Received:
31 December 2013

Revised:
21 May 2014

Accepted:
29 May 2014

doi: 10.1259/bjr.20140016

Cite this article as:

Chen W, Samuelson FW. The average receiver operating characteristic curve in multireader multicase imaging studies. Br J Radiol 2014;87: 20140016.

FULL PAPER

The average receiver operating characteristic curve in multireader multicase imaging studies

W CHEN, PhD, F W SAMUELSON, PhD

Division of Imaging and Applied Mathematics, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD, USA

Address correspondence to: Dr Weijie Chen

E-mail: weijie.chen@fda.hhs.gov

Objective: In multireader, multicase (MRMC) receiver operating characteristic (ROC) studies for evaluating medical imaging systems, the area under the ROC curve (AUC) is often used as a summary metric. Owing to the limitations of AUC, plotting the average ROC curve to accompany the rigorous statistical inference on AUC is recommended. The objective of this article is to investigate methods for generating the average ROC curve from ROC curves of individual readers.

Methods: We present both a non-parametric method and a parametric method for averaging ROC curves that produce a ROC curve, the area under which is equal to the average AUC of individual readers (a property we call area preserving). We use hypothetical examples, simulated data and a real-world imaging data set to illustrate these methods and their properties.

Results: We show that our proposed methods are area preserving. We also show that the method of averaging the ROC parameters, either the conventional bi-normal parameters (a, b) or the proper bi-normal parameters (c, d_s), is generally not area preserving and may produce a ROC curve that is intuitively not an average of multiple curves.

Conclusion: Our proposed methods are useful for making plots of average ROC curves in MRMC studies as a companion to the rigorous statistical inference on the AUC end point. The software implementing these methods is freely available from the authors.

Advances in knowledge: Methods for generating the average ROC curve in MRMC ROC studies are formally investigated. The area-preserving criterion we defined is useful to evaluate such methods.

Multireader, multicase (MRMC) receiver operating characteristic (ROC) studies have been widely used to evaluate medical imaging devices and computer-assisted detection/diagnosis devices in medical imaging.¹ Typically, multiple readers (*i.e.* radiologists) read images of multiple cases (*i.e.* patients), and, for each case, each reader rates his/her level of suspicion that a lesion is present based on the image. The diagnostic performance of each reader using an imaging device is thus characterized by a ROC curve constructed using his/her rating data. A ROC curve plots the true-positive fraction (TPF or sensitivity) as a function of the false-positive fraction (FPF or 1-specificity) as the decision threshold varies and thus illustrates the trade-off between the sensitivity and the specificity of a reader across all possible thresholds.²

The area under the ROC curve (AUC) is widely used to summarize the diagnostic performance of imaging systems.³ By analysing the sample AUCs that are obtained from a sample of readers reading a sample of cases, one makes statistical inference on the “population AUC”, which

represents the performance of the device expected (or averaged) over the population of readers and the population of cases.⁴ As such, it characterizes the device performance itself independent of the particular readers and cases used in the study. Methods are well established for the design and analysis of such studies and tutorials,⁵ and consensus papers^{1,6} are available in the literature.

The AUC is a meaningful metric, has many favourable properties and makes rigorous statistical inference possible.³ However, as a summary figure of merit, it loses some of the wealth of information that a ROC curve conveys and thus has limitations.⁷ For example, when ranking imaging systems using ROC curves, one system (A) is surely superior to the other (B) if its ROC curve is higher everywhere because, for each non-trivial sensitivity (or specificity), system A has higher specificity (or sensitivity) than system B. However, if two ROC curves cross between (0, 0) and (1, 1), then some portion of curve A is higher and some portion of curve A is lower than curve B, making the comparison dependent on the location of the operating

point. In this case, use of AUC to rank systems becomes problematic and other metrics, such as partial area under the curve⁸ or sensitivity at a fixed specificity, may be used. A question that naturally arises is why not compare the ROC curves directly instead of a summary figure of merit? One of the reasons is that it is technically much more difficult to make rigorous statistical inference on curves than on scalar variables.

In practice, when a clinical study is designed to use AUC as a study end point for comparing two imaging systems, it is assumed that the “population ROC curves” of the two systems do not cross and thus AUC would provide an unambiguous comparison. At the data analysis stage, it is desirable to plot the average ROC curve as an estimate of the population ROC curve to examine that assumption. Moreover, a summary ROC curve provides a visual representation of the operating characteristics of the readers on an imaging modality and thus provides more information than the summary metric AUC. While a summary ROC curve is not a substitute for the information in all the individual-reader curves, a summary ROC curve is often required and provided in scientific publications as a companion to the rigorous statistical inference results on the AUC in MRMC reader studies.^{9,10}

However, although some methods for generating average ROC curves are suggested,^{7,11} the properties of these methods have not been formally investigated. Section 2.3 in Swets and Pickett¹¹ briefly mentioned that an average ROC curve can be obtained by averaging ROC parameters. Metz⁷ suggested averaging (over readers) parameters of the bi-normal ROC model to obtain the average ROC curve, which was later used in scientific publications.^{9,10} A caveat of this method, as we will show in this article, is that, while the average curve is a companion to the reader-averaged mean AUC, the area under the average curve often does not equal the mean AUC. In this article, we propose curve-averaging methods to overcome this shortcoming. We use hypothetical examples, simulation data and a real-world data set to illustrate the properties of several methods that we considered for generating average ROC curves in MRMC imaging studies.

METHODS AND MATERIALS

Different methods can generate different average ROC curves from the ROC curves of individual readers. We first define an area-preserving criterion for characterizing these methods. We call a method “area preserving” if the area under the average ROC curve is equal to the average of the AUC values of readers. Area preserving would not uniquely define the population ROC curve since there are an infinite number of ROC curves that give the same area. However, we consider it a necessary property for the curve to be acceptable in the following sense. The average ROC curve is constructed in the context that the reader-averaged AUC is the study end point. When we obtain the study end point, that is, the reader-averaged AUC value, we plot the average ROC curve that corresponds to that very average AUC value. Then, we would expect that the area under the average curve be the same as the reader-averaged mean AUC value.

In the following, we consider several methods for generating average ROC curves and investigate their properties. In particular, we investigate if they satisfy the area-preserving criterion. We categorize the ROC curve-averaging methods into (i) non-parametric methods, which do averaging in the ROC space and do not need any parametric model for the averaging; and (ii) parametric methods, which do averaging in the ROC parameter space, that is, the individual ROC curves must be fitted with some parametric model. For non-parametric curve-averaging methods, we propose a general method that is area preserving and can be applied to both empirical (*i.e.* non-parametric) ROC curves and parametric ROC curves. For parametric curve-averaging methods that apply only to parametric ROC curves, we first show that the method of averaging ROC parameters (as suggested by Metz⁷) does not satisfy the area-preserving criterion, and we propose a novel method to overcome this shortcoming.

Non-parametric methods for receiver operating characteristic curve averaging

A straightforward method of averaging multiple ROC curves is to average the sensitivity (S_e) at each specificity (S_p) value or average the S_p at each S_e value.¹² Alternatively, one can average $(S_e + S_p)/2$ at each fixed $(S_e - S_p)/2$. The method is non-parametric as it is a simple arithmetic average and no modelling is needed. It applies to both parametric and non-parametric ROC curves.

To generalize, we formulated a general non-parametric algorithm for ROC curve averaging as follows. First, we rotated the (FPF, TPF) axes in the ROC space counter-clockwise for an angle θ to the (u, v) space:

$$\begin{cases} u = \text{FPF} \cos \theta + \text{TPF} \sin \theta \\ v = -\text{FPF} \sin \theta + \text{TPF} \cos \theta \end{cases}$$

Second, we averaged the ROC curves in the (u, v) space by averaging v for each u . Finally, we rotated the averaged curve in the (u, v) space back to the usual ROC space:

$$\begin{cases} \text{FPF} = u \cos \theta - v \sin \theta \\ \text{TPF} = u \sin \theta + v \cos \theta \end{cases}$$

The parameter θ characterizes the direction along which the ROC curves are averaged.

It can be shown mathematically that this general algorithm is area preserving (Appendix for a proof). In addition, with this general algorithm, the method of averaging the sensitivity at each specificity value is a special case with $\theta = 0$. Similarly, averaging the specificity at each sensitivity value corresponds to $\theta = \pi/2$ and averaging $(S_e + S_p)/2$ at each fixed $(S_e - S_p)/2$ corresponds to $\theta = \pi/4$.

Parametric methods for receiver operating characteristic curve averaging

The empirical ROC curve for each individual reader can be fitted with a parametric model, and the average ROC curve can be constructed based on the fitted ROC parameters. We begin with

Figure 1. Graphs of the dependence of area under the receiver operating characteristic curve [area under the curve (AUC)] on model parameters. (a) The dependence of the AUC of a bi-normal model on parameter a . (b) The dependence of the AUC of a proper bi-normal model on parameter c . Some portions of these curves are well approximated as linear, but, in general, they are non-linear. AUC_{CvB} , area under the conventional receiver operating characteristic curve; AUC_{proper} , area under the proper receiver operating characteristic curve.

a brief review of the two widely used ROC models: the conventional bi-normal model^{13–15} and the proper bi-normal model.¹⁶

In the conventional bi-normal model,^{13–15} the reader's rating data for the non-diseased patients are assumed to be monotonically related to a normal distribution with mean μ_x and variance σ_x^2 . Similarly, the reader's rating data for the diseased patients are assumed to be monotonically related to another normal distribution with mean μ_y and variance σ_y^2 (assuming $\mu_y > \mu_x$ without loss of generality). By defining two parameters $a \equiv (\mu_y - \mu_x)/\sigma_y$ and $b \equiv \sigma_x/\sigma_y$, the ROC curve can be represented by TPF = $\Phi[a + b\Phi^{-1}(\text{FPF})]$ and the area under the curve is $AUC_{CvB} = \Phi(a/\sqrt{1+b^2})$, where Φ is the cumulative distribution function of the standard normal distribution.

In the proper bi-normal model,¹⁶ the underlying decision variable is the likelihood ratio of the two normal distributions. As such, the ROC curve is parameterized by two parameters d_a and c , and the area under the curve is:

$$AUC_{proper} = \Phi\left(\frac{d_a}{\sqrt{2}}\right) + 2F\left(-\frac{d_a}{\sqrt{2}}, 0; -\frac{1-c^2}{1+c^2}\right)$$

where $F(x, y; \rho)$ is the cumulative distribution function of the standard bivariate normal distribution with correlation coefficient ρ .

The brief introduction of the two models here is mainly to highlight the fact that only two parameters are needed to characterize a ROC curve and the AUC under either model. Given a reader's rating data, the ROC parameters, (a, b) or (d_a, c) , can be estimated by maximum-likelihood estimation methods.^{13,14,16}

An approach has been suggested⁷ to generate the average ROC curve in MRMC studies as follows. The observed rating data for each reader are fitted with the bi-normal ROC model that results in two ROC parameters (a, b) . The ROC parameters are averaged over readers to obtain the averaged ROC parameters. Then, the average ROC curve is plotted using the averaged ROC parameters. Similarly, if the ROC data for each reader are fitted with the proper bi-normal model, then the parameters to be averaged are (d_a, c) .

The method may appear to be intuitively sound; however, it is not area preserving in general, and it may produce an average ROC curve that is significantly different from a ROC curve averaged by any other method or lower than most individual ROC curves (Figure 4). We provide a theoretical analysis here and show examples in the Examples section. Suppose there are R readers in a study and we fit their rating data individually using the conventional bi-normal model, which results in a set of fitted ROC parameters $(a_1, b_1, \dots, a_R, b_R)$. These parameters are averaged over readers to obtain $\bar{a} = \sum_{i=1}^R a_i/R$ and $\bar{b} = \sum_{i=1}^R b_i/R$. The average of the AUC values of all the readers is $\overline{AUC} = \sum_{i=1}^R \Phi(a_i/\sqrt{1+b_i^2})/R$, whereas the area under the average ROC curve is $AUC_{avg} = \Phi(\bar{a}/\sqrt{1+\bar{b}^2})$. Mathematically, $\overline{AUC} \neq AUC_{avg}$, that is, not area preserving, because the AUC is a non-linear function of its parameters. It can be similarly shown that the method of averaging the proper ROC parameters (d_a, c) is not area preserving.

To visually appreciate the non-linear nature of the AUC functions, Figure 1a plots the area under the conventional ROC curve (AUC_{CvB}) as a function of the parameter a at a fixed $b = 1$. Figure 1b plots the area under the proper ROC curve (AUC_{proper}) as a function of the parameter c at a fixed $d_a = 1$. As these plots show, the AUC (AUC_{CvB} or AUC_{proper}) is generally a non-linear function of its parameters. However, it can be approximately a linear function in some regions of the parameter space in which case the area-preserving property can be approximately maintained. For example, Figure 1a indicates that, if all the individual ROC curves have a value of 1 for parameter b and $a < 1.5$, then the AUC is approximately a linear function of parameter a and $\overline{AUC} \approx AUC_{avg}$ (which follows from the formulae in the previous paragraph). However, in general, the analytical AUC expressions are non-linear functions and the method generally loses the area-preserving property.

We propose a new parametric method that is area preserving. The basic idea is to obtain one ROC parameter by averaging over readers and compute the other ROC parameter so that the area-preserving criterion is satisfied. For example, for the conventional bi-normal model, we first compute $\overline{AUC}$, the average of the individual-reader AUC values, and $\bar{b}$, the reader-average of

Figure 2. Graphs of three receiver operating characteristic curves (dotted lines), and three different kinds of non-parametric averages of those curves (solid lines). Different types of non-parametric averaging lead to different shapes of average curves, all with the same area. AUC, area under the curve; S_e , sensitivity; S_p , specificity.

parameter b . Then, we compute $\tilde{a} = \sqrt{1 + \bar{b}^2} \Phi^{-1}(\overline{\text{AUC}})$. We use $(\tilde{a}, \bar{b})$ as the conventional bi-normal model parameters to construct the average ROC curve. Mathematically, one can use alternative ways to achieve area preserving, for example, first average parameter a then calculate parameter b to obtain the desired AUC. However, parameter a is very sensitive to the AUC (or the height) of the ROC curve, whereas parameter b is not. Primarily, b influences the shape of the curve, that is, when $b = 1$, the curve is symmetric, when $b > 1$, the curve is skewed to the upper right, and when $b < 1$, the curve is skewed to the lower left. Our choice intends to fix the shape of the curve on average and adjust the height of the curve to preserve area. This idea can be similarly applied to the proper bi-normal model as well and the detail is omitted.

Examples

We present a few hypothetical examples to further illustrate the methods described above. Suppose that we want to average three ROC curves with the following conventional bi-normal parameters: $a_1 = 1, b_1 = 1; a_2 = 2, b_2 = 1; a_3 = 3$ and $b_3 = 1$. They have AUC values of 0.760, 0.921 and 0.983, respectively, with an average AUC value of 0.888. Figure 2 shows these ROC curves (dotted line) and their non-parametric average: averaging sensitivity, averaging specificity and averaging $(S_e + S_p)/2$. They are all area preserving, yielding an AUC value of 0.888 that is the average of the three individual AUC values. The three average curves differ in their shape: the S_p -averaged curve is skewed to the upper right, the S_e -averaged curve is skewed to the lower left and the $(S_e + S_p)/2$ -averaged curve is symmetric. (Note that the original three individual ROC curves are all symmetric.)

Figure 3 shows the average ROC curves obtained using parametric methods. By directly averaging the (a, b) parameters of the three individual ROC curves, we have $\bar{a} = 2$ and $\bar{b} = 1$. The ROC curve corresponding to $(\bar{a}, \bar{b})$, which happens to be

Figure 3. Graphs of three receiver operating characteristic curves (dotted lines), and two different kinds of parametric averages of those curves. Averaging model parameters (bold dotted line) gives a different curve than the proposed method (solid line), which preserves the average area under the curve (AUC).

identical to one of the individual ROC curves, has an AUC value of 0.921, which is greater than the average AUC value of 0.888. Using our proposed method, we first obtain the average AUC value of $\overline{\text{AUC}} = 0.888$ and average b parameter to obtain $\bar{b} = 1$. Then, we compute the a parameter as $\tilde{a} = \sqrt{1 + \bar{b}^2} \Phi^{-1}(\overline{\text{AUC}}) = 1.72$. We plot the average ROC curve using parameters $(\tilde{a}, \bar{b})$ (solid line).

We use another hypothetical example to illustrate the problem of averaging the proper ROC model parameters for curve averaging. Suppose we have six individual ROC curves with the following proper ROC model parameters: $c = \pm 0.2, d_a = 1$, AUC = 0.765; $c = \pm 0.4, d_a = 1$, AUC = 0.813; and $c = \pm 0.6, d_a = 1$, AUC = 0.881. As shown in Figure 4, these six ROC curves (solid line) are three pairs with each pair having the same area corresponding to the same absolute value but opposite signs of c and are skewed toward the opposite sides of the negative 45° line. The average AUC of the six ROC curves is $\overline{\text{AUC}} = 0.820$. By averaging the proper ROC parameters c and d_a , we obtain $\bar{c} = 0$ and $\bar{d}_a = 1$, and the corresponding ROC curve is plotted in Figure 4 (dashed line). The AUC for this ROC curve is $\text{AUC}_{\text{avg}} = 0.760$, which is smaller than the AUC value of any of the six individual curves. The curve is obviously lower than the individual-reader ROC curves. For comparison, the non-parametric S_e -averaged curve is plotted (dot-dash line). It is area preserving and falls in the center of the individual-reader curves (Figure 4).

RESULTS

Simulation

To further demonstrate the methods for averaging ROC curves and investigate their properties, we simulate an application where the task is to compare imaging modality 1 with modality 2. We consider a fully paired reading design in which 15 readers

Figure 4. Graphs of six receiver operating characteristic (ROC) curves (solid lines) and two possible average ROC curves. The dashed line is the ROC curve obtained from averaging model parameters c and d_a . The dash-dotted curve is a non-parametric average, which preserves area. AUC, area under the curve; S_e , sensitivity.

read 300 cases (200 non-diseased cases and 100 diseased cases) for both modalities. Because the purpose here is for demonstrating ROC curve averaging, and we do not intend to study the statistical inference in this article, we can simulate the reader scores as if they are independent. We set a pair of parameters (a, b) for each of the 15 readers and generate reader scores from the following distributions: $N(0, 1)$ for non-diseased cases and $N(a/b, 1/b)$ for diseased cases, where $N(\mu, \sigma)$ denotes normal distribution with mean μ and standard deviation σ . We set 15 sets of (a, b) parameters that were approximated from various real-world imaging studies (for the specific values of these

parameters, see Table 1). The specified (a, b) parameters and the simulated ROC curves for individual readers are based on and similar to several real-data sets, particularly those in Rafferty et al.¹⁷

We fit the 15 individual ROC curves with the proper bi-normal model and applied the non-parametric averaging method from the Non-parametric methods for receiver operating characteristic curve averaging section to both the empirical ROC curves and the fitted curves. Figure 5a,b show the average ROC curves based on empirical ROC curves and that based on fitted ROC curves respectively. Based on the empirical ROC curves, the reader-average AUC (which is equal to the area under the average ROC curve) is 0.899 for modality 1 and 0.861 for modality 2. Based on the fitted ROC curves, the reader-average AUC (which is equal to the area under the average ROC curve) is 0.910 for modality 1 and 0.868 for modality 2. From these results, we observed that the S_e -averaged and the $(S_e + S_p)/2$ -averaged ROC curves are almost identical (the S_p -averaged curve also shows no noticeable difference from these two and is not shown). We also observed that the results are consistent between the empirical ROC curves and the fitted ROC curves.

Figure 6 plots together the $(S_e + S_p)/2$ -averaged curves with the curve obtained by averaging the proper ROC parameters for both modalities. It can be seen that the two methods yield almost identical curves for modality 1 but quite different curves for modality 2. In particular, the method of averaging ROC parameters produced a ROC curve that is substantially lower than the curve produced by the non-parametric method. To make a better visual comparison, we plot these two average curves together with the 15 individual ROC curves for modality 2 (Figure 7). Similar to the example in Figure 4, the method of averaging the ROC parameters produces a ROC curve that is not area preserving and visually is not an appropriate average of the

Figure 5. Non-parametric average receiver operating characteristic (ROC) curves. These curves are averages of 15 simulated ROC curves (the parameters are listed in Table 1). (a) Non-parametric averages of non-parametric ROC curve estimates. (b) Non-parametric averages of parametric ROC curve estimates. In these examples, the different non-parametric averaging methods give very similar results for both simulated imaging modalities. S_e , sensitivity; S_p , specificity.

Figure 6. Receiver operating characteristic (ROC) curves averaged from 15 simulated ROC curves (the parameters are listed in Table 1). Each simulated modality is averaged with non-parametric and parametric averaging [averaging the (c , d_a) parameters]. For modality 1, the parametric and non-parametric averaging give nearly identical results, and for modality 2, they give very different results. S_e , sensitivity; S_p , specificity.

individual-reader curves. The reason for this is again that the c parameters of the individual ROC curves are in a range (from -0.6714 to 0.0563), where the AUC is a non-linear function of parameter c (Figure 1).

Figure 7. Receiver operating characteristic (ROC) curves from simulated imaging modality 2 (grey lines) and averages of those curves (black lines). The non-parametric averaging gives a far more reasonable average curve than does the method of averaging ROC model parameters, which falls well below most of the individual curves. S_e , sensitivity; S_p , specificity.

Table 1. Bi-normal receiver operating characteristic parameters (a , b) used to generate rating scores for 15 simulated readers

Reader	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a	1.1901	1.5582	1.6343	1.0297	0.0000	0.0000	1.2800	0.0000	0.0000	1.1796	1.1653	1.1566	0.0006	1.4245	0.7509
b	0.6712	0.7683	0.8072	0.5101	0.2228	0.2175	0.5415	0.2359	0.2813	0.7073	0.6864	0.4534	0.3163	1.0053	0.4539

Figure 8. Receiver operating characteristic (ROC) curves and averaging for a real-world data set: radiologists read cine MRI images. (a) Non-parametric average (thick solid line) of empirical ROC curves of five radiologists (thin lines). (b) Non-parametric and parametric averages of parametric ROC curve estimates. S_e , sensitivity; S_p , specificity.

Real imaging data

To further validate the area-preserving property of ROC curve-averaging methods considered above, we apply them to a real-world data set. In this clinical example,¹⁸ the task is to evaluate the diagnostic accuracy of the cinematic presentation of spin-echo MRI for the detection of thoracic aortic dissection. 5 radiologists read images from 45 diseased patients and 69 non-diseased patients and gave each patient a rating on a 5-point ordinal scale to rate his/her confidence that the thoracic aortic dissection was present on the image.

Figure 8a shows the empirical ROC curves for the five readers together with the $(S_e + S_p)/2$ -averaged curve. The empirical AUC values for the five readers are 0.920, 0.859, 0.904, 0.973 and 0.830 with an average of 0.897. The AUC for the $(S_e + S_p)/2$ -averaged curve is exactly 0.897, validating the area-preserving property of the method.

Figure 8b shows the ROC curves for the five readers fitted with the conventional bi-normal model¹⁴ together with three average curves. The bi-normal-model-based AUC values for the five readers are 0.933, 0.890, 0.929, 0.971 and 0.833 with an average of 0.911. The first average curve (thick dashed line) is obtained by averaging the fitted a and b parameters, $\bar{a}=1.567$, $\bar{b}=0.479$, and has an AUC value of 0.921, which is greater than the reader-averaged AUC value of 0.911. The second average curve (thick solid line in the middle) is obtained by averaging $(S_e + S_p)/2$ and has exactly the same AUC value as the reader-averaged AUC value of 0.911. The third average curve (thick dotted line) is obtained using our area-preserving parametric method with bi-normal model parameters, $\tilde{a}=1.494$, $\tilde{b}=0.479$, which yields exactly the reader-averaged AUC value of 0.911.

DISCUSSION

We have found that our non-parametric method for ROC curve averaging is area preserving, computationally efficient and robust to different types of ROC curves (empirical or parametric). The reliability of the method is demonstrated in hypothetical examples, simulated data and real-world data sets.

There is a free parameter in our non-parametric method (namely the angle parameter θ), the choice of which can produce ROC

curves that are noticeably different, as we have seen in an example shown in Figure 2. However, this parameter did not produce noticeable difference in the average ROC curve for the simulation data in the Results section. In general, the parameter more likely makes a difference when the shape of the individual ROC curves is more similar. For example, the three individual ROC curves in the hypothetical example in the Examples section are all symmetric with respect to the diagonal. According to our experience, the angle parameter usually does not make much difference in real-world applications. In practical applications, to be conservative, one may vary the parameter to examine if two average ROC curves cross in worst-case scenarios.

An alternative approach to obtain an estimate of the population ROC curve is to pool all the readers' rating data together to generate one ROC curve.^{7,11} This approach assumes that all the readers use the rating scale consistently. When the assumption is violated, the pooled ROC curve is generally lower than the area-preserving average curve⁷ (also see page 133 of Pepe¹⁵).

CONCLUSION

We argue that the average ROC curve as a companion to the reader-averaged AUC in MRMC reader studies should have an AUC equalling the reader-averaged AUC (a property we call "area preserving"). We showed that the method of averaging ROC parameters, either the conventional bi-normal parameters (a , b) or the proper bi-normal parameters (c , d_a), is generally not area preserving. We proposed both a non-parametric and a parametric method for averaging ROC curves that are area preserving. Our methods are useful to make plots of average ROC curves in MRMC studies for evaluating imaging systems as a companion to the rigorous statistical inference on the AUC end point. The software implementing these methods is freely available from the authors.

ACKNOWLEDGMENTS

Certain commercial materials and equipment are identified to specify experimental procedures adequately. In no case does such identification imply recommendation or endorsement by the Federal Drugs Agency, nor does it imply that the items identified are necessarily the best available for the purpose.

REFERENCES

- Gallas BD, Chan HP, D'Orsi CJ, Dodd LE, Giger ML, Gur D, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad Radiol* 2012; **19**: 463–77. doi: 10.1016/j.acra.2011.12.016
- Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978; **8**: 283–98.
- Samuelson FW. Inference based on diagnostic measures from studies of new imaging devices. *Acad Radiol* 2013; **20**: 816–24. doi: 10.1016/j.acra.2013.03.002
- Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol* 1992; **27**: 723–31.
- Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Acad Radiol* 2007; **14**: 723–48. doi: 10.1016/j.acra.2007.03.001
- Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: lessons from recent experience. *Acad Radiol* 2002; **9**: 1264–77.
- Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* 1989; **24**: 234–45.
- Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996; **201**: 745–50. doi: 10.1148/radiology.201.3.8939225
- Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. *Radiology* 1999; **212**: 817–27.
- Chan HP, Helvie MA, Petrick N, Sahiner B, Adler DD, Paramagul C, et al. Digital mammography: observer performance study of the effects of pixel size on the characterization of malignant and benign microcalcifications. *Acad Radiol* 2001; **8**: 454–66.
- Swets JA, Pickett RM. *Evaluation of diagnostic systems: methods from signal detection theory*. New York, NY: Academic Press; 1982.
- Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006; **27**: 861–74.
- Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. *J Math Psychol* 1969; **6**: 487–96.
- Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med* 1998; **17**: 1033–53.
- Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford University Press; 2003.
- Metz CE, Pan X. "Proper" binormal ROC curves: theory and maximum-likelihood estimation. *J Math Psychol* 1999; **43**: 1–33.
- Rafferty EA, Park JM, Philpotts LE, Poplack SP, Sumkin JH, Halpern EF, et al. Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial. *Radiology* 2012; **266**: 104–13.
- Van Dyke CW, White RD, Obuchowski NA, Geisinger MA, Lorig RJ, Meziene MA. *Cine MRI in the diagnosis of thoracic aortic dissection*. 79th RSNA Meetings; Chicago, IL, 1993.

APPENDIX

Proof of area-preserving for our receiver operating characteristic curve average algorithm

First, we show that for any receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC) in the [false-positive fraction (FPF), true-positive fraction (TPF)] space ($AUC^{(TF)}$) is the same as that in the (u, v) space ($AUC^{(uv)}$). This is because the two spaces are related by a rotation:

$$\begin{cases} u = \text{FPF} \cos \theta + \text{TPF} \sin \theta \\ v = -\text{FPF} \sin \theta + \text{TPF} \cos \theta \end{cases}$$

and the Jacobian matrix is:

$$J = \begin{bmatrix} \frac{\partial u}{\partial \text{FPF}} & \frac{\partial u}{\partial \text{TPF}} \\ \frac{\partial v}{\partial \text{FPF}} & \frac{\partial v}{\partial \text{TPF}} \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

with determinant $|J|=1$. Therefore, $AUC^{(uv)} = \iint dudv = \iint |J| d\text{TPF} d\text{FPF} = \iint d\text{TPF} d\text{FPF} = AUC^{(TF)}$. Second, we show that, in the (u, v) space where the average takes place, the average is area preserving. Suppose there are R ROC curves for averaging, the AUC for each curve is $AUC_i^{(uv)} = \int u_i dv$. The AUC for the averaged curve is $AUC_{\text{avg}}^{(uv)} = \int \bar{u} dv$, where $\bar{u} = \sum_{i=1}^R u_i / R$ according to our algorithm. Then, the average of the R individual AUC values is $\overline{AUC}^{(uv)} = \sum_{i=1}^R AUC_i^{(uv)} / R = \sum_{i=1}^R \int u_i dv / R = \int \bar{u} dv = AUC_{\text{avg}}^{(uv)}$. Combining the two points above, the AUC of the averaged ROC curve is the same as the averaged AUC values for any rotation, that is, $\overline{AUC}^{(TF)} = AUC_{\text{avg}}^{(TF)}$.