Published in final edited form as:

Acad Radiol. 2008 December ; 15(12): 1554–1566. doi:10.1016/j.acra.2008.07.018.

Validation and statistical power comparison of methods for analyzing free-response observer performance studies

Dev P. Chakraborty, Ph.D.

University of Pittsburgh, Department of Radiology, 3520 Forbes Ave, Suite 109, Pittsburgh, PA 15261, dpc10@pitt.edu, 412-605-1553 (phone), 412-605-1554 (fax)

Abstract

Rationale and Objectives—The aim of this work was to validate and compare the statistical powers of proposed methods for analyzing free-response data using a search-model based simulator.

Materials and Methods—A free-response data simulator is described that can model a single reader interpreting the same cases in two modalities, or two CAD algorithms, or two human observers, interpreting the same cases in one modality. A variance components model, analogous to the Roe and Metz receiver operating characteristic (ROC) data simulator, is described, that models intra-case and inter-modality correlations in free-response studies. Two generic observers were simulated: a quasi-human observer and a quasi-CAD algorithm. Null hypothesis (NH) validity and statistical powers of ROC, jackknife alternative free-response operating characteristic (JAFROC), a variant of JAFROC termed JAFROC-1, initial detection and candidate analysis (IDCA) and a non-parametric (NP) approach were investigated.

Results—All methods had valid NH behavior over a wide range of simulator parameters. For equal numbers of normal and abnormal cases, for the human observer, the statistical power ranking of the methods was JAFROC-1 > JAFROC > (IDCA ~ NP) > ROC. For the CAD algorithm the ranking was (NP ~ IDCA) > (JAFROC-1 ~ JAFROC) > ROC. In either case the statistical power of the highest ranked method exceeded that of the lowest ranked method by about a factor of two. Dependence of statistical power on simulator parameters followed expected trends. For data sets with more abnormal cases than normal cases, JAFROC-1 power significantly exceeded JAFROC power.

Conclusion—Based on this work the recommendation is to use JAFROC-1 for human observers (including human-observers with CAD assist) and the NP method for evaluating CAD algorithms.

Keywords

observer performance; evaluation methodologies; validation; statistical power; free-response; FROC; JAFROC; CAD evaluation

1 INTRODUCTION

Free-response data consists of mark-rating pairs (1,2), classified into lesion localizations (LLs) or non-lesion localizations (NLs) according to their proximity to real lesions. The importance of assessing computer aided detection (CAD) algorithms for mammography and low-dose computed tomography (CT) screening for lung cancer, and other applications, e.g., breast

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

tomosynthesis and breast CT, has spurred renewed interest in methods for analyzing free-response data. Earlier methods (3–5) assumed that the ratings of multiple marks on the same case were independent. This assumption drew criticism (5,6) that discouraged usage, but end-users of observer performance methodologies were faced with the dilemma of using questionable analyses or ignoring the location information and using receiver operating characteristic (ROC) analysis instead. CAD algorithm evaluation for lung cancer screening presents this problem in an especially acute form. These algorithms generate tens of marks per image and ignoring them is not a credible option. Fortunately several new approaches to the analysis of free-response data are now available that do not depend on the independence assumption. However, these methods need to be validated and their statistical powers need to be compared.

Validation and statistical power studies are done via simulations which require a credible data simulator, or equivalently, a credible observer-model. For example, the binormal model is widely used in the ROC context as it has had great success in fitting ROC data (7–10). The search-model (11–13) is a candidate for a free-response data simulator. It is a mathematical parameterization of a perceptual model (14–16) supported by eye-tracking measurements of how radiologists interpret images. The aims of this work were to develop a search-model based simulator and use the simulator to validate and compare the statistical powers of proposed methods for analyzing free-response data. The methods investigated were the jackknife alternative free-response operating characteristic (JAFROC) (17), a variant of JAFROC termed JAFROC-1 (17), initial detection and candidate analysis (IDCA) (18) and a bootstrap-based non-parametric (NP) method (19,20). For comparison with an existing standard the ROC method (21) was also included.

2 MATERIALS AND METHODS

The free-response receiver operating characteristic (FROC) curve is defined (2) as the plot of lesion localization fraction (LLF) along the y-axis vs. non-lesion localization fraction (NLF) along the x-axis, where the denominators for the fractions are the total number of lesions and the total number of images, respectively. An FROC curve starts at the origin and as the confidence level is reduced the operating point moves up the curve and terminates at the *end-point* (22).

2.1 The search-model simulator

The simulator is characterized by two levels of sampling. One is the random number of *decisions-sites* per image (equivalent terms are “suspicious regions” or “regions that are considered for marking” or “perceived lesions”; the term “lesion” is reserved for a true lesion), referred to as *n-sampling*, and the other is the random *decision-variable* or *z-sample* (equivalent terms are “confidence level”, or “rating”) realized at each decision-site, referred to as *z-sampling*. The modalities and cases are indexed by *i* and *kt*, respectively. The index *t* is the *case-truth*, since it refers to the case as a whole (normal or abnormal), not to specific decision-sites (henceforth abbreviated to *sites*) in the case. For *t* = 1, N_N normal cases are indexed by $k = 1, 2, \dots, N_N$, and for *t* = 2, N_A abnormal cases are indexed by $k = 1, 2, \dots, N_A$. A combination of modality *i* and case *kt* will be abbreviated $\langle ikt \rangle$. The sites are indexed by ℓs , where *s* is the site-truth index, i.e., *s* = 1 denotes a noise-site (a normal region that was considered for marking), and *s* = 2 denotes a signal-site (a lesion that was considered for marking). Denote by n_{ikts} the number of sites with site-truth *s* on $\langle ikt \rangle$, then $\ell = 1, 2, \dots, n_{ikts}$.

In the simulator the number of noise-sites n_{ikt1} is realized by sampling a Poisson random number generator with mean λ_i :

$$n_{ikt1} \sim \text{Poi}(\lambda_i). \quad \text{Eqn. 1}$$

The number of signal-sites n_{ikt2} is realized by sampling a binomial random number generator with trial size N_{kt} and probability of success v_i , where N_{kt} is the number of lesions in case $\langle kt \rangle$ (since normal cases do not contain lesions, $N_{k1} = 0$), and v_i is the probability that a lesion is a signal-site (i.e., it is considered for marking):

$$n_{ikt2} \sim \text{binomial}(N_{kt}, v_i). \quad \text{Eqn. 2}$$

The notation allows for the possibility that λ and v may depend on the modality, but a possible dependence of λ on case-truth is not considered in this work.

An R-rating free-response study requires R ordered cutoffs $\zeta_1, \zeta_2, \dots, \zeta_R$. The simulator realizes a z-sample at each site. A site is marked if and only if its z-sample exceeds the observer's lowest reporting threshold ζ_1 . The *rating* assigned to a mark is the value of z binned according to the cutoffs, or the actual value could be reported as with a CAD algorithm. Marks made because of noise-site z-samples exceeding ζ_1 are classified as NLs and those made because of signal-site z-samples exceeding ζ_1 are LLs. Let $z_{ikt\ell s}$ denote the z-sample for modality i, case k, case-truth t, site-index ℓ , and site-truth index s, abbreviated $\langle ikt\ell s \rangle$. For example, $z_{ikt\ell 1}$ represents the z-sample from the ℓ^{th} noise-site and $z_{ikt\ell 2}$ that from the ℓ^{th} signal-site in $\langle ikt \rangle$. Since $\langle ikt \rangle$ are the same, the two samples are referring to the same physical image (or images in a multi-view display). Since a noise-site and signal-site are being realized on the same case, it must be an abnormal case ($t = 2$). Although the ℓ index is the same for the two samples, they obviously refer to different physical regions. For a given modality, case, and site-truth, identical values of ℓ refer to the same physical region, and different values of ℓ and/or different values of s refer to different physical regions. A normal case can only have noise-sites, so the site-truth index must be unity, i.e., $t = 1$ implies $s = 1$. However, an abnormal case ($t = 2$) can have both noise-sites and signal-sites.

The search-model simulator uses an *equal-variance* binormal model for the z-samples, i.e., the z-samples for noise-sites are sampled from $N(0,1)$ and z-samples for signal-sites are sampled from $N(\mu_i, 1)$, where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 , i.e.,

$$z_{ikt\ell 1} \sim N(0,1), \quad \text{Eqn. 3(a)}$$

$$z_{ikt\ell 2} \sim N(\mu_i, 1). \quad \text{Eqn. 3(b)}$$

For modality i, the simulator predicted FROC curve is defined by (12,13,22):

$$NLF_i(z) = \lambda_i \Phi(-z), \quad \text{Eqn. 4(a)}$$

$$LLF_i(z) = v_i \Phi(\mu_i - z). \quad \text{Eqn. 4(b)}$$

Inter-modality and intra-case correlations of the z_{iktl} s are modeled using a localization variance components method analogous to that developed for the ROC case (23) by Roe and Metz, see Appendix.

2.2 Simulated observers

Two classes of generic “observers” were simulated: (a) a quasi-human observer who considers for marking, on the average, 1.3 actually normal regions (noise-sites) per image in the first modality, and (b) a quasi-CAD algorithmic “observer” that considers for marking, on the average, 10 noise-sites per image in the first modality. The two modalities are labeled 1 and 2, where modality-1 had lower performance. For the human observer the term “modality” has the conventional meaning (e.g., with and without CAD assist) and for CAD it refers to two algorithms that one is interested in comparing. Modality-1 parameters, see Table 1, were chosen so that for each observer class, and one lesion per abnormal image, the area under the search-model predicted ROC curve (11,12), was 80%, i.e., $AUC_1 = 0.80$ (AUC = area under ROC curve). Parameters for the modality-2 observers were chosen to yield $AUC_2 = 0.85$, in other words the effect size was 0.05. [The CAD algorithm developer generally has access to the mark-rating pairs of many suspicious regions found by the program, the vast majority of which would not be shown to the radiologist. This type of detailed information is termed *designer-level* data (13,18). All references to CAD in this paper are to designer-level CAD.]

Simulator predicted FROC curves are shown in Fig. 1 where (a) is for the human observer and (b) is for the CAD algorithms. The lowest threshold ζ_1 determines the *truncation fraction*, Ψ (ζ_1), which is the fraction of the maximum horizontal extent of the simulator generated FROC curve that is *not* observed, see Eqn. 4(a). Values chosen for ζ_1 , namely $-\infty$, -0.674 , 0.0 , and 0.674 , corresponded to truncation fractions 0%, 25%, 50% and 75%, respectively. A larger ζ_1 characterizes a conservative observer who uses a stricter reporting style and provides an *incomplete* curve. The *complete* curve corresponds to $\zeta_1 = -\infty$. Modality-2 has smaller λ and larger v (see Table 1) and therefore both complete and incomplete curves end to the upper-left of the corresponding modality-1 curves. Fig. 1 depicts the alternative hypothesis (AH) condition where the modalities are different. For the null hypothesis (NH) condition the modality-2 parameters were set equal to the modality-1 parameters, and the modality-2 curves would collapse on to the modality-1 curves.

2.3 Analysis of free-response data

Each method involves (i) defining a figure-of-merit and (ii) prescribing a method for testing the significance of observed differences in figures-of-merit (significance testing).

2.3.1 ROC—The ROC-equivalent rating is assumed to be the rating of the highest rated mark in the image (5,11,12). The figure-of-merit is AUC. For the simulated human observer the data was binned to ≤ 6 bins and an implementation of the maximum likelihood binormal model (24) procedure was used to estimate the area under the ROC curve (AUC). The implementation of the binormal model was based on the ROCFIT Fortran source code that used to be posted on the University of Chicago ROC website, but is no longer available. It was checked against the online ROC calculator (25) which is based on ROCFIT. For the simulated CAD algorithms the data was not binned and the trapezoidal AUC was used.

2.3.2 JAFROC—In JAFROC (17) the figure-of-merit A'_1 is defined as the probability that lesions are rated higher than NL marks on *normal* images. Note that the NL marks on abnormal images are not used. A'_1 is estimated by the non-parametric Mann-Whitney-Wilcoxon U-statistic, which involves comparisons between lesion ratings and the ROC-equivalent ratings of normal images. Unmarked lesions and unmarked normal images are formally assigned the “negative infinity” rating. The figure-of-merit is defined by

$$A'_1 = \frac{1}{N_N N_L} \sum_{i=1}^{N_N} \sum_{j=1}^{N_L} \psi(X_i, Y_j) \left. \vphantom{\sum_{i=1}^{N_N} \sum_{j=1}^{N_L}} \right\} \psi(X, Y) = \begin{cases} 1.0 & \text{if } Y > X \\ 0.5 & \text{if } Y = X \\ 0.0 & \text{if } Y < X \end{cases}$$

Eqn. 5

X_i is the ROC-equivalent rating for normal case i , Y_j is the rating of the j^{th} lesion, N_N is the number of normal images and N_L is the number of lesions. [The reason for the prime is because A'_1 is the non-parametric estimate of the area under the alternative free-response receiver operating characteristic (AFROC) curve, previously denoted A_1 (3,4), *except* in JAFROC only normal images are used to estimate the x-coordinate, i.e., false positive fraction (FPF), of the AFROC curve, whereas in the original definition all images are used. The AFROC curve is the plot of LLF vs. FPF. The AFROC abscissa is identical to that of the ROC curve generated by the ROC-equivalent ratings (22).]

2.3.3 JAFROC-1—This is similar to JAFROC except that the highest rated NLs on abnormal images are included in the figure-of-merit, which is now defined by

$$A_1 = \frac{1}{(N_N + N_A) N_L} \sum_{i=1}^{N_N + N_A} \sum_{j=1}^{N_L} \psi(X_i, Y_j).$$

Eqn. 6

Here X_i ($i = N_N, N_N+1, \dots, N_N+N_A$) is the highest NL rating on abnormal case i . Since it uses the highest rated NL on all images, JAFROC-1 is expected to have greater power than JAFROC. The difference is expected to be larger when the number of normal cases is small. In the limit of no normal cases the JAFROC figure-of-merit, Eqn. 5, cannot be calculated but that for JAFROC-1, Eqn. 6, can.

2.3.4 IDCA—Define F the total number of NLs in the image set, T the total number of LLs and N_I the total number of images ($N_I = N_N + N_A$). The observed end-point is

$$\left. \begin{aligned} \lambda' &= \frac{F}{N_I} \\ \nu' &= \frac{T}{N_L} \end{aligned} \right\}.$$

Eqn. 7

The primed notation is needed because the observed end-point need not coincide with the simulator end-point, see for example the observed end-point in Fig. 1(a) which is considerably to the left of the true end-point at (1.3, 0.8). Under the binormal-model assumption the z -samples for noise-sites are sampled from $N(0,1)$ and z -samples for signal-sites are sampled from $N(\mu', \sigma'^2)$. The fitted FROC curve (13, 18) is:

$$\left. \begin{aligned} NLF(z) &= \lambda' \Phi(-z) \\ LLF(z) &= \nu' \Phi\left(\frac{\mu' - z}{\sigma'}\right) \end{aligned} \right\}.$$

Eqn. 8

The parameters μ' and σ' were estimated by regarding the NL and LL marks as normal and abnormal “images” and fitting their ratings to the binormal model (24) using the binormal model maximum likelihood program, whose output quantities, a and b , are related to μ' and σ

' by $\mu' = a/b$ and $\sigma' = 1/b$. Alternatively “proper” ROC models may be used to fit the pseudo-ROC (26–30). The fitted curve is referred to as a *pseudo-ROC* since the marks do not represent images; rather they represent reported suspicious regions in the images, i.e., parts of images. Note (i) the fitted FROC curve is a scaled replica of the pseudo-ROC, with scaling factors λ' and v' along the x and y axes, respectively; (ii) the fitted curve ends at the observed end-point (this can be seen by setting $z = -\infty$ in Eqn. 8) and cannot extend beyond it. The maximum-likelihood procedure requires the data to be binned. The binning ensured at least 5 counts in each of the two cells defined by adjacent cutoffs. For human observers the number of bins was constrained to be ≤ 6 and the pseudo-ROC curve was fitted by the binormal model. For CAD the number of bins was the maximum supported by the data subject to the limit that it be less than 20. The figure-of-merit (19) was the partial area $AUFC_\gamma$ under the *fitted* FROC curve to the left of $NLF = \gamma$, see Fig. 2(a). $AUFC_\gamma$ was evaluated by numerical integration of the curve predicted by Eqn. 8.

2.3.5 NP—In non-parametric (NP) analysis (19,20) $AUFC_\gamma$ is estimated by connecting adjacent points with straight lines and calculating the trapezoidal area under the curve. In their studies Samuelson et al used $NLF = 0.5$ NLs per image as the γ -value (20). Fig. 2(b) is intended to exaggerate the discontinuous nature of the trapezoidal area. It applies to CAD simulated data artificially binned to yield 4 operating points. In fact for CAD no binning was performed for NP analysis (for IDCA analysis the data has to be binned), and since there were many more closely spaced operating points the trapezoidal “curve” is smoother and closer to the true curve.

2.3.6 Choice of upper limit of integration γ —We anticipate a result that *for maximum statistical power it is desirable to choose γ as large as is possible*. However, if γ is chosen to the right of the observed end-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 favor of the modality with the larger γ . The maximum common value of γ is limited by the observed end-point of the observer who generates the least number of NLs per image, and the data to the right of this value provided by other observers has to be discarded from the analysis. Since the simulator generated curve ends at $NLF = \lambda$, the maximum value of γ is determined by the modality with the smaller λ , i.e., λ_2 in the present case. At the end-point ζ_1 is negative infinity. Including the effect of finite ζ_1 , the maximum value for γ is $\lambda_2\Psi(-\zeta_1)$, see Eqn. 4. Due to sampling considerations γ has to be chosen somewhat smaller as otherwise occasionally the observed end-point NLF may be less than $\lambda_2\Psi(-\zeta_1)$ in which case integration to $\lambda_2\Psi(-\zeta_1)$ would fail. Therefore we choose

$$\gamma = \lambda_2\Phi(-\zeta_1)/1.2. \quad \text{Eqn. 9}$$

If the end-point for any of the re-sampled datasets was to the left of this value, the simulation was repeated. This usually happened at the higher ζ_1 values and high correlations.

2.3.7 Significance testing—For all except the NP method, significance testing was performed using the jackknife, which is a resampling technique (31) in which cases are successively removed (“jackknifed”), the figure-of-merit is recomputed, and a pseudovalue is calculated. The pseudovalues are analyzed by an analysis of variance (ANOVA) technique, namely the Dorfman-Berbaum-Metz (DBM) method (21), which for a pair of modalities and a single reader simplifies to a paired t-test on the two pseudovalue arrays, each of length $N_N + N_A$, one per modality. If the p-value of the paired t-test is less than 5% the NH is rejected. The average fraction of rejections over 2000 simulations yields the empirical NH rejection rate. For NP the bootstrap method was used for significance testing (31). For each simulation 2000 bootstrap samples were generated and two cut-points c_1 and c_2 were determined such that 2.5%

and 97.5% of the figures-of-merit were below c_1 and above c_2 , respectively. If the 95% confidence interval defined by (c_1, c_2) does not include zero the NH is rejected. The average fraction of rejections over 2000 simulations yields the empirical NH rejection rate.

3 RESULTS

For the human observer, null hypothesis rejection rates are summarized in Table 2 for different choices of lowest cutoff ζ_1 and common correlation ($\rho = \rho_{\text{inter}} = \rho_{\text{intra}}$). Table 3 shows corresponding NH rejection rates for the CAD observer. For either human observer or CAD, provided the appropriate method was used for significance testing, as indicated in top rows of Table 2 and Table 3, all methods had excellent NH behavior.

For the human observer binned data, use of the jackknife with the NP method led to average NH rejection rate = 0.0262 (range = 0.0025, 0.092). For continuous human observer data the corresponding values were 0.0225 (range = 0.0015, 0.0925). In either case use of the jackknife led to unacceptable NH behavior. For continuous CAD data, using the jackknife for the NP method yielded an average rejection rate, of 0.0587, range = (0.0310, 0.1035). While the average is more reasonable, the jackknife method led to unacceptable NH behavior in most cases. Use of the bootstrap resolved these problems.

Tables 4 and 5 list the statistical powers for the human observer and the CAD observer, respectively. The shading indicates the section of the table expected to be most relevant for the observer in question. For the human observer $\zeta_1 = -\infty$ is not relevant as radiologists do not mark extremely low confidence level suspicious regions. Based on our experience $\zeta_1 > 0$ is a reasonable choice for the relevant region, Table 4. For designer-level CAD data the opposite is true, Table 5. For the human observer, Table 4, the statistical powers, averaged over the relevant region, for the ROC, JAFROC-1, JAFROC, IDCA and NP methods were 0.209, 0.408, 0.369, 0.316 and 0.312, respectively and the ranking was JAFROC-1 > JAFROC > (IDCA ~ NP) > ROC. The values averaged over the entire table were 0.231, 0.455, 0.403, 0.396 and 0.375, respectively, with the same ranking. For the CAD observer, Table 5, the values averaged over the relevant portion were 0.404, 0.735, 0.714, 0.866 and 0.884, respectively, with ranking (NP ~ IDCA) > (JAFROC-1 ~ JAFROC) > ROC. When averaged over the entire table the values were 0.403, 0.722, 0.701, 0.805 and 0.816, respectively, with the same ranking.

Under the AH condition and the CAD observer, as γ , the upper limit of the $\text{AUF}_{C\gamma}$ integral, increases the statistical power of IDCA increases by more than a factor of two over the indicated range, Fig. 3. With larger γ more data, i.e., mark-rating pairs, is included in the analysis, resulting in a more stable estimate of the figure-of-merit and greater sensitivity at detecting differences in figures-of-merit. A similar effect is known to exist for the partial area measure used in ROC analysis (32). This is the reason for advocating the maximum possible value for γ , see Section 2.3.6.

Fig. 4(a) shows that for the human observer JAFROC power decreases with increasing ρ_{intra} . This is because as the ratings of the marks on a case become more correlated they tend to duplicate the information instead of providing new information. This illustrates the validity of the criticism of earlier FROC analysis methods that neglected intra-image correlations: these methods would not predict the decrease in statistical power with ρ_{intra} . Power is greater for $\rho_{\text{inter}} = 0.99$ because when the ratings for a case in the two modalities are highly correlated, the inter-modality figure-of-merit difference is more stable, which allows the modality effect to be more likely to be detected. Fig. 4(b) shows the variation of JAFROC power with ρ_{inter} . Power increases with increasing ρ_{inter} and is smaller when ρ_{intra} is large, consistent with the explanations for Fig. 4(a).

Fig. 5 shows the statistical powers of JAFROC and JAFROC-1 as a function of the number of abnormal cases, *when the total number of cases was held constant at 200*. JAFROC power peaks at around 120 abnormal cases, but JAFROC-1 power increases monotonically. JAFROC-1 power is greater than that of JAFROC over the entire range, significantly so when the number of normal cases is small. In the extreme case when all cases are abnormal, JAFROC-1 is still a valid method but JAFROC is unable to analyze the data.

4 DISCUSSION

Provided the appropriate resampling method was used for significance testing, see top row of Tables 2–5, all methods were statistically valid, i.e., the null hypothesis rejection rates were close to the nominal α -value (0.05). For the human observer the statistical power ranking of the methods was JAFROC-1 > JAFROC > (IDCA ~ NP) > ROC. The power of JAFROC-1 (0.408 in the relevant region) was almost twice that of the ROC method (0.209). For the CAD observer the ranking was (NP ~ IDCA) > (JAFROC-1 ~ JAFROC) > ROC, with power of NP (0.884) also about twice that of the ROC method (0.404). These numbers demonstrate the substantial power penalty of using the ROC method to analyze free-response data. Based on this work the recommendation is to use JAFROC-1 for human observers (including human-observers with CAD assist) and the NP method for evaluating designer-level CAD algorithms (IDCA is a close second but since parametric assumptions are involved the NP method is preferable).

In using the NP method, for optimal power it is important to choose γ , the upper limit of the integration, as large as possible, and in this work values ≥ 5 NLs per image was used, see Table 5. If one used $\gamma = 0.5$ NLs per image, power would be roughly half the optimal power, see Fig. 3. The visual difference between the CAD curves shown in Fig. 1 is fairly obvious, even though the difference in ROC areas is only 0.05 (0.80 $\rightarrow$ 0.85). With 200 cases the maximum power in the relevant region was around 90%. This suggests that 200 cases may be a minimum for credible CAD evaluation studies. If the anticipated effect size is smaller, or if one desires to restrict the range of integration by choosing a smaller γ , then even more cases will be needed.

For the human observer, statistical power of JAFROC-1 (0.455) was greater than either IDCA (0.396) or NP (0.375). In the relevant range (shaded section of the table) the differences were even greater. Since JAFROC-1 considers only the highest rated NL mark, whereas IDCA and NP use all the marks, one might have expected IDCA and NP to have greater power. One reason is that while JAFROC-1 ignores all but one NL mark, the mark that *is* used is a special one, as it is the highest rated mark, and therefore contains more information about the distribution than a randomly selected mark.

Therefore the loss of statistical power due to non-use of all the NLs may be less serious than previously believed (13,17). Another reason is due to differences in the two figures-of-merit. The JAFROC-1 figure-of-merit is the total area under the AFROC (A_1) while the IDCA figure-of-merit is the partial area under the FROC, $AUFC_\gamma$. Unlike A_1 , $AUFC_\gamma$ is relatively insensitive to the *distribution of the marks between the cases*. For example, assuming the total numbers of cases, lesions, NL and LL marks are fixed, all NL marks could occur on a single case without affecting either the FROC curve or $AUFC_\gamma$, but A_1 would increase, as expected, since the observer is perfect on all but one case (a normal case with no marks is a perfect response). JAFROC-1 obtains a rating from *each and every* case, even if the case has no marks. Such cases are assigned the default $-\infty$ rating and the fact that they are less suspicious than any marked lesion tends to increase the figure-of-merit. In contrast, as shown in the above example, unmarked normal cases have no effect on the FROC curve. These effects are most pronounced for the human observer at the highest ζ_1 , which correspond to fewest marks and the most unmarked cases. They are less pronounced for the CAD observer, Table 5, but the statistical

power of JAFROC-1 (0.735) was still respectable compared to that of IDCA (0.866) or the NP (0.884) method.

In earlier work we recommended against using JAFROC-1 because based on our testing then it did not have the correct NH behavior. That study was performed with a different simulator which, for one lesion per abnormal case, forced the number of noise-sites on abnormal cases to be one less than that on normal cases. This “jerkiness” could be the reason for the relatively mild NH failures that were observed (the rejection rates were in the range 0.06 to 0.08; to be conservative we recommending using JAFROC since it had the correct NH behavior). The current simulator is “smoother”: it allows the numbers of noise-sites to differ between normal and abnormal cases, but the samples are statistically smooth since they are from the same Poisson distribution. In hindsight the earlier recommending was unfortunate as it deprived users of the benefits of a statistically valid method that had even greater statistical power than JAFROC. When few normal cases are available, the difference can be considerable, see Fig. 5. This experience also shows the need for care in choosing the data simulator – a pathological simulator can create artificial problems that could be absent in real-world data. Another possibility is to use non-parametric methods, instead of DBM, to perform the significance testing. In theory that should make the method more robust with respect to the choice of simulator, including pathological simulators.

Further testing using other distributions besides the ones considered here, e.g., non-Gaussian z-sampling, non-Poisson n-sampling, etc., is needed. The current simulator is limited to one reader and relatively simple correlations. It needs to be extended to generate multiple-reader multiple-case (MRMC) data with more complex correlations, e.g., correlations of signal and noise-sites *ratings* within the same abnormal case and inter-modality correlations of the *numbers* of noise-sites and signal-sites (19). The simulator should also allow one to study the effect of varying the acceptance radius, i.e., it should also model the spatial distribution of the marks.

Acknowledgments

The author is grateful to Hong-Jun Yoon, MS, for developing the software necessary for this work. This work was supported by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB006388. The author is grateful to Tao Song and bringing to our attention that JAFROC-1 showed valid NH behavior in his simulations.

References

1. Egan JP, Greenburg GZ, Schulman AI. Operating characteristics, signal detectability and the method of free response. *J Acoust Soc Am* 1961;33:993–1007.
2. Bunch PC, Hamilton JF, Sanderson GK, et al. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. *J of Appl Photogr Eng* 1978;4:166–171.
3. Chakraborty DP. Maximum Likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med Phys* 1989;16:561–568. [PubMed: 2770630]
4. Chakraborty DP, Winter LHL. Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment. *Radiology* 1990;174:873–881. [PubMed: 2305073]
5. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996;23:1709–1725. [PubMed: 8946368]
6. Metz, CE. Evaluation of digital mammography by ROC analysis. In: Doi, K., editor. *Digital Mammography '96*. Elsevier Science; Amsterdam, the Netherlands: 1996. p. 61-68.
7. Green, DM.; Swets, JA. *Signal Detection Theory and Psychophysics*. New York: John Wiley & Sons; 1966.
8. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978;8:283–298. [PubMed: 112681]

9. Metz CE. ROC Methodology in Radiologic Imaging. *Investigative Radiology* 1986;21:720–733. [PubMed: 3095258]
10. Metz CE. Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems. *J Am Coll Radiol* 2006;3:413–422. [PubMed: 17412096]
11. Chakraborty DP. ROC Curves predicted by a model of visual search. *Phys Med Biol* 2006;51:3463–3482. [PubMed: 16825743]
12. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol* 2006;51:3449–3462. [PubMed: 16825742]
13. Yoon HJ, Zheng B, Sahiner B, et al. Evaluating computer-aided detection algorithms. *Medical Physics* 2007;34:2024–2038. [PubMed: 17654906]
14. Kundel HL, Nodine CF. A visual concept shapes image perception. *Radiology* 1983;146:363–368. [PubMed: 6849084]
15. Kundel HL, Nodine CF. Modeling visual search during mammogram viewing. *Proc SPIE* 2004;5372:110–115.
16. Kundel HL, Nodine CF, Conant EF, et al. Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology* 2007;242:396–402. [PubMed: 17255410]
17. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis and validation. *Medical Physics* 2004;31:2313–2330. [PubMed: 15377098]
18. Edwards DC, Kupinski MA, Metz CE, et al. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. *Med Phys* 2002;29:2861–2870. [PubMed: 12512721]
19. Samuelson, FW.; Petrick, N. Comparing image detection algorithms using resampling, 2006. *IEEE International Symposium on Biomedical Imaging: From Nano to Micro*; 2006. p. 1312-1315.
20. Samuelson, FW.; Petrick, N.; Paquerault, S. Advantages and examples of resampling for CAD evaluation. 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2007. p. 492-495.
21. Dorfman DD, Berbaum KS, Metz CE. ROC characteristic rating analysis: Generalization to the Population of Readers and Patients with the Jackknife method. *Invest Radiol* 1992;27:723–731. [PubMed: 1399456]
22. Chakraborty DP, Yoon HJ. Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Med Phys* 2008;35:435–445. [PubMed: 18383663]
23. Roe CA, Metz CE. Dorfman-Berbaum-Metz Method for Statistical Analysis of Multireader, Multimodality Receiver Operating Characteristic Data: Validation with Computer Simulation. *Acad Radiol* 1997;4:298–303. [PubMed: 9110028]
24. Dorfman DD, Alf E. Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals - Rating-Method Data. *Journal of Mathematical Psychology* 1969;6:487–496.
25. Eng, J. ROC analysis: web-based calculator for ROC curves. [5/27/08]. <http://www.jrocfits.org>
26. Pan X, Metz CE. The proper binormal model: parametric receiver operating characteristic curve estimation with degenerate data. *Acad Radiol* 1997;4:380–389. [PubMed: 9156236]
27. Dorfman DD, Berbaum KS, Metz CE, et al. Proper Receiving Operating Characteristic Analysis: The Bigamma model. *Acad Radiol* 1997;4:138–149. [PubMed: 9061087]
28. Metz CE, Pan X. Proper Binormal ROC Curves: Theory and Maximum-Likelihood Estimation. *J Math Psychol* 1999;43:1–33. [PubMed: 10069933]
29. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data: Part III. Initial evaluation with detection. *Acad Radiol* 2000;7:438–447. [PubMed: 10845403]
30. Pesce LL, Metz CE. Reliable and Computationally Efficient Maximum-Likelihood Estimation of Proper Binormal ROC Curves. *Academic Radiology* 2007;14:814–829. [PubMed: 17574132]
31. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*. Montpelier: Capital City Press; 1982.
32. Zhou, X-H.; Obuchowski, NA.; McClish, DK. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons; 2002.

33. Roe CA, Metz CE. Variance-Component Modeling in the Analysis of Receiver Operating Characteristic Index Estimates. *Acad Radiol* 1997;4:587–600. [PubMed: 9261459]
34. Matsumoto M, Nishimura T. Mersenne Twister - a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation* 1998;8:3–30.
35. Galassi, M.; Davies, J.; Theiler, J., et al. GNU Scientific Library Reference Manual. Bristol, UK: Network Theory Limited; 2005.

6 APPENDIX

6.1 Variance component model for free-response tasks

Correlations are modeled using a method analogous to that developed for the ROC case (23) by Roe and Metz, and a related idea is contained in Ref. (19). The z -samples are associated with different locations in the case, and therefore a random *location* factor, denoted by L and indexed by ℓs , and associated interaction terms, are needed to describe the location-sampling contribution to the net z -sample. Analogous to the Roe and Metz model (23), the following linear model is proposed:

$$z_{ikt\ell s} = \mu_0 + \Delta\mu_{is} + C_{kt} + (\mu C)_{ikt} + (CL)_{kt\ell s} + (\mu CL)_{ikt\ell s}. \quad \text{Eqn. A1}$$

Since this is a one-reader model the reader factor is absent. Since a location sample can only occur on a case, and not in isolation, terms like L or μL term are not included. Here $(\mu_0 + \Delta\mu_{is})$ represents the population average (over readers and cases) z -sample for a site with site-truth s in modality i and $\Delta\mu_{is}$ is the modality effect that one is interested in detecting.

Specifically, $\mu_0 = 0$, $\mu_0 + \Delta\mu_{11} = 0$, $\mu_0 + \Delta\mu_{12} = \mu_1$, and $\mu_0 + \Delta\mu_{22} = \mu_2$. The remaining terms in the equation represent independent and identically distributed samples from zero-mean

Gaussian distributions with specified variances (the “variance components”) $\sigma_c^2, \sigma_{\mu c}^2, \sigma_{CL}^2$ and $\sigma_{\mu CL}^2$. Specifically $C_{kt} \sim N(0, \sigma_c^2)$ is the contribution of $\langle kt \rangle$ to the net z -sample, $(\mu C)_{ikt} \sim N(0, \sigma_{\mu c}^2)$ is the contribution of $\langle ikt \rangle$, $(CL)_{kt\ell s} \sim N(0, \sigma_{CL}^2)$ is the contribution of site ℓs on $\langle kt \rangle$; and $(\mu CL)_{ikt\ell s} \sim N(0, \sigma_{\mu CL}^2)$ is the contribution of site ℓs , on $\langle ikt \rangle$. The variances for noise-site and signal-site samples are assumed equal (e.g., $\sigma_{\mu CL}^2$ does not involve s), the variances are assumed to be independent of modality (no i -dependence), and correlations between signal-sites and noise-sites on the same case are neglected. There is a normalization constraint on the variance components associated with the case-factor and its interactions, namely

$$\sigma_z^2 \equiv \text{VAR}(z_{ijk\ell s}) = \sigma_c^2 + \sigma_{\mu c}^2 + \sigma_{CL}^2 + \sigma_{\mu CL}^2 = 1. \quad \text{Eqn. A2}$$

This ensures unit net variances of the signal and noise-site samples, as required by the search-model simulator, see Eqns. 3 (a) and 3 (b).

6.2 Correlation MM'_{LS_LS}

(Notation: primed variables are different from variables without primes, M is the modality factor indexed by i , L is the location factor indexed by ℓ and S is the site-truth factor indexed by s .) This correlation, denoted ρ_{inter} , is the *inter-modality* correlation, and is calculated (or empirically estimated from the data) as follows. Consider a pair of sites ℓs and $\ell' s'$,

corresponding to the *same* physical region in the case, one from $\langle ikt \rangle$ and the other from $\langle i' kt \rangle$, and calculate the correlation of the corresponding ratings over cases regarded as random; average over all pairs of such sites and over all modality pairings.

The correlation can be deduced from the variance components as follows. First one calculates the variance of the difference of the z-samples (33):

$$\begin{aligned} \text{VAR}(z_{iktl's} - z_{i'ktls}) &= \text{VAR} \left[\begin{array}{c} \{((\mu C)_{ikts} + (\mu CL)_{ikts})\} \\ - \{((\mu C)_{i'kts} + (\mu CL)_{i'kts})\} \end{array} \right] \\ \text{VAR}(z_{iktl's} - z_{i'ktls}) &= 2\sigma_{\mu C}^2 + 2\sigma_{\mu CL}^2. \end{aligned} \quad \text{Eqn. A3}$$

Factors not shown on the right-hand-side of the first of these equations are: (i) all fixed factors ($\mu, \Delta\mu$) since these do not contribute to the variance and (ii) factors that do not include modality, e.g., C, since they cancel in the difference. The surviving factors are μC and μCL .

The correlation of any two random variables x and y with the same variance, is given by

$$\text{COR}(x, y) = 1 - \frac{\text{VAR}(x - y)}{2\text{VAR}(y)}. \quad \text{Eqn. A4}$$

It follows that the correlation is given by,

$$\rho_{\text{inter}} = \text{COR}(z_{iktl's}, z_{i'ktls}) = \frac{\sigma_C^2 + \sigma_{CL}^2}{\sigma_C^2 + \sigma_{\mu C}^2 + \sigma_{\mu CL}^2 + \sigma_{CL}^2}. \quad \text{Eqn. A5}$$

6.3 Correlation MM_LS_L'S

This is the *intra-case correlation*, ρ_{intra} . The two values of ℓ are different and hence point to different physical locations with the same site-truth index s on the same physical image. The desired correlation is as follows:

$$\begin{aligned} \text{VAR}(z_{iktl's} - z_{iktl's'}) &= \text{VAR} \left[\begin{array}{c} \{((CL)_{kts} + (\mu CL)_{ikts})\} \\ - \{((CL)_{kts'} + (\mu CL)_{ikts'})\} \end{array} \right] \\ \text{VAR}(z_{iktl's} - z_{iktl's'}) &= 2\sigma_{\mu C}^2 + 2\sigma_{\mu CL}^2. \end{aligned} \quad \text{Eqn. A6}$$

Therefore

$$\rho_{\text{intra}} = \text{COR}(z_{iktl's}, z_{iktl's'}) = \frac{\sigma_C^2 + \sigma_{\mu C}^2}{\sigma_C^2 + \sigma_{\mu C}^2 + \sigma_{\mu CL}^2 + \sigma_{CL}^2}. \quad \text{Eqn. A7}$$

6.4 Determining the variance components from specified correlations

To use the sampling model (Eqn. A1) it is necessary to be able to calculate the variances $\sigma_C^2, \sigma_{\mu C}^2, \sigma_{LC}^2$ and $\sigma_{\mu CL}^2$ that correspond to specified inter-modality and intra-case correlations.

Trial values were searched (IDL routine AMOEBA), subject to the constraints that they must be positive and satisfy the normalization condition, until the sum of the squared differences between the predicted (Eqn. A5 and Eqn. A7) and target correlations, plus the squared difference implied by the normalization constraint (Eqn. A2), was minimized.

6.5 Technical details of the simulator

The simulator is capable of generating a random number of lesions per abnormal case subject to the constraints that it must be greater than zero, less than a specified maximum, and have a specified average. The procedure for selecting the cutoffs involved test-runs of the simulator to find modality-specific cutoffs that ensured at least 5 counts in each of the two cells defined by adjacent cutoffs. Subsequently these cutoffs were held fixed for the actual simulations. With large numbers of simulations necessary for NH validation and statistical power determination, especially with the bootstrap method, the period of the random number generator, beyond which the random number sequence repeats itself, becomes an important consideration. In this study we used a simulation quality random number generator (routine gsl_rng_mt19937) based on Ref. (34) which is available in the GNU scientific library (35). The period of this generator is approximately 10^{6000} .

Fig. 1.

Search-model predicted FROC curves for the human observer (a) and CAD (b). The sections labeled $\zeta_1 = 0$ represent the incomplete curves corresponding to 50% truncation fraction. For modality-1 the complete curve extends to NLF = 1.3 for the human observer and to NLF = 10 for CAD. The corresponding incomplete curves extend to 0.65 and 5, respectively. The parameters were chosen so that the areas under the search-model predicted ROC curves for modality 1 was 0.80, and for the modality 2 was 0.85, see Table 1. Four simulated operating points are shown for the human observer in modality-1. The simulations conditions were 100 normal cases, 100 abnormal cases with one lesion per case, $\zeta_1 = 0$, inter-modality and intra-case correlations = 0.5. Note the large difference between observed end-point and true end-

point. Two thousand (2000) simulations like these were performed to determine the NH and AH rejection rates reported in Tables 2 through 5.

Fig. 2.

(a) The IDCA partial area figure-of-merit $AUFC_\gamma$ was the integral under the fitted curve to the left of $NLF = \gamma$. (b) The NP method uses the corresponding trapezoidal area as the figure-of-merit. To exaggerate the discontinuous nature of the NP trapezoidal area construct, this figure shows the trapezoidal area for artificially binned CAD data with 4 operating points. Note that γ must be to the left of the observed end-point and a linearly interpolated area between the last two operating points needs to be included in the integral. In fact for CAD no binning was done for NP analysis, and the high density of operating points ensured that the trapezoidal area was close to the true value.

Fig. 3.

Dependence of power on γ : this figure shows that statistical power of IDCA increases with γ , the upper limit of integration for the AUC_{γ} figure-of-merit. For maximum statistical power it is desirable to include as much of the FROC curve as possible in the range of integration. This data is for the CAD observer and common correlation coefficient equal to 0.5. This data was generated with 10,000 AH trials per data point.

Fig. 4.

(a) This figure shows the variation of JAFROC power with intra-case correlation ρ_{intra} . Power decreases with increasing ρ_{intra} because the ratings of the marks on a case become more correlated and therefore redundant. Power is greater for $\rho_{inter} = 0.99$ because with highly correlated modalities the figure-of-merit difference is more stable, which allows the modality effect to be more likely to be detected. (b) This figure shows the variation of power with inter-modality correlation ρ_{inter} . Power increases with increasing ρ_{inter} and is smaller when ρ_{intra} is large, consistent with the above explanations. Simulation conditions: $\zeta_1 = 0$, and 20,000 AH trials per data point

Fig. 5.

The statistical powers of JAFROC and JAFROC-1 as a function of the number of abnormal cases, when the total number of cases is held constant at 200. JAFROC power peaks at around 120 abnormal cases, but JAFROC-1 power increases monotonically. JAFROC-1 power is always greater than that of JAFROC, but for less than 80 abnormal cases the difference is small. For relatively few normal cases, the power advantage of JAFROC-1 is substantial. The $\pm 2\sigma$ confidence intervals were determined from 21 simulations. The data points shown are averaged over all 12 conditions of correlations and cutoffs for the human observer.

This table shows simulation parameters μ , λ and v for the human observer and CAD for the two modalities (1 and 2) and the corresponding areas (AUC_1 and AUC_2) under the search-model predicted ROC curves. AUC_1 and AUC_2 were calculated numerically from the search-model parameters (11,12). These values were used to generate the data in Tables 2, 3, 4 and 5. Modality-1 parameters were chosen based on our experience with human observer and CAD generated FROC curves. For a given AUC, the choice of search model parameters is not unique. The following rule was used to generate modality-2 parameters from those of modality-1: λ was 20% smaller and v was 10% greater and μ was adjusted to get the desired AUC value.

Parameter values for the two simulated observers							
Observer	Modality 1			Modality 2			
	μ_1	λ_1	v_1	AUC_1	μ_2	λ_2	v_2
	AUC_1				AUC_2		
HUMAN	1.50	1.3	0.80	0.80	1.55	1.04	0.88
CAD	2.34	10	0.90		2.37	8	0.99

This table shows null hypothesis rejection rates for the human observer for different conditions of lowest cutoff ρ_1 and correlations $\rho = \rho_{\text{inter}} = \rho_{\text{intra}}$. The variable γ is the upper limit of integration for the AUC_{γ} figure-of-merit, which applies to IDCA and NP methods only, see Eqn. 9. The simulation conditions were 100 normal cases, 100 abnormal cases, and one lesion per abnormal case, and rejection rates were determined over 2,000 simulations. The procedure used for significance testing is indicated in the top row. The search-model predicted AUCs for both modalities was 0.80. The last row is the average NH rejection rate over all conditions. Note that all methods had excellent NH behavior provided the appropriate method was used for significance testing. Use of the jackknife for the NP method gave poor NH behavior.

Table 2

Results of NH validation for the human observer						
ζ_1 and γ	ρ	Jackknife			IDCA@ γ	Bootstrap NP@ γ
		ROC	JAFROC-1	JAFROC		
$\zeta_1 = -\infty$ $\gamma = 0.867$	0.1	0.0605	0.0513	0.0483	0.0500	0.0625
	0.5	0.0565	0.0468	0.0495	0.0565	0.0660
	0.9	0.0560	0.0528	0.0515	0.0640	0.0555
$\zeta_1 = -0.674$ $\gamma = 0.650$	0.1	0.0435	0.0520	0.0460	0.0455	0.0465
	0.5	0.0470	0.0565	0.0383	0.0505	0.0555
	0.9	0.0495	0.0505	0.0573	0.0540	0.0485
$\zeta_1 = 0.0$ $\gamma = 0.433$	0.1	0.0525	0.0480	0.0505	0.0480	0.0455
	0.5	0.0560	0.0458	0.0473	0.0415	0.0620
	0.9	0.0495	0.0560	0.0563	0.0505	0.0490
$\zeta_1 = 0.674$ $\gamma = 0.217$	0.1	0.0520	0.0548	0.0525	0.0605	0.0480
	0.5	0.0435	0.0448	0.0498	0.0555	0.0575
	0.9	0.0495	0.0458	0.0443	0.0535	0.0470
Average		0.0513	0.0504	0.0493	0.0525	0.0536

This table shows null hypothesis rejection rates for the CAD observer. Other conditions are as in Table 2. Note that all methods had excellent NH behavior.

Table 3

Results of NH validation for the CAD observer						
ζ_1 and γ	p	Jackknife				Bootstrap
		ROC	JAFROC-1	JAFROC	IDCA@ γ	NP@ γ
$\zeta_1 = -\infty$ $\gamma = 6.67$	0.1	0.0545	0.0498	0.0548	0.0495	0.0590
	0.5	0.0590	0.0485	0.0455	0.0505	0.0630
	0.9	0.0355	0.0405	0.0465	0.0460	0.0525
$\zeta_1 = -0.674$ $\gamma = 5$	0.1	0.0475	0.0425	0.0510	0.0450	0.0500
	0.5	0.0590	0.0480	0.0495	0.0550	0.0450
	0.9	0.0550	0.0485	0.0463	0.0560	0.0600
$\zeta_1 = 0.0$ $\gamma = 3.33$	0.1	0.0600	0.0565	0.0520	0.0520	0.0580
	0.5	0.0470	0.0475	0.0488	0.0425	0.0595
	0.9	0.0485	0.0525	0.0500	0.0550	0.0595
$\zeta_1 = 0.674$ $\gamma = 1.67$	0.1	0.0480	0.0458	0.0430	0.0470	0.0515
	0.5	0.0560	0.0568	0.0560	0.0590	0.0440
	0.9	0.0480	0.0493	0.0470	0.0485	0.0445
Average		0.0515	0.0488	0.0492	0.0505	0.0539

Listed are the statistical powers for the human observer for an effect size $\Delta(\text{AUC}) = 0.05$. Other conditions are as in Table 2. The shaded section of the table is relevant to human observers. The next-to-last row is the power averaged over all 12 combinations of ρ and ζ_1 and the last row is the power averaged over the shaded portion of the table. In either cases the ranking of the methods was JAFROC-1 > JAFROC > (IDCA ~ NP) > ROC. The power of the highest ranked method (JAFROC-1) is almost twice that of the lowest ranked method (ROC).

Table 4

Statistical power for the human observer						
ζ_1 and γ	ρ	Jackknife			IDCA @ γ	Bootstrap NP @ γ
		ROC	JAFROC-1	JAFROC		
$\zeta_1 = -\infty$ $\gamma = 0.867$	0.1	0.231	0.517	0.440	0.452	0.417
	0.5	0.252	0.511	0.428	0.490	0.441
	0.9	0.272	0.518	0.470	0.523	0.482
$\zeta_1 = -0.674$ $\gamma = 0.650$	0.1	0.234	0.493	0.411	0.415	0.403
	0.5	0.248	0.485	0.424	0.460	0.417
	0.9	0.283	0.490	0.457	0.516	0.467
$\zeta_1 = 0.0$ $\gamma = 0.433$	0.1	0.224	0.440	0.397	0.349	0.319
	0.5	0.242	0.453	0.377	0.393	0.390
	0.9	0.282	0.495	0.448	0.417	0.402
$\zeta_1 = 0.674$ $\gamma = 0.217$	0.1	0.183	0.314	0.293	0.205	0.210
	0.5	0.161	0.344	0.315	0.249	0.251
	0.9	0.161	0.402	0.382	0.284	0.299
Average (ALL)		0.231	0.455	0.403	0.396	0.375
Average (shaded)		0.209	0.408	0.369	0.316	0.312

Listed are the statistical powers for the CAD observer for an effect size $\Delta(\text{AUC}) = 0.05$. Other conditions are as in Table 3. The shaded section of the table is most relevant to CAD. The ranking of the methods was $(\text{NP} \sim \text{IDCA}) > (\text{JAFROC-1} \sim \text{JAFROC}) > \text{ROC}$. The power of the highest ranked method (NP) is almost twice that of the lowest ranked method (ROC).

Table 5

Statistical power for the CAD observer						
ζ_1 and γ	ρ	Jackknife			IDCA@ γ	Bootstrap NP@ γ
		ROC	JAFROC-1	JAFROC		
$\zeta_1 = -\infty$ $\gamma = 6.67$	0.1	0.251	0.578	0.537	0.815	0.847
	0.5	0.410	0.752	0.725	0.867	0.902
	0.9	0.570	0.869	0.877	0.934	0.954
$\zeta_1 = -0.674$ $\gamma = 5$	0.1	0.244	0.569	0.538	0.801	0.799
	0.5	0.366	0.764	0.727	0.858	0.859
	0.9	0.584	0.878	0.880	0.924	0.945
$\zeta_1 = 0.0$ $\gamma = 3.33$	0.1	0.239	0.600	0.559	0.700	0.712
	0.5	0.375	0.734	0.717	0.787	0.807
	0.9	0.616	0.879	0.873	0.919	0.952
$\zeta_1 = 0.674$ $\gamma = 1.67$	0.1	0.233	0.558	0.522	0.554	0.530
	0.5	0.380	0.660	0.647	0.666	0.629
	0.9	0.574	0.828	0.811	0.835	0.853
Average (ALL)		0.403	0.722	0.701	0.805	0.816
Average (shaded)		0.404	0.735	0.714	0.866	0.884