JAFROC analysis revisited: figure-of-merit considerations for human observer studies

D. P. Chakraborty*[a] and Hong-Jun Yoon[a]
[a]Dept. of Radiology, Univ. of Pittsburgh, 3520 Forbes Ave,
Pittsburgh, PA, USA 15261

ABSTRACT

Jackknife alternative free-response receiver operating characteristic (JAFROC) is a method for measuring human observer performance in localization tasks. JAFROC is being increasingly used to evaluate imaging modalities because it has been shown to have greater statistical power than conventional receiver operating characteristic (ROC) analysis, which neglects location information. JAFROC neglects the non-lesion localization marks ("false positives") on abnormal images. JAFROC1 is an alternative method that includes these marks. Both methods are *lesion-centric* in the sense that they assign equal importance to all lesions; an image with many lesions would tend to dominate the performance metric, and clinically less significant lesions are treated identically as more significant ones. In this paper *weighted* JAFROC and JAFROC1 analyses are described that treat each abnormal image (not each lesion) as a unit of measurement and account for different lesion clinical significances (weights). Lesion-centric and weighted methods were tested using a simulator that includes multiple-reader multiple-case multiple-modality location level correlations. For comparison, ROC analysis was also tested where the rating of the highest rated mark on an image was assumed to be its "ROC" rating. The testing involved random numbers of lesions per image, random weights, case-mixes (ratio of normal to abnormal images) and different correlation structures. We found that for either JAFROC or JAFROC1, both lesion-centric and weighted analyses had correct NH behavior and comparable statistical powers. For either lesion-centric or weighted analyses JAFROC1 yielded the highest power, followed by JAFROC and ROC yielded the least power, confirming a recent study using a less flexible single-reader dual-modality simulator. Provided the number of normal cases is not too small, JAFROC1 is the preferred method for analyzing human observer free-response data. For either JAFROC or JAFROC1 weighted analysis is preferable.

Keywords: Free-response, imaging system assessment, JAFROC, JAFROC1, simulators, lesion localization, lesion-centric analysis, weighted analysis

1. INTRODUCTION

In the free-response paradigm [1-3] the observer provides the locations of sufficiently suspicious findings (i.e., marks) and the associated confidence levels (ratings). By adopting a clinically relevant "nearness" criterion the investigator classifies each mark as lesion localization (LL), if the mark is sufficiently close to a true lesion, and otherwise non-lesion localization (NL). The free-response receiver operating characteristic (FROC) curve [2, 4] is the plot of lesion localization fraction (LLF) vs. non-lesion localization fraction (NLF). It is understood that the denominators in the fractions are the number of lesions and the number of images, respectively (since NLF can exceed unity it is strictly an improper fraction, but for symmetry of notation we refer to both as "fractions"). A derived ROC curve can be obtained by using the rating of the highest rated mark on an image as its "ROC-equivalent" rating and defining true positive fraction (TPF) and false positive fraction (FPF) in the usual manner and the ROC curve is the plot of TPF vs. FPF. The alternative free-response receiver operating characteristic (AFROC) curve [5] is the plot of LLF vs. FPF, i.e., it is a hybrid curve, with a location-specific y-axis and an image-specific x-axis. Since NLs can also occur on abnormal cases, there are two possibilities: counting the highest rated NL only on normal images or on all images. In Ref. [5] the latter definition was used.

JAFROC [6], a method for analyzing human observer FROC data, is being increasingly used [7-16]. (For designer-level computer aided detection algorithms an FROC curve-based non-parametric method is preferable [17].) Data analysis

*dpc10@pitt.edu; phone: 412-605-1553; fax: 412-605-1554; www.devchakraborty.com

Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment,
edited by Berkman Sahiner, David J. Manning, Proc. of SPIE Vol. 7263, 72630T
· doi: 10.1117/12.810859

involves defining a figure-of-merit (FoM) and a method for testing the significance of observed differences in FoMs. In JAFROC the FoM is the non-parametric (Mann-Whitney statistic) probability that a lesion is rated greater than the highest rated NL mark (henceforth abbreviated "highest noise") on a normal image. Equivalently, the FoM is the area under the AFROC curve, where one uses only *normal* images to determine FPF. In JAFROC1, an alternative method proposed in Ref. [6], the non-lesion marks on abnormal images are also included in the probability computation. Recently [17] a search-model based single-reader two-modality simulator (termed simulator-A) was developed using which both JAFROC and JAFROC1 were found to have correct NH behavior, but JAFROC1 had greater power especially when the relative number of normal cases was small. An aim was to extend simulator-A to accommodate multiple-reader multiple-modalities (simulator-B). Another aim was to overcome a limitation of JAFROC and JAFROC1, namely they assign equal importance to all lesions so an image with many lesions would tend to dominate the FoM, and clinically less significant lesions are treated identically as clinically more significant ones (this limitation is shared by all other methods of analyzing FROC data that we are aware of [18-23]. A weighted FoM was proposed in Ref. [6] that theoretically resolved this issue but was not evaluated; accordingly we wished to determine the NH validity and statistical powers of the lesion-centric and weighted JAFROC and JAFROC1 using simulator-B. Finally, we wished to resolve the inconsistent JAFROC1 validity finding of the study reported in Ref. [6].

2. METHODS

2.1 Notation

In this paper each case is assumed to contribute one image, so *patient*, *case* and *image* are synonymous. Modalities, readers and cases are indexed by *i*, *j* and <*kt*>, respectively. The *case-truth* index *t* refers to the case (or patient) as a whole (normal or abnormal), not to specific locations in the case. For *t* = 1, N_N normal cases are indexed by *k* = 1, 2, ..., N_N, and for t = 2, N_A abnormal cases are indexed by *k* = 1, 2, ..., N_A. N_k is the number of lesions in abnormal case *k*. The total number of lesions in the data set is N_L, and

$$\sum_{k=1}^{N_A} N_k = N_L \quad . \qquad \text{Eqn. 1}$$

Decision-sites [24, 25] (abbreviated "*sites*") are suspicious regions identified ("hit" or "seen") by the observer in the initial search phase of the interpretation. A site that corresponds to normal anatomy is a *noise-site* and one that corresponds to a lesion is a *signal-site*. In the second phase and a decision variable or *z-sample* is calculated for each site. Sites are labeled by the *location index* ℓ (ℓ = 1, 2, ...) and a *site-truth* index s which determines the *type* of the site, i.e., *s* = 1 for a noise-site and *s* = 2 for a signal-site. The combination of modality *i*, reader *j*, case <*kt*> and site <*ℓs*> is abbreviated to <*ijktℓs*>. The decision variable sample at <*ijktℓs*> is denoted z_{ijktls}. The number of noise-sites on <*ijkt*> is denoted n_{ijkt1} and the number of signal-sites is denoted n_{ijkt2}. Therefore, the range of *l* is *l* = 1, 2, …, $n_{ijkt1} + n_{ijkt2}$. If $n_{ijkt1} + n_{ijkt2} = 0$ no values of *l* are possible and there are no z-samples for that <ijkt>. In general one does not know z_{ijktls}, rather the integer ratings r_{ijktls} of marked sites. For an R-rating FROC study one defines R+2 cutoffs ζ_{ijm} (m = 0, 1, 2, …, R+1) and defines $\zeta_{ij0} = -\infty, \zeta_{ij(R+1)} = \infty$, and adopts the rule that a site is marked and rated r (r = 1, 2, ..., R) if $\zeta_{ijr} \le z_{ijktls} < \zeta_{ij(r+1)}$ (if $z_{ijktls} < \zeta_{ij1}$ the site is not marked). W_{kl} is the weight (clinical importance) of lesion *l* in abnormal case k and the weights on any given abnormal case must add up to unity

$$\sum_{l=1}^{N_k} W_{kl} = 1 \quad . \qquad \text{Eqn 2}$$

Since some lesions may be missed during the initial search, i.e., $n_{ijkt2} \equiv n_{ijk22} < N_L$, and

$$\left.\begin{array}{l}\sum_{k=1}^{N_A} n_{ijk22} \le N_L \\ \sum_{l=1}^{n_{ijk22}} W_{kl} \le 1\end{array}\right\} \quad . \qquad \text{Eqn. 3}$$

Note that s = 2 forces t = 2.

2.2 The figures-of-merit

All FoMs considered in this work are non-parametric (Mann–Whitney statistics), i.e., equivalent to the trapezoidal areas under the relevant operating characteristic.

ROC

The ROC FoM AUC_{ij}^{ROC} is defined as

$$\left.\begin{array}{l} AUC_{ij}^{ROC} = \dfrac{1}{N_A}\sum_{k_2=1}^{N_A}\dfrac{\sum_{k_1=1}^{N_N}\psi\left(max(r_{ijk_1 1*1}), max(r_{ijk_2 2**})\right)}{N_N} \\ \psi(X,Y) = \begin{bmatrix} 1.0\ if\ Y > X \\ 0.5\ if\ Y = X \\ 0.0\ if\ Y < X \end{bmatrix} \end{array}\right\} \quad . \qquad \text{Eqn. 4}$$

The max function is the maximum over the indices indicated by the asterisks. If a normal image has no marks the max function is assigned the -∞ rating as are unmarked lesions. If all lesions are marked and no noise sites are marked, signifying perfect performance, the ψ function is unity, and the FoM is unity. If no lesions are marked and the distribution of the numbers and ratings of NL marks is the same for normal and abnormal images, signifying the observer is unable to discriminate between normal and abnormal images, the ψ function comparisons between highest noise samples from the normal and abnormal images yield 0.5, on the average, implying $AUC_{ij}^{ROC} = 0.5$, which is the worst possible ROC performance. The ROC figure or merit, unlike others to be described below, ranges between 0.5 and unity. Since each image gives one rating, and one does not know which site caused that rating, it is not possible to define weighted FoMs in ROC analysis.

Weighted JAFROC and JAFROC1

In reference [6] the weighted FoM $\theta_{ij}^{JAFROC-WGHT}$ was defined as follows:

$$\theta_{ij}^{JAFROC-WGHT} = \frac{1}{N_A}\sum_{k_2=1}^{N_A}\frac{\sum_{k_1=1}^{N_N}\sum_{l=1}^{N_{k_2}} W_{k_2 l}\,\psi\left(max(r_{ijk_1 1*1}), r_{ijk_2 2l2}\right)}{N_N} \quad . \qquad \text{Eqn. 5}$$

If all lesions are marked and no normal image is marked the ψ function is unity, and it is easily seen that the FoM is unity. If no lesions are marked and every normal image has at least one mark the ψ function is zero and the FoM is zero. This figure or merit, like others to be described below, ranges between 0 and unity.

The extension to include the highest noise on abnormal images, i.e., $\theta_{ij}^{JAFROC1-WGHT}$, is

$$\theta_{ij}^{JAFROC1-WGHT} = \frac{1}{N_A}\sum_{k_2=1}^{N_A}\frac{\sum_{k_1=1}^{N_N}\sum_{l=1}^{N_{k_2}} W_{k_2 l}\psi\left(max(r_{ijk_1 1*1}), r_{ijk_2 2l2}\right) + \sum_{k_1=1}^{N_A}\sum_{l=1}^{N_{k_2}} W_{k_2 l}\psi\left(max(r_{ijk_1 2*1}), r_{ijk_2 2l2}\right)}{N_T} \quad . \qquad \text{Eqn. 6}$$

Lesion-centric JAFROC and JAFROC1

The software currently available on our website (JAFROC1 V1.1) uses lesion-centric (LC) analysis. The JAFROC lesion-centric FoM is defined as the probability that lesions are rated higher than the highest noise on *normal* images:

$$\theta_{ij}^{JAFROC-LC} = \frac{1}{N_L}\sum_{k_2=1}^{N_A}\sum_{l=1}^{N_{k_2}}\frac{\sum_{k_1=1}^{N_N}\psi\left(max(r_{ijk_1 1*1}), r_{ijk_2 2l2}\right)}{N_N} \quad . \qquad \text{Eqn. 7}$$

The corresponding JAFROC1 lesion-centric FoM, which includes the highest noise on abnormal images, is defined by

$$\theta_{ij}^{JAFROC1-LC} = \frac{1}{N_L}\sum_{k_2=1}^{N_A}\sum_{l=1}^{N_{k_2}}\left[\frac{\sum_{k_1=1}^{N_N}\psi\left(max(r_{ijk_1 1*1}), r_{ijk_2 2l2}\right) + \sum_{k_1=1}^{N_A}\psi\left(max(r_{ijk_1 2*1}), r_{ijk_2 2l2}\right)}{N_T}\right] \quad . \qquad \text{Eqn. 8}$$

2.3 The ROC multiple-reader multiple-case multiple-modality data simulator

Before describing the search model simulator (simulator-B) we summarize the conventional ROC simulator [26] which has been used to validate ROC analyses software [26-28]. The z-sample of an image is modeled as

$$z_{ijkt}^{ROC} = \mu_t + \Delta\mu_{it} + R_{jt} + C_{kt} + (\mu R)_{ijt} + (\mu C)_{ikt} + (RC)_{jkt} + (\mu RC)_{ijkt} \quad , \qquad \text{Eqn. 9}$$

with the constraint

$$\Sigma_{ROC,C|}^2 \equiv VAR_{C|}(z_{ijkt}) = \sigma_C^2 + \sigma_{\mu C}^2 + \sigma_{RC}^2 + \sigma_{\mu RC}^2 = 1 \quad . \qquad \text{Eqn. 10}$$

The term $(\mu_t + \Delta\mu_{it})$ represents the average (over reader and case populations) z-sample contribution of <*it*>; $\Delta\mu_{it}$ is the modality effect that one is interested in detecting. The remaining terms in the equation represent random samples from zero-mean Gaussian distributions with specified variances (the *variance components*). For simplicity all variances in the model are assumed to be independent of modality, reader and case-truth. Following Roe and Metz [29], the notation C| signifies that cases are regarded as random and modality and readers as fixed effects. Define $N(0,\sigma^2)$ as the zero mean normal distribution with variance σ^2. Then $R_{jt} \sim N(0,\sigma_R^2)$ is the random contribution of <*jt*>, $C_{kt} \sim N(0,\sigma_C^2)$ is the random contribution of <*kt*>, $(\mu R)_{ijt} \sim N(0,\sigma_{\mu R}^2)$ is the random contribution of <*ijt*>, $(\mu C)_{ikt} \sim N(0,\sigma_{\mu C}^2)$ is the random contribution of <*ikt*>, $(RC)_{jkt} \sim N(0,\sigma_{RC}^2)$ is the random contribution of <*jkt*>, and $(\mu RC)_{ijkt} \sim N(0,\sigma_{\mu RC}^2)$ is the random contribution of <*ijkt*>. A replication error term is included in $(\mu RC)_{ijkt}$ since without replicated readings it cannot be separated from $(\mu RC)_{ijkt}$. To assure unit overall case variance, the constraint on the variance components, as indicated in Eqn. 10, is necessary.

2.4 The FROC multiple-reader multiple-case multiple-modality data simulator (simulator-B)

The number of noise-sites n_{ijkt1} on <ijkt> is modeled as a random sample from a Poisson distribution with mean λ_{ij} : $n_{ijkt1} \sim Poi(\lambda_{ij})$ (alternative distributions, e.g., binomial [23], can be specified at this stage) and the number of signal-sites n_{ijk22} on <ijk2> is modeled as a random sample from a binomial distribution with trial size N_k and success probability ν_{ij} : $n_{ijk22} \sim binomial(N_k, \nu_{ij})$. The notation allows for the possibility that λ and ν can depend on modality and reader. The free-response data simulator is described by

$$\left.\begin{aligned} & z_{ijktls} = \mu_s + \Delta\mu_{is} \\ & +R_{js} + C_{kts} + (\mu R)_{ijs} + (\mu C)_{ikts} + (RC)_{jkts} + (\mu RC)_{ijks} \\ & +(CL)_{ktls} + (\mu CL)_{iktls} + (RCL)_{jktls} + (\mu RCL)_{ijktls} \end{aligned}\right\} , \qquad \text{Eqn. 11}$$

and

$$\left.\begin{aligned} & \Sigma^2_{FROC,C|} \equiv VAR_{C|}(z_{ijktls}) = \Sigma^2_{ROC',C|} + \sigma^2_{CL} + \sigma^2_{\mu CL} + \sigma^2_{RCL} + \sigma^2_{\mu RCL} = 1 \\ & \Sigma^2_{ROC',C|} = \sigma^2_C + \sigma^2_{\mu C} + \sigma^2_{RC} + \sigma^2_{\mu RC} \le 1 \end{aligned}\right\}. \qquad \text{Eqn. 12}$$

In Eqn. 12, $\Sigma^2_{ROC',C|} \le 1$ since the total variance $\Sigma^2_{FROC,C|}$ is distributed between location-independent $\Sigma^2_{ROC',C|}$ and location-dependent $\sigma^2_{CL} + \sigma^2_{\mu CL} + \sigma^2_{RCL} + \sigma^2_{\mu RCL}$ contributions. If $\Sigma^2_{ROC',C|} \sim 1$ most of the variance is due to the case components (terms *not* containing L), therefore the location components variances (terms containing L) must be small, which implies that the z-samples on an image are highly correlated. Conversely, if $\Sigma^2_{ROC',C|} \sim 0$, the z-samples are almost independent.

Since *sites* provide z-samples, not *cases*, each term in Eqn. 11 must have a *site-truth* index s, which distinguishes between noise and signal sites. Contrast this to Eqn. 9 where each term has a case-truth index t, which distinguishes between normal and abnormal *cases*. The terms μ_s and $\Delta\mu_{is}$ have similar meanings as in the ROC model, except they refer to the means of the noise-site and signal-site distributions, *not* the means of the ROC-equivalent ratings for normal and abnormal cases.

Each site is associated with a physical location and therefore a *location factor*, denoted by L and indexed by ℓs, is needed in the FROC simulation model. Since a site can only occur on a case, ℓs can only occur in the combination *ktℓs*, and the factor L, if it occurs, must occur in combination with C. The converse is not true; one could have a C term without the L term. As noted earlier such terms allow one to incorporate intra-image correlations. The terms in Eqn. 11 involving L are defined as follows: $(\mathrm{CL})_{\mathrm{ktls}} \sim \mathrm{N}(0, \sigma^2_{CL})$ is the random contribution of *<ktls>*, $(\mu\mathrm{CL})_{\mathrm{iktls}} \sim \mathrm{N}(0, \sigma^2_{\mu CL})$ is the random contribution of *<iktls>*, $(\mathrm{RCL})_{\mathrm{jktls}} \sim \mathrm{N}(0, \sigma^2_{RCL})$ is the random contribution of *<jktls>* and $(\mu R\mathrm{CL})_{\mathrm{ijktls}} \sim \mathrm{N}(0, \sigma^2_{\mu RCL})$ is the random contribution of *<ijktls>*. A replication error term, measuring which would require repeated samples from sites with the same set of indices *<ijktls>*, is included in the μRCL term.

Signal-site noise-site correlations on the same (abnormal) image are modeled by modifying the C_{kts}, $(\mu R)_{ijts}$ and $(\mu C)_{ikts}$ terms in Eqn. 11, by including correlation parameters ρ_C, ρ_{RC} and $\rho_{\mu C}$ and regarding C_{kts}, $(\mu R)_{ijts}$ and $(\mu C)_{ikts}$ as *bivariate* samples [6] (on normal cases all z-samples are univariate). For example, for abnormal case *<k2>*, one has $\left(C_{k21}, C_{k22}\right) \sim N_2(0, 0, \sigma^2_C, \sigma^2_C, \rho_C)$. Here $N_2(0, 0, \sigma^2_C, \sigma^2_C, \rho_C)$ is the bivariate normal distribution with zero means and common variance σ^2_C and ρ_C is the correlation. If $\rho_c \sim 1$ the high correlation between C_{k21} and C_{k22} induces positive correlations between the z-samples $z_{ijk2\ell1}$ and $z_{ijk2\ell2}$ realized on the same abnormal image.

2.5 Correlation structures

We adopt the convention that primed and un-primed indices are unequal. The model defined in Eqn. 11 implies eleven (11) correlations between different pairings of ratings. For example, $\rho_{ii'jjllss}$ is the different-modality same-reader same-site same-type correlation between ratings z_{ijktls} and $z_{i'jktls}$. Likewise $\rho_{iijj'll'ss}$ is the same-modality different-reader different-site same-type correlation between ratings z_{ijktls} and $z_{ij'ktl's}$, etc. Since reader variability is usually dominant one expects, for example $\rho_{ii'jjllss} > \rho_{iijj'llss} > \rho_{ii'jj'll'ss}$, analogous to the expected ordering of ROC correlations in the Obuchowski-Rockette model [30]. Expressions for the correlations in terms of the variance components can be derived using the procedure described by Roe and Metz [29]. For example,

$$\rho_{ii'jjllss} = \frac{\sigma_C^2 + \sigma_{RC}^2 + \sigma_{CL}^2 + \sigma_{RCL}^2}{\Sigma_{FROC,C|}^2} \quad . \qquad \text{Eqn. 13}$$

Equations like this were used to define three correlation structures (LOW, MED and HIGH) using a method similar to that described in Ref. [17]. Specific values for all variance components and correlations appearing in Eqn. 11, except σ_R^2 and $\sigma_{\mu R}^2$, were chosen to obtain three sets of 11 correlations termed LOW, MED and HIGH. The 11 correlations arose from three same-site same-type correlations, four different-site same-type correlations, and four different-site different-type correlations. For example, the three same-site same-type correlations were $\rho_{ii'jjllss}$, $\rho_{iijj'llss}$ and $\rho_{ii'jj'll'ss}$. σ_R^2 and $\sigma_{\mu R}^2$ were chosen based on the Roe and Metz variance structure HL (high data correlation, low reader variance [26]), i.e., we set $\sigma_R^2 = \sigma_{\mu R}^2 = 0.0055$.

2.6 Other simulation parameters

Two modalities were assumed and generic "human observers" were simulated with the following characteristics [17]:

$$\left.\begin{array}{llll} \lambda_{1\mathrm{j}} = 1.298 & \lambda_{2\mathrm{j}} = 1.038 & \nu_{1\mathrm{j}} = 0.80 & \nu_{2\mathrm{j}} = 0.88 \\ \mu_1 = 1.50 & \mu_2 = 1.50 & \Delta\mu_{11} = 0 & \Delta\mu_{12} = 0 \\ \Delta\mu_{21} = 0 & \Delta\mu_{22} = 0.04839 & & \end{array}\right\} \quad . \qquad \text{Eqn. 14}$$

These values yielded search-model predicted [25] areas under the ROC curve (AUC) of 0.8 in modality 1 and 0.85 in modality 2; these are AUC values, not areas under the AFROC; the latter are denoted by θ's in this paper. Note that these choices neglect likely reader-dependence of the λ and ν parameters, but the model expressed by Eqn. 11 includes the μ dependence. Eqn. 14 defines the alternative hypothesis (AH) condition. The NH condition was defined by

$$\left.\begin{array}{llll} \lambda_{1\mathrm{j}} = 1.298 & \lambda_{2\mathrm{j}} = \lambda_{1\mathrm{j}} & \nu_{1\mathrm{j}} = 0.80 & \nu_{2\mathrm{j}} = \nu_{1\mathrm{j}} \\ \mu_1 = 1.50 & \mu_2 = 1.50 & \Delta\mu_{11} = 0 & \Delta\mu_{12} = 0 \\ \Delta\mu_{21} = 0 & \Delta\mu_{22} = 0 & & \end{array}\right\} \quad . \qquad \text{Eqn. 15}$$

These values yielded search-model predicted AUC = 0.8 in both modalities. The thresholds adopted by the observers were assumed to be modality and observer independent. The lowest threshold ζ_1 determines the fraction of noise-sites that are marked, $\Phi(-\zeta_1)$, and the values chosen, $-\infty$, -0.674, 0.0, 0.674 correspond to fractions 100%, 75%, 50% and 25%, respectively. Here Φ is the cumulative unit normal distribution function. N_k was randomly sampled from a distribution (see Appendix 1) with parameters $max\left(N_k\right) = 3$ and $avg\left(N_k\right) = 1.3$, specifying the maximum and average numbers of lesions per abnormal image, respectively. The weights W_{kl} were randomly generated by the procedure described in

Appendix 2. The z-samples were binned as described in [31, 32]. NH and AH rejection rates were determined using 2000 simulations for each combination of correlation structure (3 values: LOW, MED and HIGH) and 4 lowest cutoff ζ_1 values: -∞, -0.674, 0.0, 0.674, i.e., 12 combinations in all. These yielded the data in Tables 1 and 2. The case mix was also varied; using the notation N_N/N_A to denote a particular mix, the following combinations were investigated: 100/100, 100/50, 100/25, 50/100, 50/50, 50/25, 25/100, 25/50 and 25/25, yielding the data in Table 3.

3. RESULTS

Simulator-B was used to generate the data reported in Tables 1, 2 and 3. Table 1 compares NH rejection rates for different choices of correlation structures, lowest cutoffs and methods of analyses. For Tables 1 and 2 one hundred normal and one hundred abnormal images were simulated, i.e., $N_N = N_A = 100$. The analyses considered were (i) ROC analysis, (ii) lesion-centric implementations of JAFROC and JAFROC1, and weighted JAFROC and JAFROC1. The shaded section of the table is more relevant to human observers [17] who generate few marks. The next-to-last row is the NH rejection rate averaged over all combinations of ρ and ζ_1 and the last row is the corresponding average over the shaded portion of the table. For all methods, correlation structures and cutoffs NH rejection rates were close to the nominal α-value (5%) of the test.

Table 1: NH rejection rates for ROC analysis, lesion-centric JAFROC and JAFROC1, and weighted JAFROC and JAFROC1. The shaded section of the table is relevant to human observers. NH conditions: AUC = 0.8 for both modalities; $N_N = N_A = 100$.

Lowest cutoff ζ_1	Correlation Structure		Lesion-centric		Weighted	
		ROC	JAFROC	JAFROC1	JAFROC	JAFROC1
-∞	LOW	0.0710	0.0460	0.0530	0.0525	0.0550
	MED	0.0565	0.0360	0.0575	0.0565	0.0575
	HIGH	0.0600	0.0505	0.0510	0.0550	0.0510
-0.674	LOW	0.0525	0.0475	0.0425	0.0435	0.0530
	MED	0.0445	0.0585	0.0520	0.0695	0.0575
	HIGH	0.0535	0.0600	0.0570	0.0570	0.0485
0.0	LOW	0.0590	0.0485	0.0500	0.0480	0.0595
	MED	0.0685	0.0480	0.0570	0.0500	0.0525
	HIGH	0.0615	0.0545	0.0585	0.0505	0.0655
0.674	LOW	0.0490	0.0525	0.0620	0.0640	0.0645
	MED	0.0675	0.0520	0.0575	0.0525	0.0720
	HIGH	0.0730	0.0640	0.0595	0.0420	0.0575
Average (ALL)		0.0597	0.0515	0.0548	0.0534	0.0578
Average (shaded)		0.0631	0.0533	0.0574	0.0512	0.0619

Table 2 compares AH rejection rates (i.e., statistical powers) for the same choices of correlation structures, cutoffs and methods of analyses shown in Table 1. The statistical power ordering of methods using the lesion-centric versions of JAFROC and JAFROC1 (JAFROC1 > JAFROC > ROC) confirmed those reported earlier [17] with simulator-A. For JAFROC and JAFROC1 the statistical powers of the lesion-centric and weighted methods are almost identical and the overall ordering is JAFROC1 > JAFROC > ROC. The power averaged over the shaded area is smaller than that averaged over the entire table since there are more marks (more information) at the lower ζ_1 values.

Table 2: AH rejection rates (statistical power) for ROC analysis, lesion-centric JAFROC and JAFROC1, and weighted JAFROC and JAFROC1. The shaded section of the table is relevant to human observers. AH conditions: AUC = 0.80 for first modality and AUC = 0.85 for second modality; $N_N = N_A = 100$.

Lowest cutoff ζ_1	Correlation Structure		Lesion-centric		Weighted	
		ROC	JAFROC	JAFROC1	JAFROC	JAFROC1
-∞	LOW	0.4815	0.8495	0.9310	0.8640	0.9135
	MED	0.4665	0.8925	0.9335	0.8255	0.8880
	HIGH	0.5905	0.8920	0.9470	0.8770	0.9305
-0.674	LOW	0.3870	0.8170	0.8930	0.8180	0.8875
	MED	0.5285	0.8700	0.9110	0.8270	0.8880
	HIGH	0.6045	0.8630	0.9385	0.8410	0.9115
0.0	LOW	0.4155	0.7655	0.8480	0.7660	0.8265
	MED	0.4595	0.7715	0.8800	0.7665	0.8400
	HIGH	0.5300	0.8435	0.9010	0.8255	0.8770
0.674	LOW	0.3785	0.6115	0.6875	0.5935	0.6550
	MED	0.4265	0.6580	0.6890	0.6415	0.6625
	HIGH	0.4550	0.7125	0.7930	0.6870	0.7575
Average (ALL)		0.4770	0.7955	0.8627	0.7777	0.8365
Average (shaded)		0.4442	0.7271	0.7998	0.7133	0.7698

So far all results reported were for 100 normal and 100 abnormal cases. In Table 3 we report the effects of *case-mix* on NH and AH behaviors of ROC and *weighted* JAFROC and JAFROC1. For each case mix, i.e., a specified pair of values for N_N / N_A, tables like those shown in Tables 1 and 2 were generated. In Table 3 we report only the average values over the shaded region of these tables, as these are most relevant to the human observer. N_T is the total number of cases. NH behavior was within acceptable limits. For all case mixes, the ordering of the methods in terms of statistical power was JAFROC1 > JAFROC > ROC. For all methods, statistical power increased with the total number of cases. For example, JAFROC1, 100 / 100 yielded 77% power compared to 40% power for 25/25. When the total number of cases is the same, the abnormal image richer mix ($N_A > N_N$) gives greater JAFROC1 power, as expected, since JAFROC1 uses the highest rated marks on abnormal images, which are ignored in JAFROC. For example, for JAFROC1, the case mix 100 / 50 yielded 57% power but 50 / 100 yielded 74% power. JAFROC power is affected similarly but to a smaller degree: 54% vs. 59% and ROC power is unaffected: 36% vs. 35%.

Table 3: Effect of case-mix on NH and AH behaviors of ROC and weighted JAFROC and JAFROC1. Data from tables like 1 and 2, averaged over the shaded region is shown for different combinations of numbers of normal cases (N_N) and abnormal cases (N_N); $N_T = N_N + N_A$ is the total number of cases.

N_T	N_N / N_A	NH			AH		
		ROC	JAFROC	JAFROC1	ROC	JAFROC	JAFROC1
200	100 / 100	0.0631	0.0512	0.0619	0.4442	0.7133	0.7698
150	100 / 50	0.0498	0.0538	0.0510	0.3558	0.5376	0.5699
125	100 / 25	0.0522	0.0532	0.0452	0.2468	0.4083	0.4177
150	50 / 100	0.0563	0.0550	0.0608	0.3459	0.5940	0.7380
100	50 / 50	0.0471	0.0552	0.0507	0.2665	0.4686	0.5273
75	50 / 25	0.0527	0.0460	0.0478	0.2195	0.3708	0.3875
125	25 / 100	0.0527	0.0523	0.0566	0.2397	0.4417	0.7368
75	25 / 50	0.0500	0.0510	0.0559	0.2162	0.3379	0.4974
50	25 / 25	0.0517	0.0540	0.0542	0.1937	0.3232	0.3983

4. DISCUSSION

Both JAFROC and JAFROC1 apply to a fully-crossed factorial study design in which all readers interpret all cases in all modalities. For significance testing the images are sequentially jackknifed, i.e., removed from the analysis. Each time the FoM is recomputed and a jackknife pseudovalue is calculated. The pseudovalue matrix, which has dimensions (# modalities) x (# readers) x (# images) is analyzed by a mixed-model ANOVA in a manner completely analogous to the DBM MRMC method [33, 34] for analyzing ROC studies, the only difference being the FoM. The advantage of jackknifing the data is that when an image is removed, all marks associated with the image are removed. No assumptions are made regarding correlations between the marks on the same image. This leads to correct NH behavior over a wide range of inter-modality, intra-image and lesion vs. non-lesion mark correlations as confirmed in prior studies [6, 17] and in this work.

For either JAFROC or JAFROC1 weighted and lesion-centric methods yielded correct NH behavior (weighted and lesion-centric analyses do not apply to the ROC method). This was true over a wide range of numbers of lesions per case, weights, case-mixes, and correlations. The results of the lesion-centric analysis were consistent with those reported earlier using different simulators [6, 17]. The non-parametric FoMs and the resampling-based significance testing may explain the insensitivity of the NH results to the details of the simulators and parameter combinations. Note that the model parameters were assumed to be independent of the weights, which implies that the probability that a lesion is marked is independent of its clinical significance. It is possible that more significant lesions are, on the average, harder to locate than less significant ones, or vise versa, but this was not accounted for in the simulations. The near-equality of statistical powers of lesion-centric and weighted methods could change if the detectability of lesions depends on their weights. However, for reasons stated above, NH validity may not be compromised.

This study confirms the statistical power ordering (JAFROC1 > JAFROC > ROC) reported in Ref. [17] which used simulator-A and the lesion-centric method. The ordering was the same for both lesion-centric and weighted analysis over a wide range of numbers of lesions per case, weights, case-mixes, and correlations. Although JAFROC1 power is greater than JAFROC, especially when the number of normal cases is relatively small, one needs to exercise caution in applying it to datasets with few normal cases, as the results may not generalize to the case population.

Two earlier studies [6, 17] have yielded conflicting results regarding the null hypothesis validity of JAFROC1. In the first study [6] JAFROC1 did not appear to have correct NH behavior. The failures were relatively mild with worst-case rejection rates in the range 0.07–0.08 for a nominal test size of 5%. To be on the conservative side we recommended against using it. The simulator used in that study regarded each image as the union of T non-overlapping regions, each of which could contain noise or a lesion, and T was assumed constant for all images. The more recent study [17], which used the search-model based single-reader two-modality simulator (simulator-A) showed that JAFROC1 had correct NH behavior (i.e., observed NH rejection rates were in the acceptable range, i.e., 0.04 to 0.06 for 2000 simulations) and this was reconfirmed in this study over wide range of simulation conditions. To determine the reason for this inconsistency, we performed an independent re-implementation of the method described in Ref. [6] including the original simulator, and found that JAFROC1 had correct NH behavior. This suggests that the discrepant JAFROC1 result in Ref. [6] was probably due to a programming error.

Simulator-B (multiple-reader multiple-case multiple-modality) is the most versatile known to us at this date. It reduces to the Roe and Metz simulator [26] if the location factor is ignored, and it reduces to the one-reader two-modality simulator (simulator-A) if the reader factor is ignored. Unlike the simulator used in Ref. [6] it accounts for random numbers of NL and LL marks per image. It predicts FROC curves similar to those observed, in particular it does not force the curve to end at $(\infty,1)$; it predicts proper ROC curves, and fits experimentally observed ROC, AFROC and LROC curves [35]. The simulator has a *physical basis* in eye-tracking data obtained while radiologists interpret images. It makes distributional assumptions (Poisson and binomial) which can be relaxed; any integer valued distribution with non-negative samples $n > 0$ can be used for the number of noise sites, and any integer valued distribution with $0 \leq n \leq N_k$ can be used for the number of signal sites. The simulator does not include non-independence effects such as satisfaction of search [36], but it does includes eleven location-specific correlations of the type outlined in Section 2.5. While the simulator can take into account modality and reader dependence of λ and ν, in this study this dependence was ignored; if they were included the powers reported in Tables 2 and 3 would decrease due to the greater variability of FoM induced by the reader and modality dependence of λ and ν, but NH validity is unlikely to be affected. More sophisticated simulators will be needed to properly model the search process in clinical interpretations.

Weighted JAFROC and JAFROC1 solve a limitation of the lesion-centric methods (giving excessive importance to images with many lesions and equal importance to all lesions, regardless of their clinical significances). Unlike the lesion-centric approach weighted JAFROC and JAFROC1 treat each abnormal image (patient) as a unit of measurement, which is desirable based on general statistical considerations. The weighted FoMs imply a different definition for the y-axis of the FROC (and AFROC) curve, but pursuing this aspect of the problem was outside the scope of this work.

Acknowledgments: This work was supported in part by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688.

5. APPENDICES

Appendix 1

N_k was randomly sampled from a binomial-based distribution with parameters $N_L^{\max}$ and N_L^{avg}, specifying the maximum and average numbers of lesions per image, respectively. Defining $p = \dfrac{N_L^{avg} - 1}{N_L^{\max}}$

$N_k \equiv \min(\sim Bin(N_L^{\max}, p) + 1, N_L^{\max})$. Eqn. 16

where ~Bin(N,p) is a sample from the binomial distribution with trial size N and probability of success p. It is seen that

$\langle N_k \rangle = \min(Np + 1, N) = \min(N_L^{avg} - 1 + 1, N) = N_L^{avg}$. Eqn. 17

Eqn. 16 ensures that $\max(N_k) \le N_L^{\max}$.

Appendix 2

For an image with N_k lesions the weights were randomly sampled from a distribution based on the binomial distribution with trial size N_k and probability of success A as follows.

$$
\begin{aligned}
w_{kl'} &= \binom{N_k}{l'} A^{l'} (1 - A)^{N_k - l'} && l' = 0, 1, ..., N_k \\
W_{kl} &= \frac{w_{kl}}{\sum_{l=1}^{N_k} w_{kl}} && l = 1, 2, ..., N_k
\end{aligned}
$$

. Eqn. 18

In this study we set A = 0.5. The weights were randomly sampled for each image.

REFERENCES

[1] 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).

[2] 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 of Appl Photogr. Eng. 4 (4), 166-171 (1978).

[3] Chakraborty, D. P., Breatnach, E. S., Yester, M. V., Soto, B., Barnes, G. T. and Fraser, R. G., "Digital and Conventional Chest Imaging: A Modified ROC Study of Observer Performance Using Simulated Nodules", Radiology 158, 35-39 (1986).

[4] 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", Proc. SPIE 127, 124-135 (1977).

[5] Chakraborty, D. P., "Maximum Likelihood analysis of free-response receiver operating characteristic (FROC) data", Med. Phys. 16 (4), 561-568 (1989).

[6] Chakraborty, D. P. and Berbaum, K. S., "Observer studies involving detection and localization: Modeling, analysis and validation", Medical Physics 31 (8), 2313-2330 (2004).

[7] Penedo, M., Souto, M., Tahoces, P. G., Carreira, J. M., Villalon, J., Porto, G., Seoane, C., Vidal, J. J., Berbaum, K. S., Chakraborty, D. P. and Fajardo, L. L., "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).

[8] Ruschin, M., Timberg, P., Bath, M., Hemdal, B., Svahn, T., Saunders, R., Samei, E., Andersson, I., Mattsson, S., Chakraborty, D. P. and Tingberg, A., "Dose dependence of mass and microcalcification detection in digital mammography: free response human observer studies", Med. Phys. 34, 400 - 407 (2007).

[9] Svahn, T., Hemdal, B., Ruschin, M., Chakraborty, D. P., Andersson, I., Tingberg, A. and Mattson, S., "Dose reduction and its influence on diagnostic accuracy and radiation risk in digital mammography: an observer performance study using an anthropomorphic breast phantom", British Journal of Radiology 80, 557–562 (2007).

[10] 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).

[11] Zanca, F., Chakraborty, D. P., Van Ongeval, C., Jacobs, J., Claus, F., Marchal, G. and Bosmans, H., "An improved method for simulating microcalcifications in digital mammograms", Medical Physics 35 (9), 4012-4018 (2008).

[12] Sahiner, B., Hadjiiski, L. M., Chan, H.-P., Shi, J., Cascade, P. N., Kazerooni, E. A., Zhou, C., Wei, J., Chughtai, A. R., Poopat, C., Song, T., Nojkova, J. S., Frank, L. and Attili, A., "Effect of CAD on radiologists' detection of lung nodules on thoracic CT scans: observer performance study", SPIE Medical Imaging 2007: Perception, Observer Performance, and Technology Assessment 6515 (2007).

[13] McEntee, M. F., Ryan, J., Evanoff, M. G., Keeling, A., Chakraborty, D., Manning, D. and Brennan, P. C., "Ambient lighting: setting international standards for the viewing of softcopy chest images", Proc. SPIE Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment 6515 (2007).

[14] 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), W177-180 (2007).

[15] Volokh, L., Liu, C. and Tsui, B. M. W., "Exploring FROC paradigm - initial experience with clinical applications" Medical Imaging 2006, Image Perception, Observer Performance and Technology Assessment, 6146, (2006).

[16] Timberg, P., Ruschin, M., Båth, M., Hemdal, B., Andersson, I., Mattsson, S., Chakraborty, D., Saunders, R., Samei, E. and Tingberg, A., "Potential for lower absorbed dose in digital mammography: a JAFROC experiment using clinical hybrid images with simulated dose reduction", SPIE Medical Imaging 2006: Perception, Observer Performance, and Technology Assessment 6146 (2006).

[17] Chakraborty, D. P., "Validation and Statistical Power Comparison of Methods for Analyzing Free-response Observer Performance Studies", Academic Radiology 15 (12), 1554-1566 (2008).

[18] 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).

[19] Bornefalk, H., "Estimation and Comparison of CAD System Performance in Clinical Settings", Acad Radiol 12, 687–694 (2005).

[20] Bornefalk, H. and Hermansson, A. B., "On the comparison of FROC curves in mammography CAD systems", Med. Phys. 32 (2), 412-417 (2005).

[21] Popescu, L. M. and Lewitt, R. M., "Small nodule detectability evaluation using a generalized scan statistic model", Phys. Med. Biol. 51 (23), 6225-6244 (2006).

[22] Song, T., Bandos, A. I., Rockette, H. E. and Gur, D., "On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data", Medical Physics 35 (4), 1547-1558 (2008).

[23] Bandos, A. I., Rockette, H. E., Song, T. and Gur, D., "Area under the Free-Response ROC Curve (FROC) and a Related Summary Index", Biometrics EPUB, xx (2008).

[24] Chakraborty, D. P., "ROC Curves predicted by a model of visual search", Phys. Med. Biol. 51, 3463–3482 (2006).

[25] 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).

[26] 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).

[27] Dorfman, D. D., Berbaum, K. S., Lenth, R. V., Chen, Y.-F. and Donaghy, B. A., "Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: Factorial experimental design", Acad. Radiol. 5, 591-602 (1998).

[28] Hillis, S. L. and Berbaum, K. S., "Monte Carlo validation of the dorfman-berbaum-metz method using normalized pseudovalues and less data-based model simplification", Academic Radiology 12 (12), 1534-1541 (2005).

[29] Roe, C. A. and Metz, C. E., "Variance-Component Modeling in the Analysis of Receiver Operating Characteristic Index Estimates", Acad. Radiol. 4 (8), 587-600 (1997).

[30] Obuchowski, N. A. and Rockette, H. E., "Hypothesis Testing of the Diagnostic Accuracy for Multiple Diagnostic Tests: An ANOVA Approach with Dependent Observations", Communications in Statistics: Simulation and Computation 24, 285-308. (1995).

[31] Yoon, H. J., Zheng, B., Sahiner, B. and Chakraborty, D. P., "Evaluating computer-aided detection algorithms", Medical Physics 34 (6), 2024-2038 (2007).

[32] Chakraborty, D. P., "Independent Versus Sequential Reading in ROC Studies of Computer-Assist Modalities: Analysis of Components of Variance", Acad Radiol. 10 (2), 212 (2003).

[33] 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).

[34] Hillis, S. L., Berbaum, K. S. and Metz, C. E., "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis", Acad Radiol 15 (5), 647-661 (2008).

[35] Chakraborty, D. P. and Yoon, H. J., "Operating characteristics predicted by models for diagnostic tasks involving lesion localization", Med. Phys. 35 (2), 435-445 (2008).

[36] Berbaum, K. S., Franken, E. A., Dorfman, D. D., Rooholamini, S. A., Kathol, M. H., Barloon, T. J., Behlke, F. M., Sato, Y., Lu, C. H., El-Khoury, G. Y., Flickinger, F. W. and Montgomery, W. J., "Satisfaction of Search in Diagnostic Radiology", Invest. Radiol. 25 (2), 133-140 (1990).