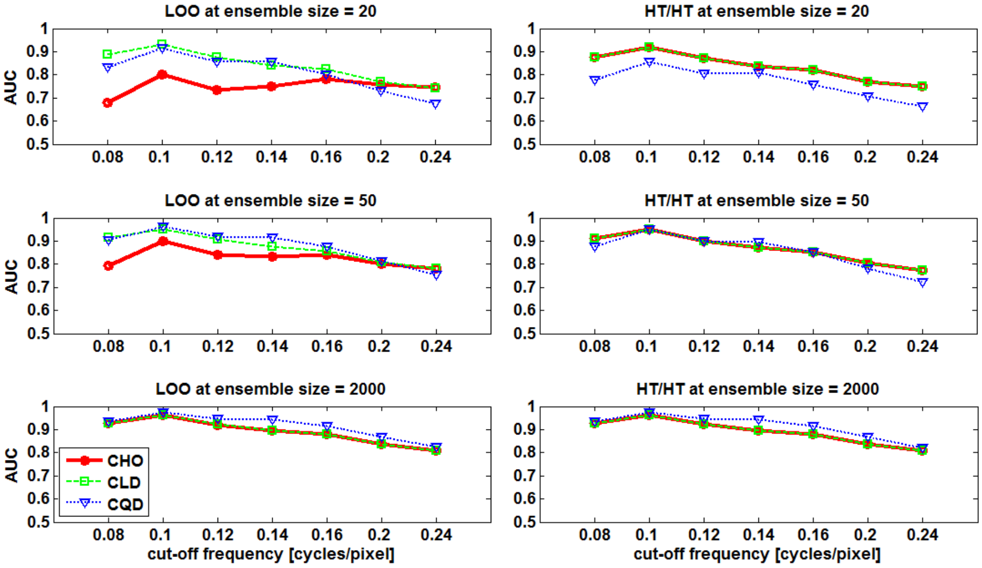

Figure 12. The estimated mean AUC values as functions of the cut-off frequency of the post-reconstruction filter using the F-MVNUNEQ ensemble are shown. The plots are for six different combinations of observers and resampling schemes for various ensemble sizes.

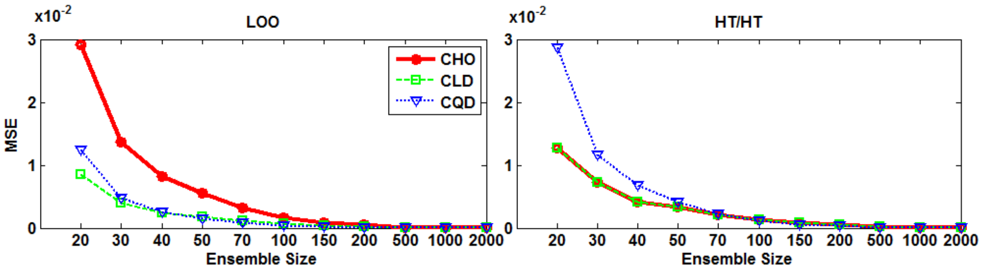

Figure 13. The MSEs of the estimated AUC values using the F-MVNUNEQ ensemble as functions of the ensemble size for a cut-off of 0.14 cycles/pixel are shown.

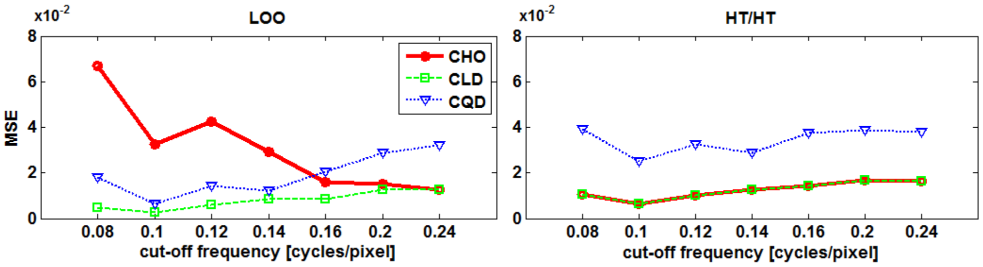

Figure 14. The MSEs of the estimated AUC values using the F-MVNUNEQ ensemble as functions of the cut-off frequency for an ensemble size of 20 samples/class are shown.

defect-present classes. This indicates that the RMSD in the test statistics estimated from the smaller ensemble was larger for the CHO than for the CLD when using the LOO resampling scheme. This additional error in test statistic values can explain the large difference in the estimated AUCs between the CHO and CLD at lower cut-offs, when small ensemble sizes were used (see figure 4).

5. Discussion

It has been assumed in the literature that the performance of the CHO and the CLD are the same (Myers and Barrett 1987) because the difference between the test statistics from these two observers is the extra term Δ defined in equation (6), which is independent of the test data. In this work, it was demonstrated that the use of the CLD with the described LOO scheme has major advantages in the case of a binary classification task when small ensembles of MPS images, with only uptake variability, were used.

The performances of all observers (as measured by the AUC values) were similar at large ensemble sizes, except that the CQD outperformed the CHO and CLD for the F-MVNUNEQ data at some cut-offs. To explain this, recall that the CQD will have optimal performance (i.e. equal to the performance of the IO) when the data from the two classes are MVN distributed with unequal covariance matrices, which was the case of F-MVNUNEQ data at large ensemble size (Chan *et al* 1999). The CLD and the CHO will be suboptimal in the case of F-MVNUNEQ because of the unequal covariance matrices.

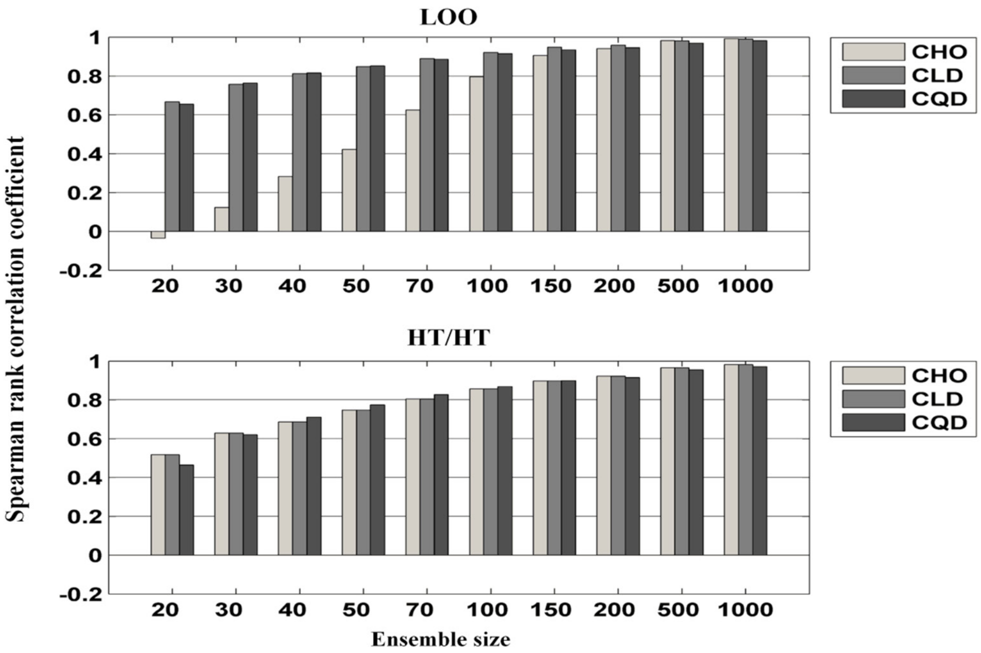

Figure 15. The Spearman's rank correlation coefficients of the AUCs as functions of the ensemble size using the F-MVNUNEQ ensemble are shown. The plots represent the mean of the 1000 bootstrap repetitions. The standard error was approximately on the order of magnitude of 10^{-4} to 10^{-2} and is thus not displayed.

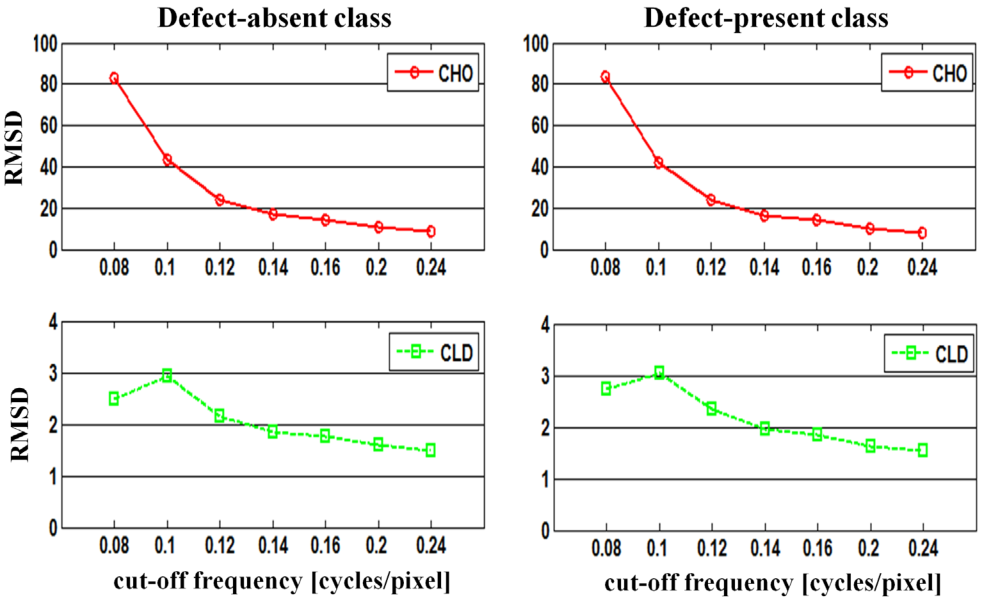

Figure 16. The RMSD of the estimated test statistics using the F-MPS ensemble are shown. Note that the vertical scale is smaller by a factor of 25 for the CLD compared to the CHO.

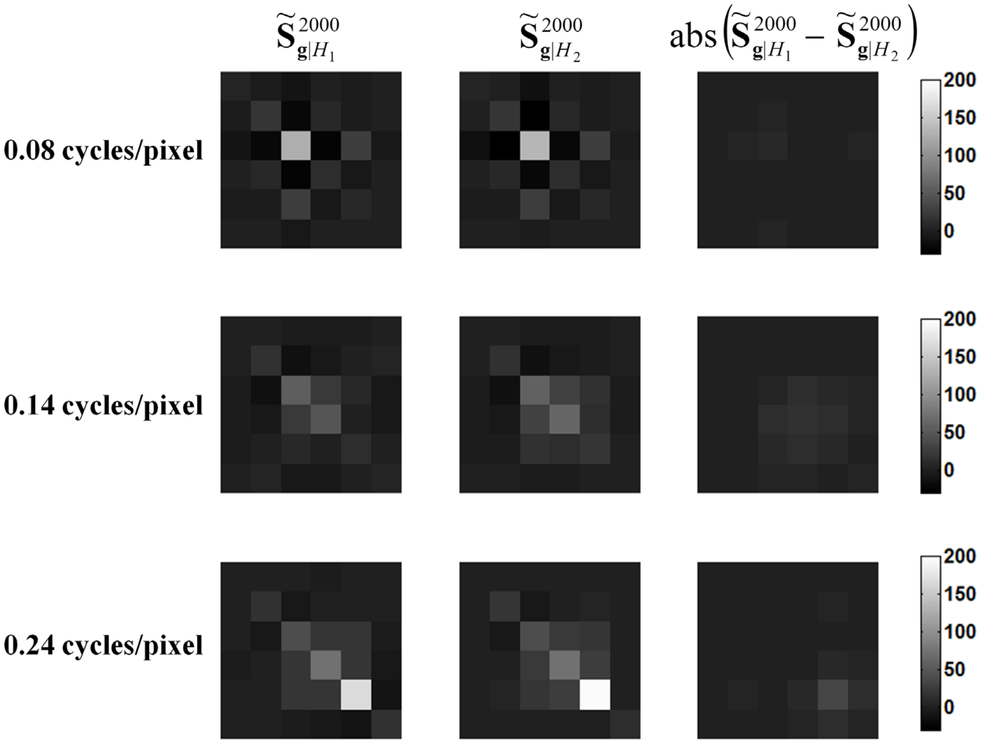

Figure 17. Images of the covariance matrices for the defect-absent (left column) and defect-present (middle column) classes, and images of the absolute difference between the covariance matrices (right column) are shown.

Figure 17 shows images of the covariance matrices for both classes, and images of their absolute differences at cut-offs 0.08, 0.14, and 0.24 cycles/pixel. It is observed from figure 17 that the covariance matrices of the two classes had different structures and this difference in structure changed as a function of the cut-off frequency. This could explain why the CQD had higher AUC values than the CHO and CLD for some cut-offs at large ensemble size (see figure 12, bottom row).

The results reported in section 4 showed that, in general, the CQD required larger ensemble size than that required by the CLD. Our observations were consistent with previous findings in Marks and Dunn (1974), Wahl and Kronmal (1977) and Fukunaga (1990).

Although we have used the rotationally symmetric channels with square profiles, the analysis and principles developed in this work can be applied to other channel types. However, further experiments are necessary to draw conclusions about the different channel models, especially for the case of [CLD, LOO] and [CHO, LOO].

In this work, the goal was to find a strategy (i.e. an observer and a resampling scheme) that provides an improved precision for small ensemble sizes. Thus, we used an ensemble size of 2000 samples/class as our gold standard and we did not model internal noise in the observer. To match the human observer performance, internal noise can be added (Brankov 2013). The work presented in this paper could be extended to include internal noise in the test statistics calculations. We anticipate that the addition of internal noise can change the underlying AUC values, but the precision of the AUC estimates may not be much affected.

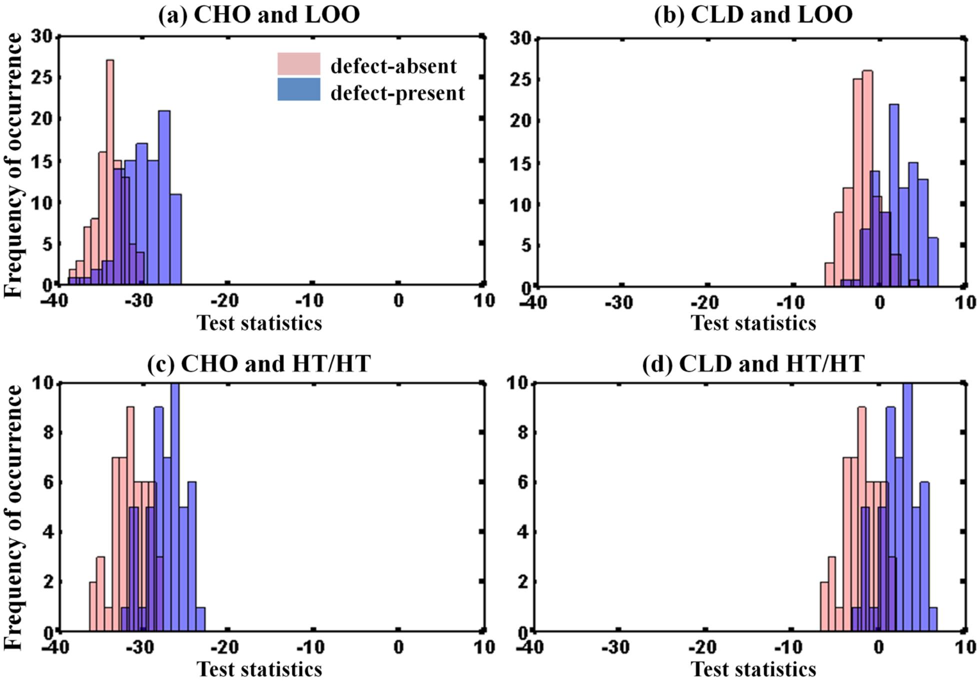

Figure 18. Histograms of the test statistics of the CHO and the CLD using both resampling schemes for the F-MPS ensemble using 100 samples/class are shown. (a) CHO and LOO, (b) CLD and LOO, (c) CHO and HT/HT and (d) CLD and HT/HT.

5.1. Comparison between the CHO and CLD

As the ensemble size decreased, the AUC values became more negatively biased for all observers and data distributions. It was observed that the CLD observer trained and tested using the described LOO scheme gave better performance for small ensemble sizes, regardless of the distribution of the data (see figures 4–15). In other words, this combination better preserved the AUC values because the MSE was small and almost constant over the different frequencies (see figures 5, 6, 9, 10, 13 and 14). Consequently, the rankings of the cut-off frequencies were better preserved (see figures 7, 11 and 15).

Although the CHO and the CLD are both linear classifiers and the only difference between them is the extra term Δ defined in equation (6), their numerical behavior depends on the training and testing scheme. Figure 18 shows the histograms of the test statistics using 100 samples/class. It was observed that the extra term in the CLD affected the shape of the distribution of the test statistics compared to that of the CHO, for the LOO scheme (see figures 18(a) and (b)). However, the same was not true for the HT/HT scheme (see figures 18(c) and (d)). The term Δ depends on the mean vectors and covariance matrices of the data computed during the training phase. For the HT/HT scheme, this term was computed once from half the available samples and then used to calculate the test statistics for the remaining samples. Thus, the distribution of the test statistics using the CLD will be a shifted version of that obtained using the CHO. For the LOO scheme, the extra term was computed for each of the $2n$ experiments (as previously described in section 3.5). This would result in $2n$ different values of Δ and each test statistic was thus calculated using a different Δ. Thus, the distribution of the test statistics using the CLD would not be only a shifted version of that obtained using CHO, but the shape of the distribution may also be different.

6. Conclusions

In this work, we assessed the performance of three channelized model observers, the CHO, CLD, and CQD, using two resampling schemes, LOO and half train/half test (HT/HT), for different ensemble sizes.

For the task considered, the results showed that the combination [CLD, LOO] better preserved the performance rank for small ensembles, followed by either the [CQD, LOO], the [CHO, HT/HT] or [CLD, HT/HT]. The combination [CHO, LOO] had the worst performance both in terms of preserving AUC and performance rank for small ensemble size. The performance of CHO and CLD were the same when HT/HT resampling was used, as expected. The combination [CQD, HT/HT] had a higher MSE than the [CHO, HT/HT] and [CLD, HT/HT] combinations for the small ensembles, likely reflecting the larger number of parameters that must be estimated from the training set. These observations held for all three datasets investigated.

The results of this study suggest that the CLD combined with the LOO scheme is better able to handle small ensemble sizes for SKE/BKS task evaluation than the other methods investigated. The use of this combination has the potential to provide more statistical power for a given number of images, and thus allow for the study of a larger range of parameters in optimization studies.

Acknowledgments

This work was supported by National Institutes of Health grants R01 EB016231 and R01 EB013558. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID iDs

Michael Ghaly https://orcid.org/0000-0002-3043-8105

References

Barrett H H and Myers K J 2004 *Foudations of Image Science* (New York: Wiley)

Barrett H H, Yao J, Rolland J P and Myers K J 1993 Model observers for assessment of image quality *Proc. Natl Acad. Sci. USA* **90** 9758–65

Brankov J G 2013 Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection *J. Phys. Med. Biol.* **58** 7159–82

Chan H P, Sahiner B, Wagner R F and Petrick N 1999 Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers *Med. Phys.* **26** 2654–68

Daniel W W 1990 *Applied Nonparametric Statistics* (Boston, MA: Cengage Learning)

Elshahaby F E A, Ghaly M, Jha A K and Frey E C 2015a The effect of signal variability on the histograms of anthropomorphic channel outputs: factors resulting in non-Normally distributed data *Proc. SPIE* **9416** 94160P–94160P-6

Elshahaby F E A, Ghaly M, Jha A K and Frey E C 2016 Factors affecting the normality of channel outputs of channelized model observers: an investigation using realistic myocardial perfusion SPECT images *J. Med. Imaging* **3** 015503

Elshahaby F E A, Ghaly M, Li X, Jha A K and Frey E C 2015b Estimating model observer performance with small image ensembles *J. Nucl. Med.* **56** 540

Frey E C, Gilland K L and Tsui B M 2002 Application of task-based measures of image quality to optimization and evaluation of three-dimensional reconstruction-based compensation methods in myocardial perfusion SPECT *IEEE Trans. Med. Imaging* **21** 1040–50

Fukunaga K 1990 *Introduction to Statistical Pattern Recognition* (New York: Academic)

Fukunaga K and Hayes R R 1989a Effects of sample size in classifier design *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 873–85

Fukunaga K and Hayes R R 1989b Estimation of classifier performance *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 1087–101

Ge D, Zhang L, Cavaro-Ménard C and Gallet P L 2014 Numerical stability issues on channelized Hotelling observer under different background assumptions *J. Opt. Soc. Am.* A **31** 1112–6

Ghaly M, Du Y, Fung G S, Tsui B M, Links J M and Frey E 2014 Design of a digital phantom population for myocardial perfusion SPECT imaging research *Phys. Med. Biol.* **59** 2935–53

Ghaly M, Du Y, Links J M and Frey E C 2016 Collimator optimization in myocardial perfusion SPECT using the ideal observer and realistic background variability for lesion detection and joint detection and localization tasks *Phys. Med. Biol.* **61** 2048–66

Ghaly M, Links J M and Frey E C 2015 Optimization of energy window and evaluation of scatter compensation methods in myocardial perfusion SPECT using the ideal observer with and without model mismatch and an anthropomorphic model observer *J. Med. Img.* **2** 015502

Gifford H C, King M A, De Vries D J and Soares E J 2000 Channelized hotelling and human observer correlation for lesion detection in hepatic SPECT imaging *J. Nucl. Med.* **41** 514–21

Gilland K L, Tsui B M W, Qi Y and Gullberg G T 2006 Comparison of channelized hotelling and human observers in determining optimum OS-EM reconstruction parameters for myocardial SPECT *IEEE Trans. Nucl. Sci.* **53** 1200–4

He X, Frey E C, Links J M, Gilland K L, Segars W P and Tsui B M 2004 A mathematical observer study for the evaluation and optimization of compensation methods for myocardial SPECT using a phantom population that realistically models patient variability *IEEE Trans. Nucl. Sci.* **51** 218–24

He X, Links J M and Frey E C 2010 An investigation of the trade-off between the count level and image quality in myocardial perfusion SPECT using simulated images: the effects of statistical noise and object variability on defect detectability *Phys. Med. Biol.* **55** 4949–61

He X, Links J M, Gilland K L, Tsui B M and Frey E C 2006 Comparison of 180 degrees and 360 degrees acquisition for myocardial perfusion SPECT with compensation for attenuation, detector response, and scatter: Monte Carlo and mathematical observer results *J. Nucl. Cardiol.* **13** 345–53

Kupinski M A, Clarkson E and Hasterman J Y 2007 Bias in hotelling observer performance computed from finite data *Proc. Soc. Photo Opt. Instrum. Eng.* **6515** 1–7

Li X, Jha A K, Ghaly M, Elshahaby F E A, Links J M and Frey E C 2017 Use of sub-ensembles and multi-template observers to evaluate detection task performance for data that are not multivariate normal *IEEE Trans. Med. Imaging* **36** 917–29

Marks S and Dunn O J 1974 Discriminant functions when covariance matrices are unequal *J. Am. Stat. Assoc.* **69** 555–9

Metz C E, Herman B A and Shen J-H 1998 Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data *Stat. Med.* **17** 1033–53

Myers K J and Barrett H H 1987 Addition of a channel mechanism to the ideal-observer model *J. Opt. Soc. Am.* A **4** 2447–57

Park S, Clarkson E, Kupinski M A and Barrett H H 2005a Efficiency of human and model observers for signal-detection tasks in non-Gaussian distributed lumpy backgrounds *Proc. SPIE* **5749** 138–49

Park S, Clarkson E, Kupinski M A and Barrett H H 2005b Efficiency of the human observer detecting random signals in random backgrounds *J. Opt. Soc. Am.* A **22** 3–16

Sahiner B, Chan H P and Hadjiiski L 2008 Classifier performance prediction for computed-aided diagnosis using a limited dataset *Med. Phys.* **34** 1559–70

Sankaran S, Frey E C, Gilland K L and Tsui B M 2002 Optimum compensation method and filter cutoff frequency in myocardial SPECT: a human observer study *J. Nucl. Med.* **43** 432–8

Sgouros G, Frey E C, Bolch W E, Wayson M B, Abadia A F and Treves S T 2011 An approach for balancing diagnostic image quality with cancer risk: application to pediatric diagnostic imaging of 99mTc-dimercaptosuccinic acid *J. Nucl. Med.* **52** 1923–9

Tseng H W, Fan J and Kupinski M A 2016 Design of a practical model-observer-based image quality assessment method for x-ray computed tomography imaging systems *J. Med. Imaging* **3** 035503

Wahl P W and Kronmal R A 1977 Discriminant functions when covariances are unequal and sample sizes are moderate *Biometrics* **33** 479–84

Wollenweber S D, Tsui B M W, Lalush D S, Frey E C, Lacroix K J and Gullberg G T 1999 Comparison of Hotelling observer models and human observers in defect detection from myocardial SPECT imaging *IEEE Trans. Nucl. Sci.* **46** 2098–103

Wunderlich A and Noo F 2009 Estimation of channelized hotelling observer performance with known class means or known difference of class means *IEEE Trans. Med. Imaging* **28** 1198–207

Yihuan L, Lin C and Gene G 2014 Collimator performance evaluation for In-111 SPECT using a detection/localization task *Phys. Med. Biol.* **59** 679– 96