Dartmouth College

Dartmouth Digital Commons

Open Dartmouth: Published works by
Dartmouth faculty

Faculty Work

9-1-2007

Receiver Operating Characteristic and Location Analysis of Simulated Near-Infrared Tomography Images

Xiaomei Song
Dartmouth College

Brian W. Pogue
Dartmouth College

Hamid Dehghani
University of Exeter

Shudong Jiang
Dartmouth College

Keith D. Paulsen
Dartmouth College

See next page for additional authors

Follow this and additional works at: https://digitalcommons.dartmouth.edu/facoa

Part of the Engineering Commons, and the Medicine and Health Sciences Commons

Authors

Xiaomei Song, Brian W. Pogue, Hamid Dehghani, Shudong Jiang, Keith D. Paulsen, and Tor D. Tosteson

Receiver operating characteristic and location analysis of simulated near-infrared tomography images

Xiaomei Song
Brian W. Pogue
Hamid Dehghani
Shudong Jiang
Keith D. Paulsen
Dartmouth College
Thayer School of Engineering
Hanover, New Hampshire 03755

Tor D. Tosteson
Dartmouth Medical School
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire 03756

Abstract. Receiver operating characteristic (ROC) analysis was performed on simulated near-infrared tomography images, using both human observer and contrast-to-noise ratio (CNR) computational assessment, for application in breast cancer imaging. In the analysis, a nonparametric approach was applied for estimating the ROC curves. Human observer detection of objects had superior capability to localize the presence of heterogeneities when the objects were small with high contrast, with a minimum detectable threshold of CNR near 3.0 to 3.3 in the images. Human observers were able to detect heterogeneities in the images below a size limit of 4 mm, yet could not accurately find the location of these objects when they were below 10 mm diameter. For large objects, the lower limit of a detectable contrast limit was near 10% increase relative to the background. The results also indicate that iterations of the nonlinear reconstruction algorithm beyond 4 did not significantly improve the human detection ability, and degraded the overall localization ability for the objects in the image, predominantly by increasing the noise in the background. Interobserver variance performance in detecting objects in these images was low, suggesting that because of the low spatial resolution, detection tasks with NIR tomography is likely consistent between human observers. © 2007 Society of Photo-Optical Instrumentation Engineers.
[DOI: 10.1117/1.2799197]

1 Introduction

Near-infrared (NIR) tomographic image reconstruction methods can be used to recover the distribution of absorber, scatterer, or fluorophore concentrations in living tissue, using noninvasive diffusely transmitted measurements. This provides a diagnostic modality to noninvasively quantify oxygen saturation, hemoglobin concentration, water concentration, scattering, and potentially exogenous chromophores in tissues.[1–6] It has been shown that tumors tend to have a higher level of vascularity and cellular stroma relative to normal tissues[7] due to hyperactive growth and angiogenesis, and this leads to significant contrast in the near-infrared spectrum,[3] or light between 650-, and 850-nm wavelengths. Images of tumors within normal breast tissue have been demonstrated, and clinical trials are ongoing to interpret the potential clinical role of this type of imaging device. The purpose of this study is to provide a preliminary understanding of how NIR images can be used in detection tasks, using observer performance assessment.[8,9] The capabilities of NIR tomography reconstruction algorithms are analyzed, and the results are put into context with how such a system might be used in breast cancer imaging.

Using frequency domain NIR tomography imaging, amplitude modulated light signals are transmitted through tissue to quantify both absorption and scattering images of the breast. The intensity and phase-shift information of the transmitted near-infrared signal provides this information. While several promising NIR reconstruction schemes have been developed and demonstrated in preclinical and clinical studies, little consideration has been given to the measurement of observer performance on the NIR tomographic images of a specific method. NIR tomography offers an interesting test case for such observer studies, in that the reconstruction algorithm is a nonlinear iterative approach that has niche uses in other clinical modalities, but has not been systematically tested for how it affects receiver operating characteristic (ROC) responses and local receiver operating characteristic (LROC) responses. Additionally, one major issue in diffuse tomography, which is problematic for this type of analysis, is that generally there is nonlinear response across the imaging field. Thus, evaluation of this type of iterative-based reconstruction method is complicated because there are many parameters to examine and few systematic studies of how they impact detectability. Following the established methodologies in radiology, observer studies can be carried out to assess the effects of iteration and location of objects in the imaging field. In this study, this

---

Address all correspondence to Brian W. Pogue, Dartmouth College, Thayer School of Engineering, Hanover, NH 03755. Tel: 603-646-3861; Fax: 603-646-3856; E-mail: pogue@dartmouth.edu

evaluation is done using the standard approach to the ROC curve analysis, and the LROC. These are used to study several parameters including: 1. the minimum object size detectable, 2. object-background contrast levels, 3. reconstruction process efficacy, and 4. the confidence level of the outcome from the system. Human observers with and without NIR tomography imaging experience and computational measures based on contrast-to-noise ratio (CNR) were used in this study. CNR is defined as the relative difference between the average property values within the ROI and those within the background region, divided by the average variation in the background. While CNR is a simple objective basis for measuring image quality in an automated way, it is generally an inferior surrogate to human observer or ideal observer estimates.[10,11] Nonetheless, it is quantified here, because if the purpose of the system ultimately shifts from one in which detection of regions is not the goal, but rather quantification of region values is the goal, then knowledge of the CNR does provide a useful measure of the system performance.

The ROC methodology has been widely used to address the clinical efficacy of medical imaging systems.[11–14] In an ROC study, the reader views images, some of which contain single or multiple abnormalities, while the remaining images are normal. The reader then assigns numeric ratings to each image as an indication of their confidence level that the image is abnormal. The resulting rating data are then plotted on an ROC curve, which entails the true-positive fraction plotted against the false-positive fraction, as the positivity criterion is varied across the range of the rating scale. An example of this analysis was recently reported by Chance et al.,[4] using reflectance data from breast tumor measurements, without tomographic reconstruction. In this work, however, this concept is extended for tomography assessment. Summary measures of the curve, including the partial area under the curve in a particular region of interest or the area under the entire curve, are typically used as an objective measure of the ability of the reader to detect objects in the images.[11] The area under the curve must be greater than 0.5 for a greater than 50% chance of detecting objects, and can achieve a maximum of 1.0 for perfect detection ability.

Typically, ROC analysis is a way to represent the image quality of the medical modality for a specific human detection task. In standard ROC methods, the complexity of the target object location is often eliminated by clearly specifying the possible ROI (region of interest) in the images. However, in more complex medical imaging applications where the expected image resolution is spatially dependent, the applicability of standard ROC analysis is very limited. Recent developments in localization-response ROC (LROC) analysis statistically offer more understanding of medical imaging methodology, in terms of measuring the conjoint ability of detecting and correctly localizing the actual targets in medical images. These developments include simultaneous ROC/LROC fitting[9] and alternative free-response ROC (AFROC) analysis.[11,15]

The LROC plots the probability of both detecting and locating objects in images with abnormalities versus the probability of falsely detecting objects in normal images as the detection criteria is varied. Several models have been proposed to be used in LROC analysis, including the discrete-location models[16] and the general detection-localization model;[15] however, both models hold their particular assumptions. In this study, both human readers with or without medical imaging background and a "computational reader" were required to specify the location of the suspicious area of abnormalities, which makes both ROC and LROC techniques applicable. While ROC and LROC analysis have been used in iterative image formation analysis before, they have not been systematically examined in an algorithm that uses iterative refinement of an ill-posed image reconstruction problem. The nature of the physical attenuation of NIR light decreases the sensitivity by over an order of magnitude with every centimeter of penetration into the tissue, and thus the sensitivity matrix contains values that vary by many orders of magnitude. This type of a sensitivity matrix is highly ill-posed, and the images formed in this process are not well characterized in terms of how humans interpret them. This work is the first published systematic study of this type of algorithm.

Despite the essential simplicity of the fundamental concepts of ROC analysis and CNR analysis in medical imaging, professionals designing and performing ROC studies often find that many subtle issues related to experimental design and data analysis must be confronted in practice. Such issues include: 1. case selection, 2. collection, 3. presentation; 4. observer selection and grouping; 5. system error and tolerance analysis; 6. localization of data, 7. the strategy of data analysis and curve fitting, and 8. confidence interval of the test results. All of these issues are considered here in the process of this initial study design, with the goal of developing an efficient and well-understood way to interpret NIR tomography images and the NIR image reconstruction process.

2 Methods

2.1 Near-Infrared Tomography Simulation and Image Preparation

In previous publications, NIR frequency domain absorption and scatter tomography reconstruction has been demonstrated by several researchers, using finite element models of the diffusion approximation, and iterative reconstruction algorithms.[17–21] In the work proposed in this study, the focus is on a particular experimental system design used at Dartmouth, which can produce 2-D or 3-D images of absorption and scatter coefficients of the objects examined by the noninvasive imaging array. The top level of the tomographic system detection interface is shown in Fig. 1, with a circular arrangement of 16 linear translation stages that allow direct contact of the optical fiber bundles to the tissue being imaged. While experimental images are not used in the analysis of this study, the configuration of sources, detectors, and data noise are simulated particularly to match this experimental system, such that future studies could focus on the ROC and LROC analysis of the clinical data generated by this system.

A circular 86-mm-diam field was chosen as the background in which to place objects and test the methodology based on computed images. The finite element mesh used was a 2000 node forward 2-D geometry, which is commonly used for breast image reconstruction, which is an average spatial resolution of 1.7 mm between nodes. This field had an absorption coefficient $\mu_a=0.005$ mm^{-1}, and a reduced scattering coefficient $\mu'_s=1.0$ mm^{-1}, and within this field a spherical object with fixed $\mu'_s=1.0$ mm^{-1} and a variable μ_a was placed
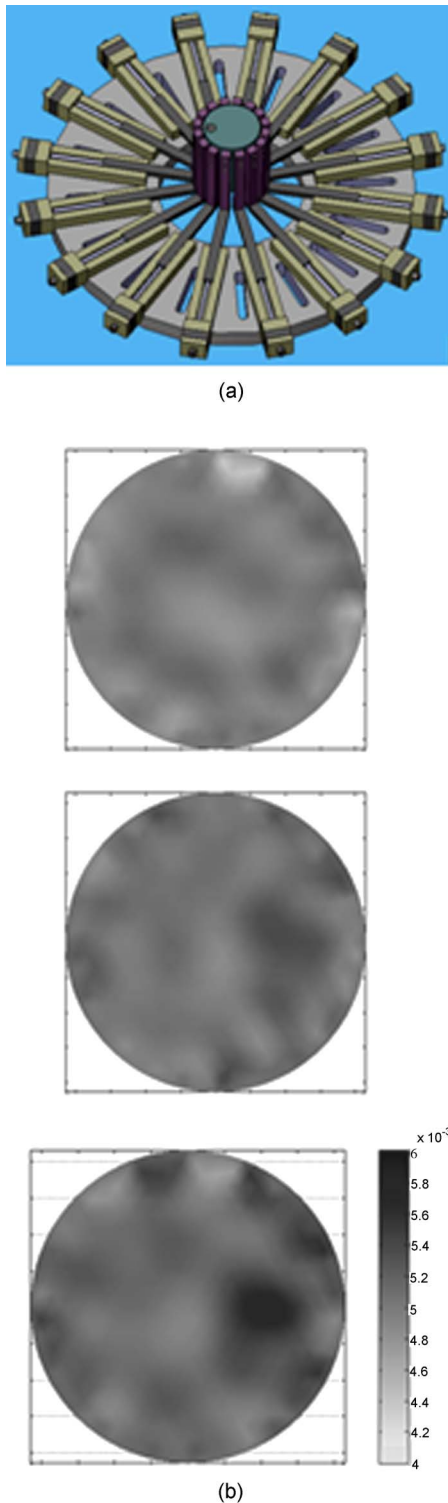
(a)

(b)

Fig. 1 (a) A mechanical model view of the NIR imaging fiber interface is shown, using a set of 16 fiber bundles for imaging the tissue in a circular tomographic geometry, which are all moved in and out on linear translation stages, allowing imaging of different sized breasts in a circular geometry.[5,6] (b) Typical simulated NIR tomography images are shown with the same size and location of ROI, iteration number, and reconstruction algorithm but different contrast levels, which indicates different image quality for a typical reconstructed absorption coefficient image. At the top, the contrast was $C=1.1$; in the middle the contrast was $C=1.4$, and at the bottom the contrast was $C=2.0$.

to provide a localized heterogeneity. The μ_a value of the object was varied to simulate changes in absorption contrast, with values ranging from 1.06:1 (6%) up to to 1.3:1 (30%). The size of the object ranged from 4 to 16 mm in diameter. Forward calculations were based on this specific NIR tomography system geometry and diffusion theory solution using the finite element numerical method, and were used along with zero-mean Gaussian noise of 1% in amplitude and 1 deg in phase shift, to create simulated measurement data. The reconstruction algorithm was then applied to generate the 2-D absorption and scattering coefficient images. In the same manner, thousands of reconstructed images were automatically created, having the same noise level and size of heterogeneities but with different locations of the ROI and different contrasts. Similar images were created with no objects inside, to simulate normal tissue, or the "control" images. The amount and constitution of the reconstructed image pool was designed specially for several different studies presented here, and these contained images of different ROI size (4, 6, 10, and 16 mm) or contrast, or different variations of the reconstruction algorithm, i.e., different iteration numbers of the reconstruction program or different contrast levels between the ROI and the background.

The output of the forward problem, also known as the calculated system measurement, was compared to the measured data and thereafter to determine the quality of the reconstruction process. An important measure of error in the reconstruction process is the objective function defined as:

$$\chi_k^2 = \frac{\|\Phi_M - \Phi_C(\mu^k)\|^2}{M\sigma^2}, \qquad (1)$$

where Φ_M is the measured set of data, $\Phi_C(\mu^k)$ is the calculated measurements at the k'th iteration, M is the number of measurements, and σ is the entire image standard deviation. This is also sometimes referred to as the projection error, as it is a direct measurement of the squared error of measured and simulated projection data through the tissue.

2.2 Human Observer Detection and Localization Tasks

Four human observers participated in this study, of which two observers had the NIR tomography imaging experience and the other two observers did not. A MATLAB program made with a graphical user interface was used as the main test program, and is shown in Fig. 2. This controlled the data collection stage of the observer performance studies. The ROC experiment was completed by displaying a fixed number of reconstructed homogeneous or heterogeneous images in random order. After a computer-controlled training session, in which the observer was presented with sample images, then information was asked of the observed as to whether the images contained heterogeneity and also what the possible region size was. The observers were asked to view 600 different reconstructed images in one ROC experiment session. Instead of just deciding whether the viewed image was homogeneous or heterogeneous, the observers were asked to rate the probability that the image was heterogeneous, using a continuous scale value between 0 and 1, where 0 represented "absolutely homogeneous," 1 represented "absolutely heterogeneous," and 0.5 was a guess of "it not being known." The observer
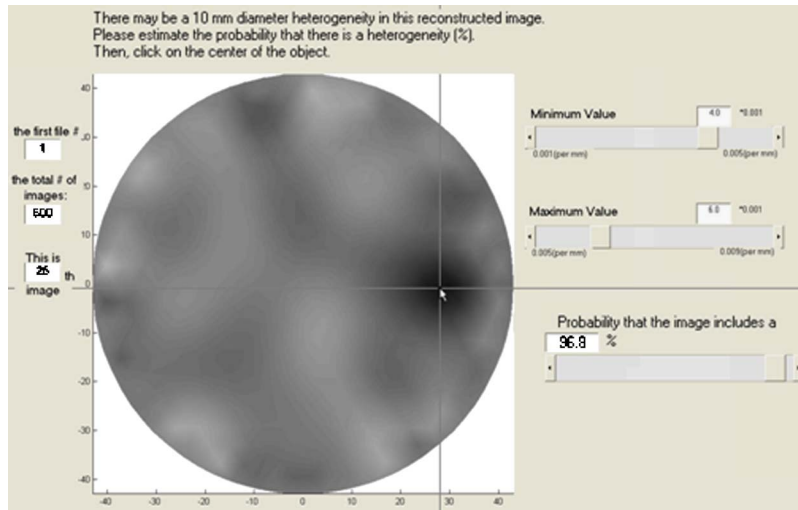
Fig. 2 The image detection program is shown. Observers can adjust the image gray map scale with the top two slide bars, and point to the estimated heterogeneity location with the computer mouse.

was also asked to pick up a possible location of the ROI by selecting the centroid of the region with a mouse cursor. The observer was able to control the image contrast and brightness by adjusting the minimum and maximum display parameters of the MATLAB image profile. For every viewed image, the program recorded all of the information about this image and the decision of the observer.

2.3 Computational Assessment of Contrast-to-Noise Ratio

In addition to human observer assessment, the contrast-to-noise ratio (CNR) was also used as a computational measure throughout this study. Similar to the signal-to-noise ratio in digital signal processing theory, the CNR is defined as the relative difference between the ROI and the background region values of the property, divided by the average variation in the background.[22–24] There are different choices for the background, and here the entire region outside of the target ROI was used as the background, and thus the CNR is defined as[25]

$$CNR_{III} = \frac{\mu_{ROI} - \mu_{background}}{\sqrt{w_{ROI}\sigma_{ROI}^2 + w_{background}\sigma_{background}^2}}, \qquad (2)$$

where μ_{ROI} is the mean of the node values in the target; $\mu_{background}$ is the mean value over the variable background; σ_{ROI} and $\sigma_{background}$ are the standard deviations of the target and the whole background areas, respectively; and w_{ROI} and $w_{background}$ are the weights in the target and background, which are defined as the fractional size of the target and background in the image field. Given the ROI size, CNR was calculated at each node and the maximum CNR and its correlated location was considered as the real ROI CNR and location. Using the images generated in Sec. 2.2 and given the real diameter and location of the target, CNR was calculated as the absorption contrast and the ROI size was systematically varied.

2.4 Parametric and Nonparametric Receiver Operating Characteristic Analysis

ROC curves are usually generated by a model-based fit to the reader data, as illustrated in Fig. 3. Several methods have been published to plot ROC curves based on discrete or continuous test data. These methods can be divided into two basic categories, nonparametric or parametric. The empirical nonparametric method is used to calculate the ROC curve using empirical histogram distributions, in which there is no need for structural assumptions nor parameters for model fitting. Though the empirical nonparametric method is robust and easy in some cases, it is not a smooth fitted curve leading to less conclusive analysis in the case of sparse datasets.

An improvement to the empirical nonparametric method is the nonparametric kernel smoothing technique.[13,26,27] In this method, a local density function, or the so-called kernel function, and a bandwidth are introduced to estimate the observer decision distribution function for diseased and healthy images. The kernel function and the bandwidth are optimized to numerically represent the distribution functions, thereby plotting a smooth and optimal ROC curve. More objective parametric methods require that some assumption be made regarding the functional form of the ROC curve. Several functional models have been proposed since ROC analysis was first developed.[28] One of the most popular models is the binormal ROC method, which assumes that a pair of latent normal decision variable distributions underlies the ROC data, and this has been widely used for ROC curve fitting.[12,29–32] According to this two-parameter model, each ROC curve is assumed to have the same functional form as that implied by two Gaussian decision variable distribution functions. Within this binormal model, the task of curve fitting becomes one of choosing numerical values for the two parameter pair to best represent the measured data in observer performance studies.

The software prepared for this study includes a MATLAB command line based program, in which the empirical nonparametric approach and the maximum-likelihood estimation ROC curve fitting methods[29] were applied (Fig. 3). In the
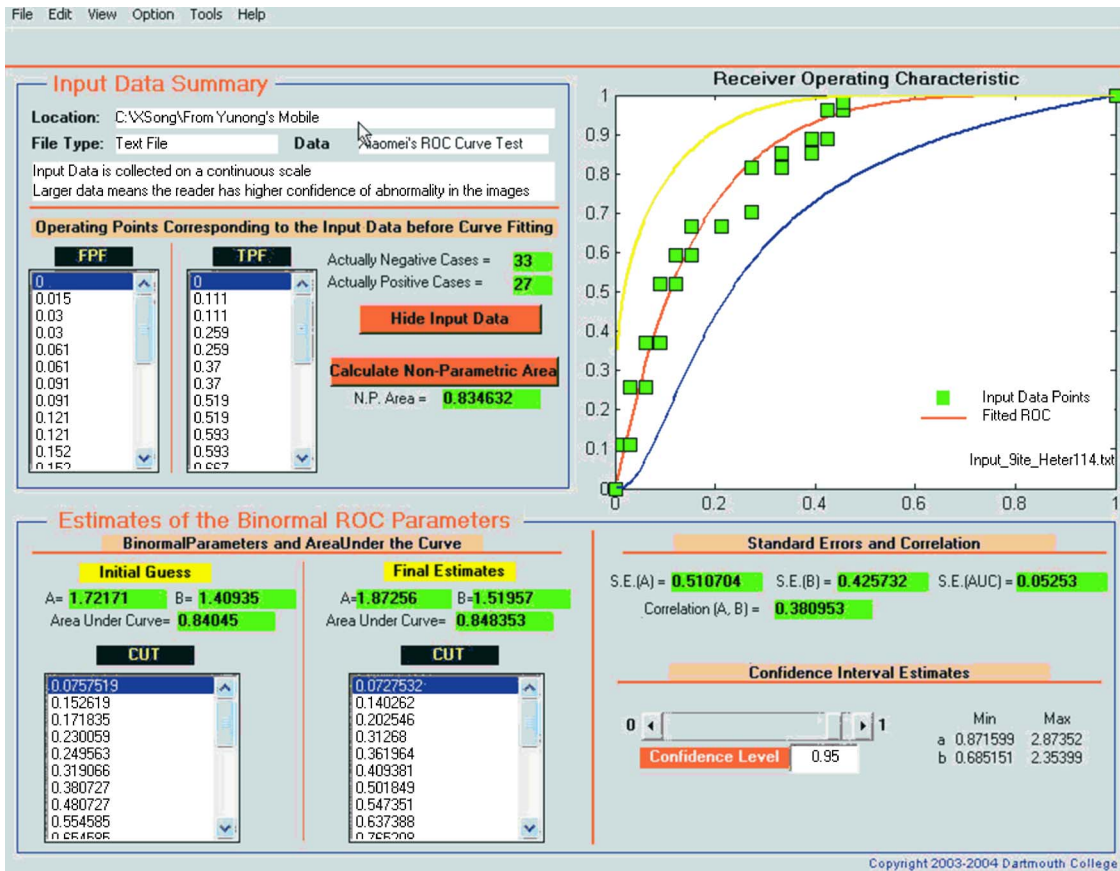
Fig. 3 A screen view of the graphical user interface (GUI)-based ROC fitting platform is shown, which includes empirical nonparametric initial analysis, maximum-likelihood estimation-based binormal ROC analysis, and related statistical parameter analysis such as the ROC curve confidence interval boundary, standard error, data correlation, and parameter confidence intervals.

ROC fitting process, each image reading was assumed to be independent of all other images, thus the whole observer practice followed an independent multinomial distribution formed after classifying the continuous rating according to an ordinal classification. It was also assumed that two latent normal distributions for homogeneous and heterogeneous tissue underlie the ordinal classification, leading to a multinomial probability model for the ROC curve. Based on these two assumptions, the parameters of the ROC curve were determined via maximum-likelihood estimation ROC analysis. In the LROC fitting process, observer performance information for reconstructed images consists of two parts: 1. a continuous confidence rating of possible heterogeneity presence, and 2. coordinates representing the possible object location. The one-stage plug-in method with a biweight kernel function[13] was used to generate the distribution bandwidth, and these were then applied to fitting the LROC curve, as illustrated in Fig. 4. This kernel function is as described in Ref. 12, with the optimal bandwidth chosen to smooth the data. The diagnostic probability density distribution functions of normal and diseased cases are estimated to be in the form of

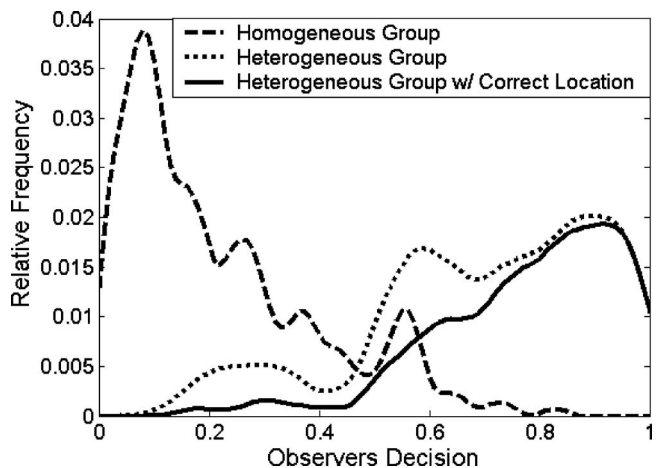$$f(t) = \frac{1}{n_0 h} \sum_{i=1}^{n_0} K\left(\frac{t - x_i}{h}\right), \qquad (3)$$

Fig. 4 The distributions graph of human observer responses are shown, and fitted with kernel density estimation. The window width was chosen from the direct plug-in method, and the Epanechnikov kernel was used.[9]

Table 1 Observer performance results for receiver operating characteristic (ROC) and location ROC (LROC) for human observers numbered 1 through 4. Note that observers 1 and 2 are experienced, and observers 3 and 4 are inexperienced.

Observer (note)	ROC area under the curve with error correction			LROC area under the curve with error correction		
	ROI=6 mm	ROI=10 mm	ROI=16 mm	ROI=6 mm	ROI=10 mm	ROI=16 mm
1	0.90±0.012	0.99±0.008	0.99±0.003	0.23±0.026	0.71±0.023	0.90±0.013
2	0.89±0.015	0.98±0.009	0.99±0.004	0.23±0.026	0.69±0.024	0.89±0.01
3	0.88±0.014	0.99±0.005	1.00±0.001	0.23±0.026	0.70±0.024	0.91±0.013
4	0.93±0.009	0.99±0.003	0.96±0.008	0.25±0.028	0.68±0.025	0.89±0.014

$$g(t) = \frac{1}{n_1 h} \sum_{i=1}^{n_1} K\left(\frac{t - y_i}{h}\right), \qquad (4)$$

where x_i and y_i are the assessed data points for normal and abnormal probabilities, with sample sizes n_0 and n_1. In this approach, the integration of K over the space provides a mean-zero value, and h is the bandwidth that controls the degree of smoothing. By using these models to represent the data, a smoothed version of the data is provided, in a nonparametric manner.

3 Results

In the subsections here, the results of the overall study are described. Four human observers and CNR values were calculated for all image sets. Human observer and CNR values are reported in terms of ROC and LROC curves and the related area-under-curve value with error correction. The results of the human observers with or without medical imaging background were compared to each other and then compared with the CNR-based computational observer. As discussed in Sec. 2.1, given a fixed system noise level, three key parameters determine the quality of a reconstructed image: 1. the size of the heterogeneities, 2. the absorption and scattering coefficient contrast between the heterogeneities and the homogeneous tissue, and 3. the number of the reconstruction iterations used in the image formation. In other words, the detectable object size and object-background contrast level need to be determined in this approach, and the reconstruction

process efficacy needs to be analyzed. In this section, the results of a number of studies designed to evaluate the impacts of these three parameters are reported for a standard version of the NIR tomography reconstruction algorithm. While other parameters such as signal to noise, artifact presence, etc., may also be important, these parameters have been found to be dominant factors qualitatively in changing the nature of the recovered images, and so are examined in detail here.

3.1 Human Observer Performance Comparisons

In trying to minimize the effect of the observer's background experience as a factor in the observer performance results, a training session was given to all participants prior to the final image presentation and detection process. In the training session, the trainee was asked to review sample images with the same reconstruction condition, and were notified if the images contained heterogeneity, and if so, what the location of the heterogeneity was. The trainee was allowed to review and analyze as many as 200 images before the real detection task began. Table 1 summarizes the area under the curve (AUC) values and associated standard error for both ROC and LROC curves for different observers in the three heterogeneity-size studies. The standard deviation between observers on the whole is less than 1% of the ROC and LROC area under the curve values, indicating that the observer performances are quite similar in this group of four subjects.

Table 2 Parameter settings used in the study of heterogeneity size (iteration number=6).

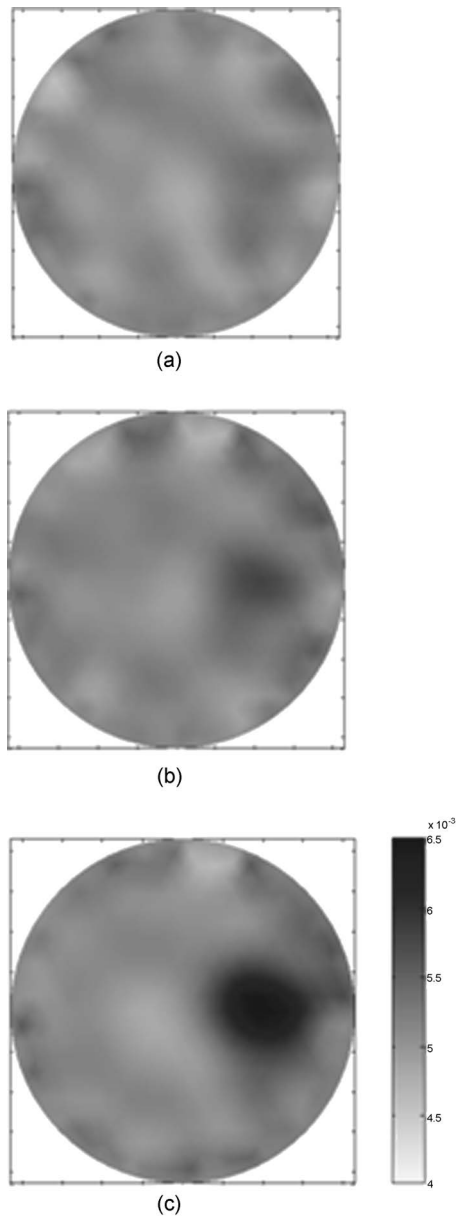| Study	Noise level (random normally distributed)	Heterogeneity properties		Background properties
		μ_a (mm^{-1})	Diameter (mm)	μ_a (mm^{-1})
1	1% amplitude and 1-deg phase shift	0.0055 to 0.01	4	0.0050
2	1% amplitude and 1-deg phase shift	0.0055 to 0.01	6	0.0050
3	1% amplitude and 1-deg phase shift	0.0055 to 0.01	10	0.0050
4	1% amplitude and 1-deg phase shift	0.0055 to 0.01	16	0.0050

(a)

(b)

(c)

Fig. 5 Sample images from heterogeneity-size studies are shown. The objects all have the same contrast of $C=2.0$, the same reconstruction process iteration number $I=6$, but with different object sizes. In (a), a small-sized object was used, with diameter of 6 mm. In (b), a medium-sized object was used, with diameter of 10 mm. In (c), a large object of diameter 16 mm was used.

3.2 Human Observer Versus Computational Contrast-to-Noise Ratio Performance Comparison

3.2.1 Heterogeneity-size study

This study was designed to evaluate how the NIR tomography imaging reconstruction algorithm response to different sized heterogeneities would affect the human perception from the images. For the sake of simplicity, a circular geometry was chosen for the heterogeneities with a diameter ranging from 4 to 16 mm, compared to an 86-mm-diam region for the background homogeneous tissue. Table 2 summarizes the relevant parameters of four observer performance studies, of

which each study contained 600 images with optical property ranges listed in the table. To maintain experimental consistency, zero-mean Gaussian noise of 1% amplitude and 1 deg in phase shift were added to the calculated boundary data of all studies to simulate a realistic dataset from our NIR tomography system. Sample images from heterogeneity-size studies, in which objects were sized from 4 to 16 mm and the contrast level defined in the next section is ranged from 1.1 to 2.0, are shown in Fig. 5.

Figure 6 summarizes the human observer and computed CNR performance in the heterogeneity-size studies. In Fig. 6(a), as the diameter of the target decreases down to 6 mm, the CNR value decreases below a level from which the heterogeneity can be efficiently detected, corresponding to when the AUC of the ROC is equal to 0.5. However, human observer detection is sufficiently high (above 0.5) for all heterogeneous images, even below 6 mm, having AUC values of 0.82 for 4-mm objects and 0.88 for 6-mm objects. As the size of the targets increases, the CNR values increase to allow for automated detection of objects in the image. Interestingly, as shown in Fig. 6(b), the localization accuracy is considerably worse in these same sets of images, with objects smaller than 10 mm not being able to be localized accurately.

3.2.2 Heterogeneity contrast level study

The optical property contrast between the object and background was provided by the absorption coefficient difference between the heterogeneity and the background. The heterogeneity-to-background contrast was defined as

$$C = \frac{\mu_{\text{ROI}}}{\mu_{\text{background}}}, \tag{5}$$

where μ_{ROI} is the mean absorption coefficient of the node values in the target and $\mu_{\text{background}}$ is the mean absorption coefficient value over the variable background. Using fixed object diameter $D=10$ mm and reconstruction process iteration number $I=6$, and varied contrast ranging from 1.1 to 2.0 in steps of 0.1, resulted in 600 images in total. The same human and computational observers are used in these ten studies. An example image from $C=1.4$, 1.7, and 2.0 is shown in Fig. 7.

Figure 8 summarizes the human observer and computed CNR performance in the heterogeneity contrast level studies for ROC detection task assessment [Fig. 8(a)] and LROC localization task assessment (Figure 8(b)). In Figure 8(a), where the ROI size was fixed at 10 mm, the human observer performed quite well, even at low contrast levels, and achieved perfect detection (AUC=1.0) performance when the contrast level was above 1.6. The computed CNR, on the other hand, was unacceptably low for the 10 to 50% contrast range, but was sufficiently accurate above this range. Human observers have more capability to detect heterogeneities with small contrast levels, yet this difference was negligible if the heterogeneity was at higher contrast levels compared to the homogeneous background.

3.2.3 Analysis of the reconstruction process iteration

The most dominant factors in the reconstruction process that affect image quality is the regularization parameter and the
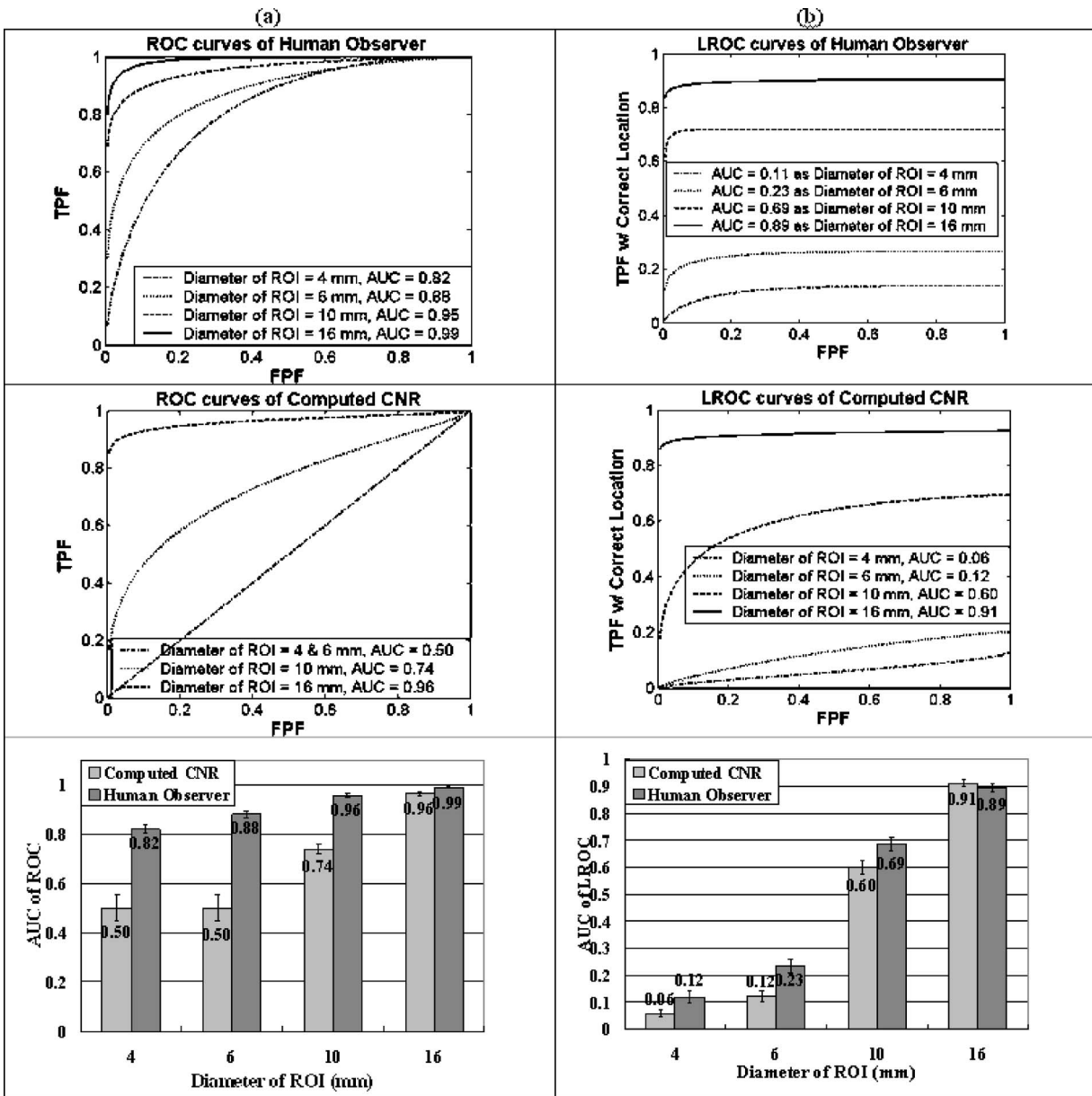
Fig. 6 The heterogeneity-size study for both human and computational observers is shown in (a) with ROC and area under curve data, and in (b) with LROC and AUC data. Both of these are shown for datasets where the absorption contrast ranged from 1.1 to 2.0, with a fixed six iterations.

number of iterations used in their formation. In the presented work a modified Levenberg Marquardt scheme is used, in which the regularization parameter starts out at a high value relative to the normalized Hessian diagonal (near $100\times$), and then is systematically decreased by a factor of 3 at each successive iteration. Using this approach, the iteration number is the major factor influencing the detection of objects, and so this was studied here. The outline of the iteration study is summarized in Table 3, where 600 images were used in total. During a single reconstruction process step, the 2-D image was first constructed and used as the input data for the forward problem.

In practice, as the iteration number increases, the objective function first decreases and then eventually reaches a lowest point, and then is increased very slowly, as shown in Fig. 9.

Here the ROI was 10 mm diam and the contrast ranged from 1.1 to 2.0. The error was estimated from a number of reconstructions, and the mean error is plotted as a function of iteration number. In Fig. 9, the minimum objective function value occurs at an iteration number of 9. Theoretically, the reconstructed image parameter value should be closest to the true image value at the point where the objective function is at its lowest value, but in practice, the reconstruction iteration process also inevitably introduces some high-frequency spatial noise, so that it affects the judgment of both the human observers and the computational observer. Sample images from iteration numbers 4, 6, and 10 are shown in Fig. 10.

Figure 11 summarizes the human observer and the computed CNR values for performance as a function of the iteration number for ROC detection tasks [Fig. 11(a)] and LROC
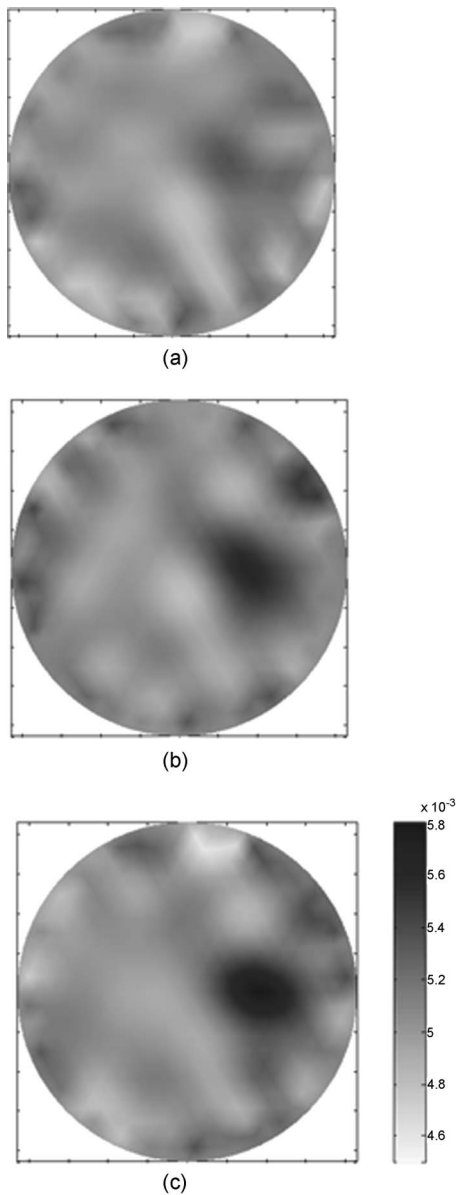
Fig. 7 Sample images from heterogeneity contrast studies are shown. The objects had the same size, with diameter of 10 mm, same reconstruction process with iteration number $I=6$, but with different contrasts. In (a), the contrast was $C=1.4$; in (b), the contrast was $C=1.7$; and in (c), the contrast was $C=2.0$.

localization tasks [Fig. 11(b)]. From Fig. 11, where the ROI size was fixed at 10 mm and the contrast level was varied from 1.1 to 2.0, it can be seen that the human observer always has better performance than this computed value of CNR, in terms of detectability and localization accuracy. In the range of 4 to 10 iterations, the iteration number has either negligible impact [see human observer data in Fig. 11(a)] or a slightly negative impact on the computed CNR data (Fig. 11) for both human observers and CNR detection and localization decisions.

4 Discussion

4.1 Assessment of the Human Observer Data

In the interhuman observer performance comparison study, AUC values for the ROC and LROC curves are summarized in Table 1. It is well known that the area under the curve of an ROC and LROC curve is a reasonable indicator of the quality of the observer performance. Given the fact that observers 1 and 2 had NIR medical experience and observers 3 and 4 were inexperienced, the main observation from Table 1 is that after the training procedure, readers without NIR imaging experience had negligible performance difference compared to readers with NIR imaging experience (i.e., less than 1% standard deviation in the AUC values overall). This trend is also observed when tested with different heterogeneity contrast levels and when the reconstruction process iteration number was varied (see Table 1). Summarizing these different tests, it was concluded that the four observers had less than 4% total variation from one another in any category of test, and the standard deviation in each test was less than 1%. These observations were important to then allow averaging of the ROC and LROC data results processed independently for all four human readers, in subsequent experimental datasets, thereby providing better power to assess the human observer results. It is always problematic to determine when averaging data from different readers is possible. However in this study all observers were given the same control of the images and all performed in a substantially similar manner, thus averaging the ROC data is a logical choice to assess the repeatability between readers. Analysis of the individual observer's datasets were carried out, and the computed ROC AUC values did not show substantial difference from one another (less than 5% in AUC).

4.2 Human Observer Decision Versus Computed Contrast-to-Noise Ratio

In the human observer detection process, readers were required to continuously rate the possibility of a heterogeneity being present in the images. In the computational observer detection process, the CNR value was calculated at each node and the maximum value was considered in the "decision" or "rating" of the computational observer for the NIR tomography images. Both the human observer and computational observer decision data are plotted as scatter pairs for the two image groups, including datasets where the ROI size was 10 and 16 mm. These are shown in Fig. 12. The solid line in Fig. 12 is a linear fit of human observer and CNR values, of which the fitted equation and R square value are also listed in the diagram. From Fig. 12, the main observation is that the CNR of the reconstructed images and the human observer responses are directly linearly correlated when the diameters of the targets are larger, and there is an enhanced ability of the human reader to detect smaller objects as lower CNR values. Interestingly, while the fit to the data is quite good in Fig. 12(b) with $R2=0.91$, the fit to the data in Fig. 12(a) is not as ideal, and the shape of the data might indicate that a higher order fit is in order, perhaps with a saturation below human observer probabilities of 0.5. Further studies of this observation are ongoing, and would be consistent with the need for more complex models of human observers that require spatial
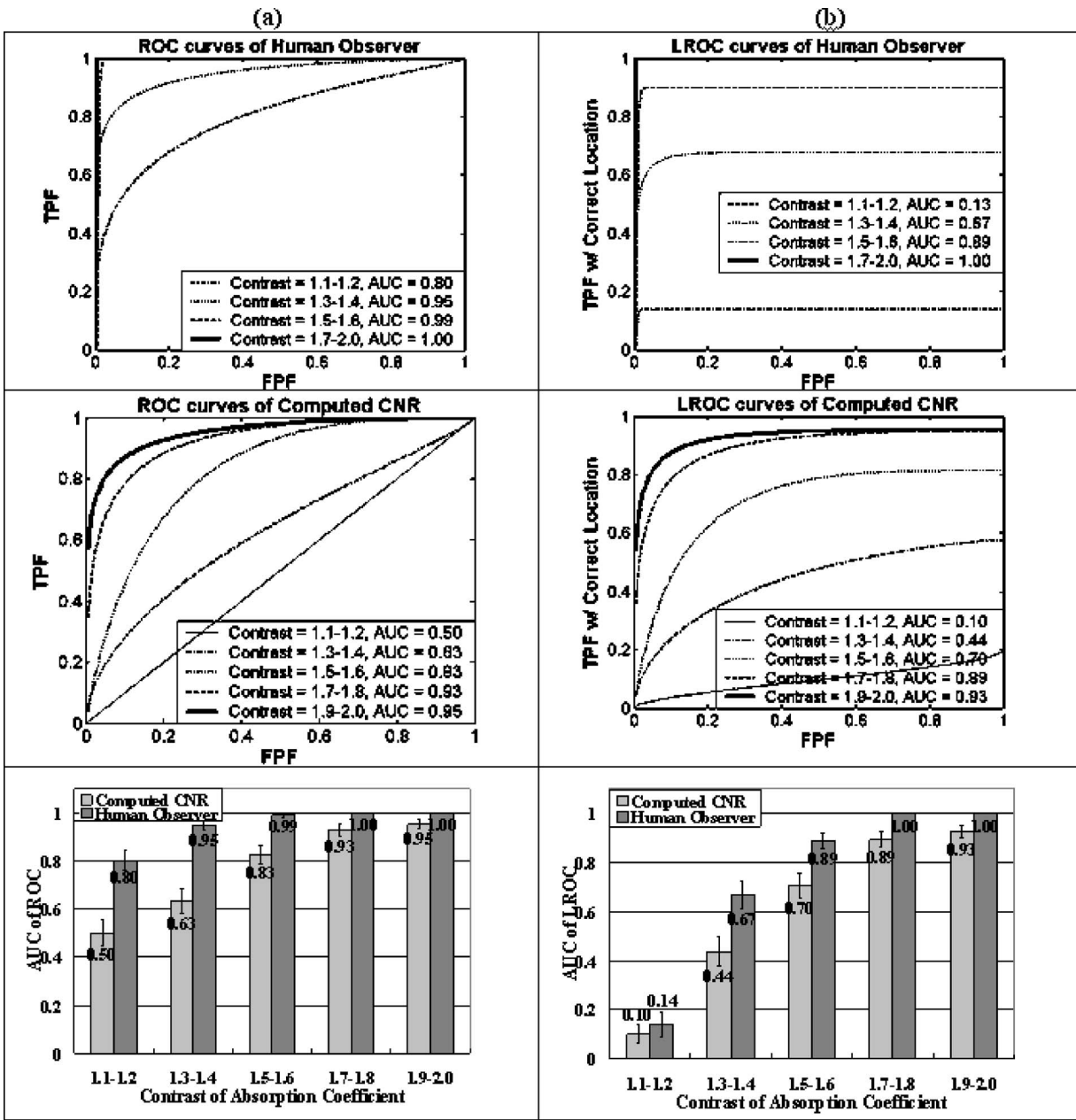
Fig. 8 The heterogeneity contrast study for both human and computational observers is shown, with (a) being the ROC curves and AUC, and (b) showing the LROC curves and area under curve. The diameter of the ROI was equal to 10 mm for all tests, and six iterations were used.

matched filter analysis to accurately mimick the detection rate of a human observer.

Examining Fig. 12, the data lines plotted cross the 50% point on the x axis at a level near 3.0 to 3.3 in CNR. This represents the CNR value at which the human observer has a 50% probability of detecting an object present in the image, and thus represents the threshold level at which observers can detect objects. Below this CNR value, the human observer cannot be expected to detect the object, as they have an equal probability of not detecting it. This observation is used as a lower limit rule of thumb in interpreting the data.
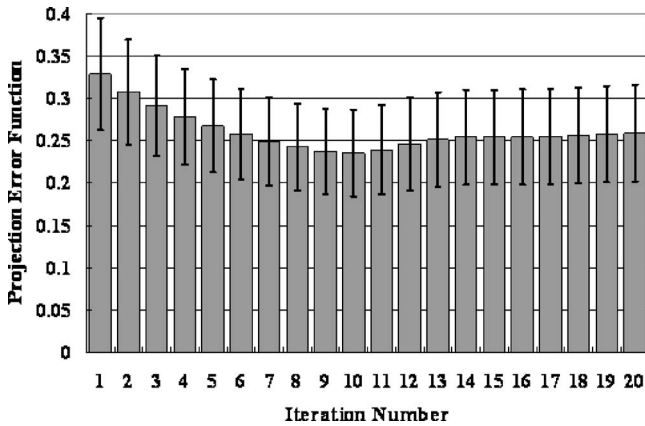
Table 3 Parameter settings used in the reconstruction process for the study of iteration numbers.

		Heterogeneity properties		
Study	Noise level	Contrast	Diameter (mm)	Iteration number
1	1% amplitude and 1-deg phase shift	1.1 to 2.0	10	3
2	1% amplitude and 1-deg phase shift	1.1 to 2.0	10	6
3	1% amplitude and 1-deg phase shift	1.1 to 2.0	10	8

Fig. 9 The mean of the projection error function of 294 heterogeneous reconstructed images is shown as the function of iteration number, using a diameter ROI of 10 mm and an absorption contrast range from 1.1 to 2.0. The error bars are the average standard deviations from all images.

4.3 Analysis of Human Observer Versus Computed Contrast-to-Noise Ratio Decisions

Given NIR images with a suspicious object in them, it is clear that the human observer would perform better than a measure of CNR; however, the human ability to localize where the region is may not be as developed as the ability to determine if an object is present. When the object size is as small as 6 mm, the existence of this object will change the diffuse light image, but its impact is not significant enough to be detected from within the noise pattern of the image. Iterations beyond the fourth iteration tend to degrade the ability to find the location as well. Though the object is not big enough to be seen on the reconstructed image at the object true location, its presence generally makes the image seem noisier than the average homogeneous image. This entire image profile change is detectable by a human observer whose strategy and overall ability are more sensitive to noise appearance. As the heterogeneity size increases to 10 mm, it dominates the image sufficiently enough to be accurately found, and as a result, the human observers have adequate detection capability and localization accuracy. Interestingly, this phenomenon is not apparent for the CNR value, where the AUC value does not dramatically change from ROC to LROC plots. This latter point indicates that the use of CNR for tumor detection in NIR tomography images would only provide a good indicator at larger sizes and higher contrast values, whereas localization of the region may be similar to that of human observers.

An interesting area that requires further investigation is the choice of size for "detecting" a region with LROC analysis. In this study, the choice of detection versus nondetection of the location was based on using the physical size of the known object as a measure. If the observer chose outside this region, then it was counted as a mistake in localization. Given that NIR tomography has a Guassian blurred imaging field, the size of the region that is suitable to use for LROC could arguably be larger than the physical size of the existing region. However, since the resolution of the image varies with radial location, it is not obvious what choice of distance would be optimal for LROC analysis. The results here repre-
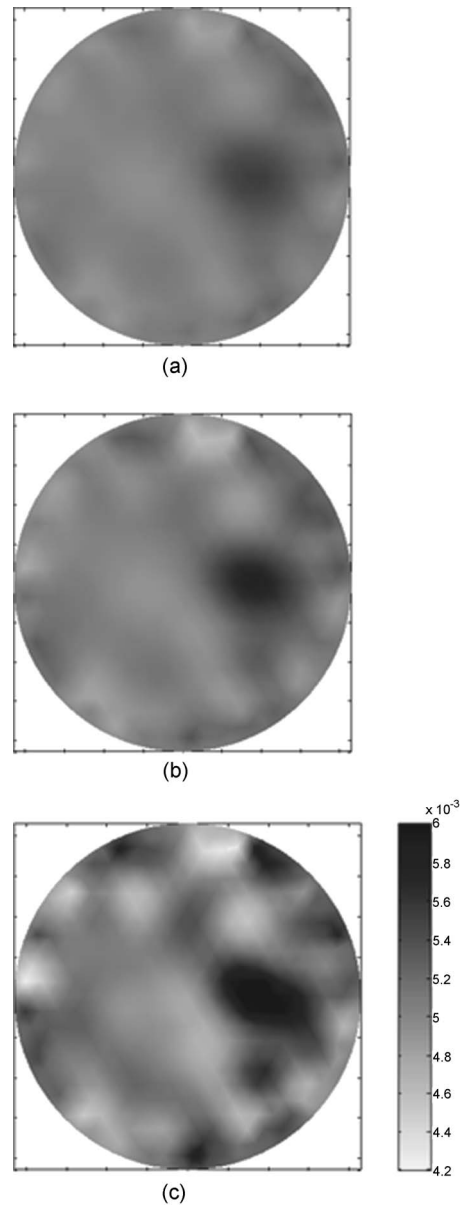

Fig. 10 Sample images from the reconstruction process are shown for different iteration numbers. The images have the same object size, $D=10$ mm, the same object-to-background contrast of $C=2.0$, but are from different numbers of iteration in the algorithm. In (a), the result from four iterations is shown. In (b), the result from six iterations is shown. And in (c), the result from ten iterations is shown.

sent one interpretation. However, if the size was increased, it is likely that the statistics of LROC values would improve monotonically with the increase in size. A further analysis could be carried out where the LROC AUC is plotted as a function of the choice of diameter for "detection."

The analysis here was completed on embedded inclusions where the contrast always increased relative to the background. In clinical studies, tumors always appear to have equal or larger vascularity, and thus have increases in contrast. So while the ROC analysis may be slightly different for situations where the contrast decreases relative to the background, this remains to be analyzed fully. Similarly, the noise
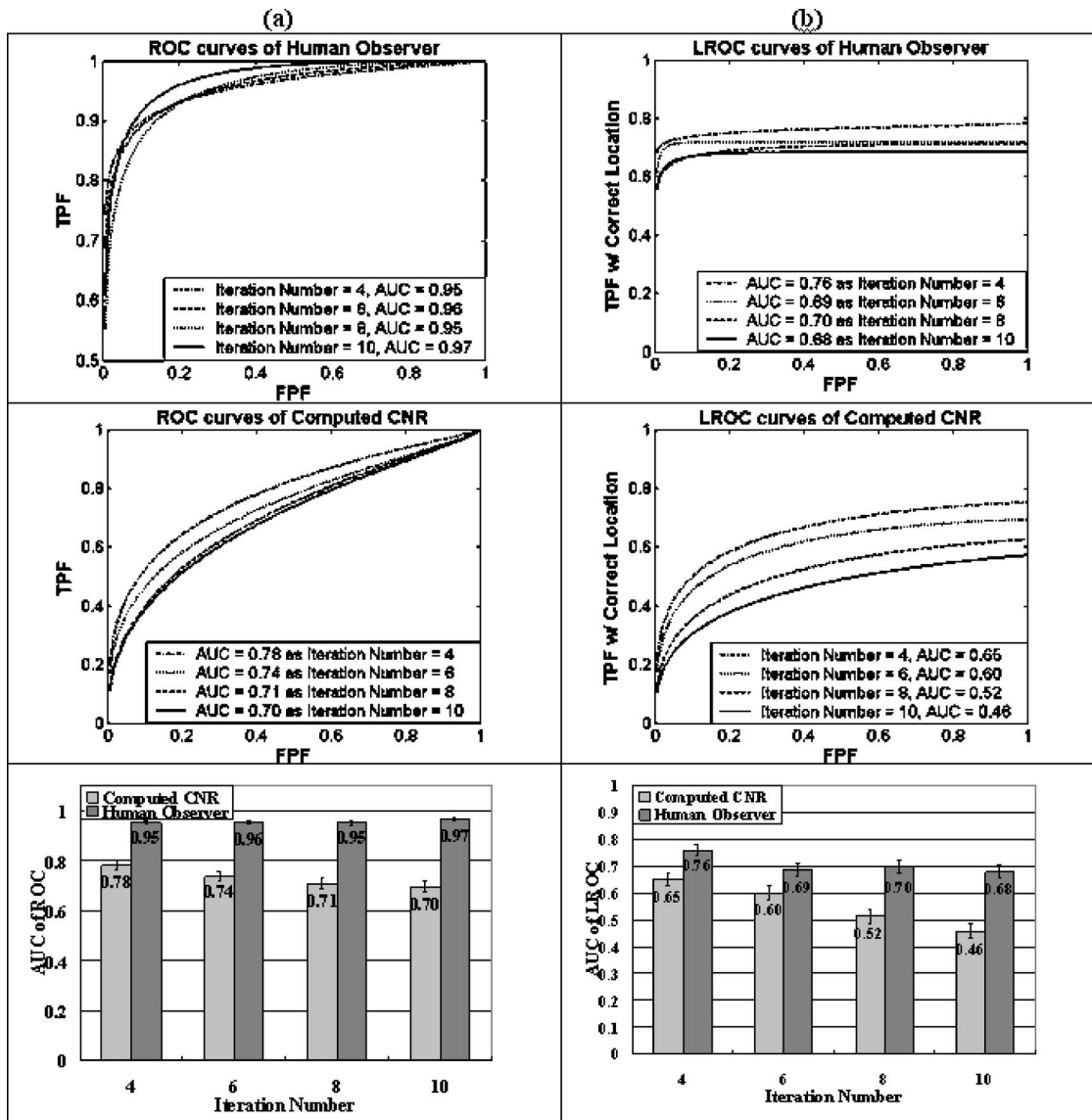
Fig. 11 The reconstruction process iteration number was studied for both human and computational observers. In (a), the ROC and AUC are shown, and in (b), the LROC and AUC are shown, for a fixed diameter equal to 10 mm and with varying absorption contrast from 1.1 to 2.0.

levels were fixed in this analysis, and were chosen to be representative of the clinical instrument that is being used at our medical center. However, it should be expected that lower signal-to-noise level values would lead to degradation of the images and hence poorer performance in the ROC and LROC curves.

5 Conclusion

In this study, NIR tomography image analysis is completed with observer performance assessment. Using results of experienced and inexperienced human observers and contrast-to-noise ratio computational calculations, NIR tomography images are assessed in terms of sensitivity to human observer discrepancy, detectable object size, object-to-background contrast levels and reconstruction algorithm parameters. While the computed CNR and human observer detection data have

similar capabilities to accurately localize the heterogeneity, human observers are more capable of detecting objects in NIR images with inclusion sizes of 6 mm diameter or less. Human observers are also more capable of detecting heterogeneities with contrast levels as small as 1.1, yet this difference is negligible if the heterogeneity is at higher contrast levels compared to the homogenous background. Human observer detection of objects appears to converge to a probability of 50% when the contrast-to-noise ratio of the region is near 3.0, indicating that images with lower than $CNR = 3.0$ could not be expected to have detectable regions in them. The results also indicate that effects of iteration and algorithm performance alter detectability of objects in NIR tomography images for both human observers and in the computed CNR value. Four interactions of the reconstruction process are sufficient to achieve maximum detection levels, and further iterations thereafter have negligible impact on the human observer de-
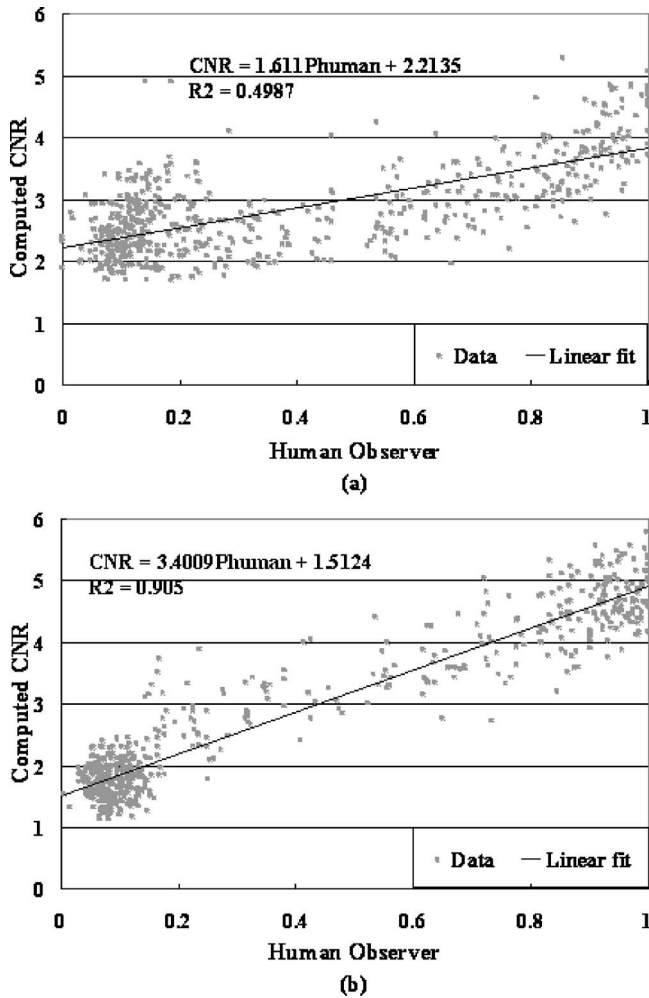
Fig. 12 Human and computational observer performance comparisons are shown for paired sets of images analyzed. The human observer decision (confidence level that the image is abnormal) and computational observer decision (contrast-to-noise ratio) for the same group of images are plotted. A linear fitting algorithm was then applied, and the R^2 values are present for the two situations of (a) where contrast was set to 2.0, iterations set at six, and the object size was set at 10 mm. In (b), the contrast was still 2.0, iterations were the same, but the object size was increased to 16 mm diameter.

tection rate and a distinct detrimental impact on the computed CNR and localization decisions.

References

1. B. J. Tromberg, O. Coquoz, J. B. Fishkin, T. Pham, E. R. Anderson, J. Butler, M. Cahn, J. D. Gross, V. Venugopalan, and D. Pham, "Non-invasive measurements of breast tissue optical properties using frequency-domain photon migration," *Philos. Trans. R. Soc. London, Ser. B* **352**, 661–668 (1997).
2. V. Ntziachristos and B. Chance, "Probing physiology and molecular function using optical imaging: applications to breast cancer," *Breast Cancer Res.* **3**(1), 41–46 (2001).
3. B. W. Pogue, S. P. Poplack, T. O. McBride, W. A. Wells, K. S. Osterman, U. L. Osterberg, and K. D. Paulsen, "Quantitative hemoglobin tomography with diffuse near-infrared spectroscopy: pilot results in the breast," *Radiology* **218**(1), 261–266 (2001).
4. B. Chance, S. Nioka, J. Zhang, E. F. Conant, E. Hwang, S. Briest, Schnall G. Orel, S. M. D., and B. J. Czerniecki, "Breast cancer detection based on incremental biochemical and physiological properties of breast cancers: a six-year, two-site study," *Acad. Radiol.* **12**(8), 925–933 (2005).
5. S. P. Poplack, K. D. Paulsen, A. Hartov, P. M. Meaney, B. W. Pogue, T. D. Tosteson, S. K. Soho, and W. A. Wells, "Electromagnetic breast imaging—pilot results in women with abnormal mammography," *Radiology* **243**(2), 350–359 (2007).
6. S. P. Poplack, K. D. Paulsen, A. Hartov, P. M. Meaney, B. W. Pogue, T. D. Tosteson, M. R. Grove, S. K. Soho, and W. A. Wells, "Electromagnetic breast imaging: average tissue property values in women with negative clinical findings," *Radiology* **231**(2), 571–580 (2004).
7. W. A. Wells, C. P. Daghlian, T. D. Tosteson, M. R. Grove, S. P. Poplack, S. Knowlton-Soho, and K. D. Paulsen, "Analysis of the microvasculature and tissue type ratios in normal vs. benign and malignant breast tissue," *Anal Quant Cytol. Histol.* **26**(3), 166–174 (2004).
8. S. E. Seltzer, R. G. Swensson, P. F. Judy, and R. D. Nawfel, "Size discrimination in computed tomographic images. Effects of feature contrast and display window," *Invest. Radiol.* **23**(6), 455–462 (1988).
9. R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.* **23**(10), 1709–1725 (1996).
10. H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," *Proc. Natl. Acad. Sci. U.S.A.* **90**(21), 9758–9765 (1993).
11. H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions," *J. Opt. Soc. Am. A Opt. Image Sci. Vis* **15**(6), 1520–1535 (1998).
12. C. E. Metz, "ROC Methodology in radiologic imaging," *Invest. Radiol.* **21**, 720–733 (1986).
13. X. H. Zhou and J. Harezlak, "Comparison of bandwidth selection method for kernel smoothing of ROC curves," *Stat. Med.* **21**, 2045–2055 (2002).
14. D. P. Chakraborty, "Statistical power in observer performance studies: a comparison of the ROC and free-response methods in tasks involving localization," *Acad. Radiol.* **9**, 147–156 (2002).
15. D. P. Chakraborty and L. Winter, "Free-response methodology: alternate analysis and a new observer-performance experiment," *Radiology* **174**, 873–881 (1990).
16. S. J. Starr, C. E. Metz, L. B. Lusted, and D. J. Goodenough, "Visual detection and localization of radiographic images," *Radiology* **116**(3), 533–538 (1975).
17. K. D. Paulsen and H. Jiang, "Spatially varying optical property reconstruction using a finite element diffusion equation approximation," *Med. Phys.* **22**(6), 691–701 (1995).
18. S. R. Arridge, J. C. Hebden, M. Schwinger, F. E. W. Schmidt, M. E. Fry, E. M. C. Hillman, H. Dehghani, and D. T. Delpy, "A method for three-dimensional time-resolved optical tomography," *Int. J. Imaging Syst. Technol.* **11**(1), 2–11 (2000).
19. B. W. Pogue, S. Geimer, T. O. McBride, S. Jiang, U. L. Österberg, and K. D. Paulsen, "Three-dimensional simulation of near-infrared diffusion in tissue: boundary condition and geometry analysis for finite element image reconstruction," *Appl. Opt.* **40**(4), 588–600 (2001).
20. H. Dehghani, B. W. Pogue, J. Shudong, B. Brooksby, and K. D. Paulsen, "Three-dimensional optical tomography: resolution in small-object imaging," *Appl. Opt.* **42**(16), 3117–3128 (2003).
21. A. P. Gibson, J. Riley, M. Schweiger, J. C. Hebden, S. R. Arridge, and D. T. Delpy, "A method for generating patient-specific finite element meshes for head modelling," *Phys. Med. Biol.* **48**(4), 481–495 (2003).
22. R. F. Wagner, S. W. Smith, J. M. Sandrick, and H. Lopez, "Statistics of speckle in ultrasound B-scans," *IEEE Trans. Sonics Ultrason.* **30**, 156–161 (1983).
23. J. J. Rownd, E. L. Madsen, J. A. Zagzebski, G. R. Frank, and F. Dong, "Phantoms and automated system for testing the resolution of ultrasound scanners," *Ultrasound Med. Biol.* **23**(2), 245–260 (1997).

24. H. Lopez, M. H. Loew, and D. G. Goodenough, "Objective analysis of ultrasound images by use of a computational observer," *IEEE Trans. Med. Imaging* **11**(4), 496–506 (1992).
25. X. Song, B. W. Pogue, S. Jiang, M. M. Doyley, H. Dehghani, T. D. Tosteson, and K. D. Paulsen, "Automated region detection based on the contrast-to-noise ratio in near-infrared tomography," *Appl. Opt.* **43**(5), 1053–1062 (2004).
26. K. H. Zhou, W. J. Hall, and D. E. Shapiro, "Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests," *Stat. Med.* **16**, 2143–2156 (1997).
27. C. J. Lloyd and Z. Yong, "Kernel estimators of the ROC curve are better than empirical," *Stat. Probab. Lett.* **44**, 221–228 (1999).
28. J. A. Swets, "Indices of discrimination or diagnostic accuracy: their ROCs and implied models," *Psycholog. Bull.* **99**(1), 100–117 (1986).
29. C. E. Metz, B. A. Herman, and J. H. S. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Stat. Med.* **17**, 1033–1053 (1998).
30. J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Invest. Radiol.* **14**(2), 109–121 (1979).
31. J. A. Swets, "Sensitivities and specificities of diagnostic tests," *JAMA, J. Am. Med. Assoc.* **248**(5), 548–550 (1982).
32. J. A. Swets, D. J. Getty, R. M. Pickett, C. J. D'Orsi, S. E. Seltzer, and B. J. McNeil, "Enhancing and evaluating diagnostic accuracy," *Med. Decis Making* **11**(1), 9–18 (1991).