Radiation Protection Dosimetry (2010), Vol. 139, No. 1–3, pp. 20–25 doi:10.1093/rpd/ncp305
Advance Access publication 18 January 2010

A STATUS REPORT ON FREE-RESPONSE ANALYSIS

D. P. Chakraborty*
Department of Radiology, University of Pittsburgh, Pittsburgh, PA, USA

*Corresponding author: dpc10@pitt.edu

The purpose of this paper is to summarise recent progress in free-response receiver operating characteristic (FROC) methodology. These are: (1) jackknife alternative FROC analysis including recent extensions and alternative methods; (2) the search-model simulator that enables validation and objective comparison of different methods of analysing the data; (3) case-based analysis that has the potential of greater clinical relevance than conventional free-response analysis; (4) a method for collectively analysing the multiple lesion types in an image (e.g. microcalcifications, masses and architectural distortions); (5) a method for sample-size estimation for FROC studies; and (6) a method for determining an objective proximity criterion, namely how close must a mark be to a true lesion in order to credit the observer for a true localisation. FROC analysis is being increasingly used to evaluate the imaging systems and understanding of recent progress should help researchers conduct better FROC studies.

INTRODUCTION

Receiver operating characteristic (ROC) methodology is a widely accepted method for measuring imaging system performance[1–3]. Unlike physical measurements it takes into account the entire imaging chain, including the radiologist. ROC is a binary paradigm: the patient does or does not have disease (binary truth), and the radiologist's task is to state whether the patient does or does not have disease (binary response). The resulting 2×2 truth table defines good decisions, i.e. true positive, true negative and bad decisions, i.e. false positive and false negative. Because it applies to a binary task, ROC cannot take into account multiple suspicious regions and/or multiple lesions that may be present in the image. The free-response paradigm is an extension of ROC to include lesion detection and localisation tasks. In this paradigm the radiologist does not know *a priori* how many lesions may be present in an image, if any, and therefore must search the image for lesions and mark regions that are suspicious. Because it takes more information into account the free-response paradigm is more sensitive at detecting small differences between performances of modalities, i.e. it has higher statistical power[4,5]. Statistical power is an important consideration in observer performance studies as it determines the probability of detecting a true difference between modalities while controlling the probability of detecting non-existent differences. It is directly related to the precision of the measurement. Due to its higher statistical power, the free-response method is being increasingly used to compare imaging systems[6–9]. The purpose of this paper is to summarise the current status of free-response analysis. The authors start by giving a brief account of the status prior to 2004 when methods were available for analysing individual reader free-response data in one modality, but correlated multiple reader data in two or more modalities could not be analysed. The latter is the most common application of observer performance studies. Next, the two current approaches were described to analyse free-response data for multiple readers and multiple modalities, which are implemented in freely available software (www.devchakraborty.com). Other methods will be discussed followed by a description of a free-response data simulator used to validate these methods and results of the validation testing. A new approach termed case-based analysis will be described that can potentially extend the clinical relevance of free-response studies and make it possible to simultaneously analyse the multiple lesion types. Finally, sample-size estimation (How many readers and cases are needed to detect a specified difference?) and an approach to more objectively choosing the proximity criterion will be described.

STATUS PRIOR TO 2004

The free-response paradigm was introduced more than four decades ago[10] but subsequent progress has been slow. The free-response receiver operating characteristic (FROC) curve was introduced[11] in 1978. It is a visual summary of observer performance in the free-response task but a method for fitting the data points to a theoretical curve was not available. The free-response data unit is a mark-rating pair, where a mark is the location of a suspicious region and the rating is the confidence level that a real lesion is present at the indicated location. The number of mark-rating pairs per image is a random integer (0, 1, 2, ...) and the number of lesions per image (0, 1, 2, ...) and their locations

are assumed to be known to the investigator, but the observers are, of course, blinded to this information. The marks are scored as lesion localisations (LLs) or non-lesion localisations (NLs) according to their proximity to real lesions (close marks are LLs). This requires specification of a proximity criterion, which is currently arbitrarily selected. The FROC curve[11] is defined as the plot, as the threshold confidence level is varied, of LL fraction (LLF = number of LLs/total number of lesions) vs. non-LL fraction (NLF = number of NLs/total number of images). The alternative FROC (AFROC) curve is defined as the plot of LLF vs. false-positive fraction (FPF = number of normal images with highest rated mark greater than the threshold confidence level/total number of normal images); FPF is the x-axis of the traditional ROC curve if it is assumed that the highest rating reflects the single rating that would have been given in an ROC task. FROC and AFROC curves are different ways of visualising free-response data, and based on current understanding the latter is preferable because unlike the FROC curve, which can theoretically extend infinitely to the right, the AFROC curve, like the ROC curve, is contained within the unit square. Beginning in 1989, parametric methods for fitting FROC and AFROC curves, termed FROCFIT and AFROC fitting, respectively[12,13] were developed, which give good fits to human observer data. In 1996 Swensson[14] described a parametric model for the localisation ROC (LROC) paradigm (in which the observer gives a single ROC-like rating and indicates the location of the most suspicious region) that also predicted FROC and AFROC curves. In 2002 the initial detection and candidate analysis (IDCA) method was proposed[15] for fitting a computer-aided detection (CAD) algorithm generated FROC curve for a single modality. All of these methods allowed fitting of individual reader FROC or AFROC curves, but they are not able to analyse the correlated data in which a number of readers interpret the same cases in all modalities, commonly termed the multiple-reader multiple-case multiple-modality (MRMC) study design. This design is needed to be able to generalise the conclusions of a study to the populations of readers and cases.

STATISTICAL ANALYSIS OF OBSERVER PERFORMANCE DATA

Free-response analysis, or for that matter analysis of any other observer performance paradigm, consists of defining a figure of merit (FOM) and a method for testing the significance of differences between observed FOMs for two or more modalities, i.e. analysis = FOM + significance testing. The FOM quantifies observer performance by rewarding 'good' decisions (finding lesions, LLs, and/or not marking NLs), and penalising for 'bad' decisions (missing lesions and/or marking NLs). The parametric methods described above allow one to calculate FOM for a given data set or one can use non-parametric (NP) methods. There are two possibilities, the FOM difference is zero, i.e. the null hypothesis (NH) is true, or the difference is non-zero, i.e. the alternative hypothesis (AH) is true. One selects a test size α (e.g. $\alpha = 5$ %), which is a control on the probability of declaring modalities different when in fact they have the same FOM. The analysis yields a test statistic (e.g. the t-statistic for a single reader and two modalities or an F-statistic for multiple readers and multiple modalities) and if the test statistic falls in the acceptance region of the NH, which is determined by α and the statistical distribution of the test statistic in question, the NH is not rejected and otherwise it is rejected. Equivalently the test-statistic can be converted to a p-value and if $p < \alpha$ the NH is rejected and the modalities are declared significantly different. Determining if the NH should be rejected, or equivalently the p-value, is termed significance testing. Statistical power is defined as the probability of rejecting the NH when it is false. In planning an observer performance study, it is desirable to have high statistical power and one typically aims for 80 % power. If statistical power is low, then true differences between modalities may not be detected (i.e. one may falsely accept the NH). Dorfman-Berbaum-Metz (DBM) solved the significance testing problem for MRMC-ROC studies in 1992[16]. The significance testing procedure uses a re-sampling technique known as the jackknife[17]. The cases are sequentially jackknifed, i.e. removed from the analysis, and the FOM is re-computed for each jackknife. If θ_{ij} is the FOM for modality-i and reader-j when all cases are included, and $\theta_{ij(k)}$ is the corresponding FOM [e.g. the area under the ROC curve (AUC)] when case-k is deleted, and N_T is the total number of cases, the jackknife pseudovalue P_{ijk} for modality-i, reader-j and case-k is defined by:

$$P_{ijk} = N_T\theta_{ij} - (N_T - 1)\theta_{ij(k)}.$$

The procedure is repeated for all cases, readers and modalities yielding a three-dimensional matrix of pseudovalues with ijk elements, where i is the number of modalities, j the number of readers and k the number of cases. The pseudovalue-matrix is analysed using a mixed model analysis of variance (ANOVA), which calculates a p-value for rejecting the NH that all modalities have identical FOMs. The procedure also calculates confidence intervals for FOM differences for all pairings of modalities.

JAFROC AND JAFROC1

The DBM method is applicable to the free-response parametric fitting methods described previously, with the ROC FOM replaced by an appropriate free-response FOM calculated from the fitted curves, but this possibility went unnoticed until 2004 when the jackknife alternative FROC (JAFROC) method was developed and validated[4]. Analogous to DBM, which applies to MRMC-ROC studies, JAFROC applies to MRMC-FROC studies. The JAFROC FOM θ^{JAFROC} is the NP area (i.e. the trapezoidal area obtained by joining the operating points, including the upper right corner, with straight lines) under the AFROC curve. Recall that the AFROC curve is defined as the plot of LLF vs. FPF, where FPF is computed in the usual (ROC) sense over *normal* cases, i.e. it is the number of normal cases with highest rated NLs rated higher than a threshold confidence level, divided by the total number of normal cases. By changing the threshold confidence level from $-\infty$ to $+\infty$ one traces out the AFROC curve from top to bottom. An additional analysis method was described in ref.[4] termed JAFROC1, which used a slightly different FOM defined as the NP area under the AFROC1 curve, $\theta^{\mathrm{JAFROC1}}$. The AFROC1 curve is defined as the plot of LLF vs. FPF1, where FPF1 is the FPF computed over normal vs. abnormal cases, i.e. it is the number of normal and abnormal cases with highest rated NLs rated higher than a threshold confidence level divided by the total number of cases (FPF1 can only be defined in the free-response context since in ROC abnormal images cannot yield false positives). In either case the significance testing procedure was identical to DBM-MRMC except for the FOM (θ^{JAFROC} or $\theta^{\mathrm{JAFROC1}}$) being used instead of AUC. The JAFROC x-axis is identical to that of the ROC curve, namely FPF. However, the JAFROC y-axis is LLF, while the y-axis of the ROC curve is TPF. LLF takes LL into account but TPF does not. For example, if on an abnormal image an observer bases the ROC rating on a highly suspicious normal region but does not see the lesion, the ROC rating would be counted as a high confidence level TP, which raises AUC. However in FROC scoring these count as two mistakes: a missed lesion and an NL ('false-positive'). The NL is not used in JAFROC but the fact that the lesion was missed lowers LLF, which lowers the area under the AFROC curve. Similarly with multiple lesions ROC obtains one rating for the image but JAFROC uses ratings from all lesions. On normal images JAFROC offers no advantage over ROC, since both obtain one rating.

VALIDATION AND POWER

Validation and statistical power determination of these methods require a simulator that generates synthetic free-response data under controlled conditions. The simulator described in ref. (4) allowed the deliberate introduction of intra-image and inter-modality correlations. Specifics of the simulator are described in the referenced paper, but an important point was that it assumed equal numbers of decision sites (suspicious regions at which decisions to mark or not mark are made) on all images, and that all lesions were 'found', i.e. were considered for marking. The first assumption does not take into account expected random case-sampling effects (some images are expected to have more suspicious regions than others). The second assumption leads to the prediction that at sufficiently low confidence level the FROC curve attains $\mathrm{LLF} = 1$ for finite NLF, which prediction is not supported by data. Both JAFROC and JAFROC1 passed the NH validity test using this simulator. Both JAFROC and JAFROC1 have recently been re-validated[5] using a simulator described below that does not make these assumptions.

IDCA AND NP ANALYSIS

In parametric IDCA and NP analysis the FOM is the area AUFC_γ under the FROC curve to the left of a specified value $\mathrm{NLF} = \gamma$[15,18]. Recall the FROC curve is the plot of LLF vs. NLF. In IDCA the FOM AUFC_γ is estimated by fitting the observed operating points to a smooth curve and calculating the area by numerical integration. In the NP method, AUFC_γ is estimated by connecting adjacent operating points with straight lines and calculating the trapezoidal area under the line segments. For CAD data, the operating points are closely spaced and the trapezoidal estimate introduces minimal error. However, for human observers the operating points are sparse and the trapezoidal area underestimates the true area; moreover the jackknife pseudovalues can be confined to a small number of values. For the parametric IDCA method, the jackknife pseudovalue-based DBM significance testing method led to correct NH behaviour[5], but the NP method led to incorrect NH behaviour that was corrected by using the bootstrap for significance testing. In the bootstrap significance testing method for each simulated data set 2000 bootstrap samples were generated (i.e. re-sampled with replacement) and the distribution of the difference of the FOMs (not pseudovalues) is percentiled to determine two cut-points c_1 and c_2 such that 2.5 % of the differences are below c_1 and 2.5 % of the differences are above c_2. If the 95 % confidence interval for the FOM difference, namely (c_1, c_2), does not include zero the NH is rejected. The cut-points can be converted to a p-value. The procedure was repeated for 2000 simulated data sets to obtain the average NH rejection rate.

For simulated CAD data the two methods (IDCA and NP) had similar statistical powers, but since it makes fewer assumptions, the NP method may be preferable even though the bootstrap is computationally more intensive than the jackknife. It was found that for maximum statistical power it was desirable to choose γ as large as possible. However, if γ is chosen larger than the observed end-point (uppermost operating point) then $AUFC_\gamma$ cannot be calculated. For a fair comparison the same γ must be chosen for both modalities, as otherwise one will introduce bias in favour of the modality with the larger γ. Choosing a large common value for γ is relatively easy with CAD algorithms since they generate many NLs but is more difficult with human observers since conservative observers may not provide appreciable numbers of NLs. In that case the maximum common value of γ would be limited by the NLF of the observer, who generates the least number of NLs, and the data to the right of this value provided by other observers has to be excluded from the analysis, which would compromise power. Similar problems also affect ROC/JAFROC analysis since the FOMs (AUC/θ^{JAFROC}) is the total area under the ROC/AFROC curve, and if the observed operating points are clustered near the lower left corner, the FOM estimate involves an undesirable larger extrapolation to the upper right corner. This shows the need for careful data collection in ROC/FROC studies so that the full range of the operating characteristic is adequately sampled; alternatively one could use partial area measures (θ_γ^{JAFROC} or $\theta_\gamma^{JAFROC1}$) which are similar to $AUFC_\gamma$, but will sacrifice power because γ is determined by the most conservative observer.

FREE-RESPONSE DATA SIMULATOR

The free-response simulator is based on a search-model for the mark-rating pairs[19,20], which in turn is based on the Kundel-Nodine model for radiologist interpretations[21]. The search-model assumes that each image yields a finite number of decision sites (suspicious regions at which decisions to mark or not mark are made), which are termed noise sites or signal sites if they correspond to normal anatomy or lesions, respectively. The number of noise sites on an image is assumed to be a random sample from a Poisson distribution with mean λ; the number of signal sites on an abnormal image is assumed to be sampled from a binomial distribution with mean $s\,\nu$ and trial size s, where ν is the probability that a lesion is a decision site (i.e. it is 'found' or considered for marking) and s is the number of lesions in the image. The decision variable (z-sample or confidence level) from a noise site is sampled from $N(0,1)$ and that from a signal site is sampled from $N(\mu,1)$ where $N(\mu,\sigma^2)$ is the normal distribution with mean μ and variance σ^2. The simulator can be used to predict the FROC curves for a single reader, and it is the free-response analogue of the binormal model[22] used extensively in the ROC analysis. However, to test methods for analysing the MRMC-FROC data, it is necessary to extend the search model simulator to multiple modalities and readers. The corresponding problem was solved in the ROC context[23] and the resulting ROC simulator is termed the Roe and Metz simulator. This simulator was recently extended to MRMC free-response studies[24]. The essential modification was the introduction of a location factor (L), which accounts for the locations of the marks.

CASE-BASED ANALYSIS

A universal limitation[24] of current free-response figures of merit is that they are *lesion-based* giving equal importance to all *lesions*. Consequently, a case with a large number of lesions contributes more to the FOM than a case with only one lesion. To ensure that the case is the unit of analysis one should be giving equal importance to each *case* (i.e. each patient). Another limitation is that lesion-based methods do not account for different cancer risks associated with the lesions. A relatively benign lesion has lower cancer risk than a highly malignant lesion. Using the JAFROC-FOM as an example, in case-based analysis one performs a risk-weighted-average over all lesions on an abnormal case, of the probability that a lesion is rated higher than the highest rated non-lesion on a normal case, and averages over all cases. The risks of lesions on an image must add up to unity (so that each patient gets equal importance in determining the FOM) and can be assigned by a truth panel (not the readers in the study) based upon their clinical knowledge of survival statistics associated with different lesion types. One could adopt a five- point scale from 1: lesion is relatively low risk and the patient needs follow-up in 6 months to 5: lesion is high risk and action needs to be taken soon. The risk rating can be converted to normalised risks that add up to unity. For an image with s lesions risk-rated r_i ($i = 1, 2, \ldots, s$) the weights are $w_i = r_i/\Sigma r_i$. Any lesion-based free-response FOM (θ^{JAFROC}, $\theta^{JAFROC1}$ or $AUFC_\gamma$) has its case-based counterpart. Case-based analysis could potentially extend the capability of relatively inexpensive laboratory free-response measurements to better correlate with the higher levels in the[25] six-level hierarchy of clinical efficacies of measurements. Showing the higher correlation requires a clinical trial which the authors hope will be conducted soon. The case-based method can also be used to analyse different lesion-types. Different types of lesions may be present in the same breast, e.g. microcalcifications, masses, architectural distortions. Currently analysis

is generally conducted separately for the different types, e.g. one analysis for masses, one for microcalcifications, etc. This underestimates the probability of a Type I error and additionally one lacks a measure of overall performance. Case-based analysis offers a solution. Given the normalised risk of the different types of lesions the case-based FOM takes all lesions into account and the analysis yields a single FOM and p-value for overall performance.

SAMPLE SIZE ESTIMATION

Sample size estimation seeks to predict the numbers of cases and readers necessary if one is to have a reasonable chance (i.e. statistical power) of detecting (i.e. rejecting the NH) a specified difference (i.e. effect size) in performance between two modalities. With the high cost of conducting observer performance studies, sample-size estimation plays an important role at the planning stage of an observer study. Hillis and Berbaum[26] (HB) have described a sample size estimation procedure for the ROC paradigm. The procedure uses the DBM-MRMC-ANOVA calculated pseudovalue variance components, which are input to SAS software that can be downloaded from the University of Iowa website: http://perception.radiology.uiowa.edu/Software/. Since the underlying FOM (not decision variable) variance-components models are the same, the HB sample-size estimation procedures should work with the FROC data (although the z-sampling models are different, the FOM models are the same, i.e. in spite of the complication introduced by the multiple marks and ratings, it remains true that each data set yields a single FOM). Sample size estimation code for both ROC and FROC studies can be downloaded from the author's web site.

ADDRESSING THE ARBITRARINESS OF THE PROXIMITY CRITERION

One way of defining a proximity criterion is via an acceptance radius (AR), defined such that marks within this distance from lesion centres are classified as lesion localisations. Currently there is no objective guideline on how to choose AR and this is a fundamental limitation of all proposed methods of analysing the FROC data. Increasing AR causes the end-point of the FROC to move to the upper left, implying improved performance and vice-versa. One approach to this problem is as follows[27]. Suppose the observer marks an abnormal image and the suspicious finding that caused the mark was in fact a lesion, i.e. the observer saw the lesion. The mark will likely fall close to the lesion but for obvious reasons will not coincide with the exact centre of the lesion: for large irregular lesions there is subjectivity regarding the meaning of lesion 'centre', in addition there is hand 'jitter', etc. Since the observer saw the lesion he/she should get credit since the fact that the mark is not at the exact centre will likely not affect the clinical outcome (a radiologist and a surgeon looking at the same image do not have to agree on the exact centre of the lesion in order to agree that they are dealing with the same lesion and the clinical follow-up, e.g. biopsy, will be unaffected). If the suspicious finding is in fact a normal region the observer should not get credit since such marks will tend to lead to inappropriate clinical follow-up (the biopsy could occur at the wrong location and the real lesion may not be biopsied). The two types of marks are termed perceptual hits and perceptual misses, respectively; on normal images all marks are by definition perceptual misses. Perceptual hits and perceptual misses are not the same as LL and NL, because the scoring of a physical mark into LL or NL involves defining AR and appropriately classifying the mark. Consider the histogram of the radial distances of marks relative to the nearest lesion centres. The histogram consists of a narrow peak corresponding to perceptual hits that tend to cluster around lesion centres followed by a minimum and a subsequent broad peak corresponding to perceptual misses, which bear no fixed spatial relationship to lesions. The abscissa of the 'valley' between the two peaks is an objective empirical choice for acceptance radius (AR_{avg}). This should be regarded as an average choice, since it includes all lesion sizes. AR_{avg} can be scaled to account for lesion size. It is reasonable to impose a tighter limit for smaller lesions, but not excessively tight since marking 'jitter' makes it impossible to position the mark exactly where the observer intends. It is good practice to repeat the analysis with different choices for AR_{avg}, e.g. 0.8 and 1.2 AR_{avg}, this will test the robustness of the final conclusions with respect to choice of AR.

ACKNOWLEDGEMENT

The author is grateful to Hong-Jun Yoon, MS, for developing the software necessary for this work.

FUNDING

This work was supported by a grant from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688.

REFERENCES

1. Metz, C. E. *ROC methodology in radiologic imaging*. Invest. Radiol. **21**(9), 720–733 (1986).

2. Metz, C. E. *Some practical issues of experimental design and data analysis in radiological ROC studies.* Invest. Radiol. **24**, 234–245 (1989).
3. Metz, C. E. *Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems.* J. Am. Coll. Radiol. **3**, 413–422 (2006).
4. Chakraborty, D. P. and Berbaum, K. S. *Observer studies involving detection and localization: modeling, analysis and validation.* Med. Phys. **31**(8), 2313–2330 (2004).
5. Chakraborty, D. P. *Validation and statistical power comparison of methods for analyzing free-response observer performance studies.* Acad. Radiol. **15**(12), 1554–1566 (2008).
6. Penedo, M. *et al*. *Free-response receiver operating characteristic evaluation of Lossy JPEG2000 and object-based set partitioning in hierarchical trees compression of digitized mammograms.* Radiology **237**(2), 450–457 (2005).
7. Ruschin, M. *et al*. *Dose dependence of mass and microcalcification detection in digital mammography: free response human observer studies.* Med. Phys. **34**, 400–407 (2007).
8. Brennan, P. C., McEntee, M., Evanoff, M., Phillips, P., O'Connor, W. T. and Manning, D. J. *Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist.* Am. J. Roentgenol. **188**(2), W177–W180 (2007).
9. Vikgren, J., Zachrisson, S., Svalkvist, A., Johnsson, A. A., Boijsen, M., Flinck, A., Kheddache, S. and Bath, M. *Comparison of chest tomosynthesis and chest radiography for detection of pulmonary nodules: human observer study of clinical cases.* Radiology **249**(3), 1034–1041 (2008).
10. Egan, J. P., Greenburg, G. Z. and Schulman, A. I. *Operating characteristics, signal detectability and the method of free response.* J. Acoust. Soc. Am. **33**, 993–1007 (1961).
11. Bunch, P. C., Hamilton, J. F., Sanderson, G. K. and Simmons, A. H. *A free-response approach to the measurement and characterization of radiographic-observer performance.* J. Appl. Photogr. Eng. **4**(4), 166–171 (1978).
12. Chakraborty, D. P. *Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data.* Med. Phys. **16**(4), 561–568 (1989).
13. Chakraborty, D. P. and Winter, L. H. L. *Free-response methodology: alternate analysis and a new observer-performance experiment.* Radiology **174**, 873–881 (1990).
14. Swensson, R. G. *Unified measurement of observer performance in detecting and localizing target objects on images.* Med. Phys. **23**(10), 1709–1725 (1996).
15. Edwards, D. C., Kupinski, M. A., Metz, C. E. and Nishikawa, R. M. *Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model.* Med. Phys. **29**(12), 2861–2870 (2002).
16. Dorfman, D. D., Berbaum, K. S. and Metz, C. E. *ROC characteristic rating analysis: generalization to the population of readers and patients with the Jackknife method.* Invest. Radiol. **27**(9), 723–731 (1992).
17. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans* (Montpelier: Capital City Press) (1982).
18. Samuelson, F. W. and Petrick, N. *Comparing image detection algorithms using resampling.* 2006 IEEE International Symposium Biomedical Imaging: From Nano to Micro, pp. 1312–1315 (2006).
19. Chakraborty, D. P. *A search model and figure of merit for observer data acquired according to the free-response paradigm.* Phys. Med. Biol. **51**, 3449–3462 (2006).
20. Chakraborty, D. P. *ROC curves predicted by a model of visual search.* Phys. Med. Biol. **51**, 3463–3482 (2006).
21. Kundel, H. L., Nodine, C. F., Conant, E. F. and Weinstein, S.P. *Holistic component of image perception in mammogram interpretation: gaze-tracking study.* Radiology **242** (2), 396–402 (2007).
22. Dorfman, D. D. and Alf, E. *Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data.* J. Math. Psychol. **6**, 487–496 (1969).
23. Roe, C. A. and Metz, C. E. *Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation.* Acad. Radiol. **4**, 298–303 (1997).
24. Chakraborty, D. P. and Yoon, H. J. *JAFROC analysis revisited: figure-of-merit considerations for human observer studies.* Proceeding of the SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment, **7263**, 72630T (2009).
25. Fryback, D. G. and Thornbury, J. R. *The efficacy of diagnostic imaging.* Med. Decis. Making **11**(2), 88–94 (1991).
26. Hillis, S. L. and Berbaum, K. S. *Power estimation for the Dorfman-Berbaum-Metz method.* Acad. Radiol. **11**(11), 1260–1273 (2004).
27. Chakraborty, D. P., Yoon, H. J. and Mello-Thoms, C. *Spatial localization accuracy of radiologists in free-response studies: inferring perceptual FROC curves from mark-rating data.* Acad. Radiol. **14**, 4–18 (2007).