Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Bounding sample size projections for the area under a ROC curve

Jeffrey D. Blume

Center for Statistical Sciences, Brown University, Providence, RI 02912, USA

ARTICLE INFO

Article history:
Received 9 October 2006
Accepted 28 September 2007

Keywords:
Receiver operating characteristic (ROC) curve
Multireader study
Area under the curve (AUC)
Bounds

ABSTRACT

Studies of diagnostic tests are often designed with the goal of estimating the area under the receiver operating characteristic curve (AUC) because the AUC is a natural summary of a test's overall diagnostic ability. However, sample size projections dealing with AUCs are very sensitive to assumptions about the variance of the empirical AUC estimator, which depends on two correlation parameters. While these correlation parameters can be estimated from the available data, in practice it is hard to find reliable estimates before the study is conducted. Here we derive achievable bounds on the projected sample size that are free of these two correlation parameters. The lower bound is the smallest sample size that would yield the desired level of precision *for some model*, while the upper bound is the smallest sample size that would yield the desired level of precision *for all models*. These bounds are important reference points when designing a single or multi-arm study; they are the absolute minimum and maximum sample size that would ever be required. When the study design includes multiple readers or interpreters of the test, we derive bounds pertaining to the average reader AUC and the 'pooled' or overall AUC for the population of readers. These upper bounds for multireader studies are not too conservative when several readers are involved.

1. Introduction

Receiver operating characteristic (ROC) curves assess the ability of a diagnostic test to discriminate between two populations. See Pepe (2003) and Zhou et al. (2002) for a comprehensive description of methods for diagnostic tests. The area under the ROC curve (AUC) is the probability that a randomly selected observation from one population scores less on the test than a randomly selected observation from the other population. If the AUC is one (or zero) then the test discriminates perfectly, but if the AUC is one-half then the test has no discriminative ability whatsoever. The AUC is an easily understood metric that summarizes the test's overall diagnostic ability. Studies that seek to assess the ability of a diagnostic test are typically designed to estimate the AUC within a fixed margin of error (Hanley and McNeil, 1982, 1983, 1984; Obuchowski, 1994, 1998; Owen et al., 1964; Janes and Pepe, 2006).

The AUC estimator of choice is the empirical (scaled Mann–Whitney–U) estimator because it is a consistent and efficient estimator under very general conditions (Hanley and McNeil, 1982, 1984; Lehmann, 1998). Its variance depends on the true AUC and two additional correlation parameters, both of which can be estimated non-parametrically from the data at hand (Mee, 1990; Hanley and McNeil, 1982; Bamber, 1975). However, at the design stage, there is seldom any data available to estimate these parameters. Investigators may have an idea of what the AUC might be, but they seldom have a reliable guess for the correlation parameters. Unfortunately these correlation parameters play an important role in determining the variance, making the subsequent sample size projections quite sensitive to the initial inputs.

E-mail address: jblume@stat.brown.edu.

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.09.015

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

However, it is possible to determine the maximum and minimum sample size required to achieve a certain level of efficiency, over all models, without having to specify the correlation parameters. To derive these bounds, we use the fact that the maximum and minimum variance of the empirical AUC estimate, over all models, depends only on the true AUC and the sample size. Specification of the two key correlation parameters is, in this sense, eliminated. The minimum sample size will yield the desired level of precision *for some model*, while the maximum will yield the desired level of precision *for all models*. These bounds are useful for determining the sensitivity of sample size projections to modeling assumptions and for determining the extent of resources that may be needed for a given study.

We then extend these bounds to complex designs involving multiple tests and multiple readers of each test. These situations require specification of the correlation structure between the tests and readers. We examine the bounds in relation to projections obtained according to the standard ANOVA approach for multireader studies. This approach is based on a random effects ANOVA model for each reader's empirical AUC, which does account for correlation between readers. These methods apply, in principle, to any large sample situation under certain conditions (see Thompson and Zucchini, 1989; Obuchowski, 1995; Roe and Metz, 1997). However, the ANOVA approach is highly sensitive to assumptions about the components of variance and may yield projections greater than the upper bound. We illustrate this behavior and discuss when these bounds are themselves sufficient projections.

Section 2 is background and establishes notation. Sample size projections for estimating a single AUC are in Section 4 and projections for estimating the difference between two AUCs are in Section 5. Extensions to designs involving multiple readers are also considered. Closing comments are in Section 7.

2. Background

A sample of test outcomes from the 'non-diseased' or normal population will be represented by random variables $X_1, X_2, \ldots, X_m$ with cumulative distribution function F. Likewise, represent the test outcomes from the 'diseased' or abnormal population as $Y_1, Y_2, \ldots, Y_n$ with CDF G. Denote the AUC as $\theta = P(X < Y)$ (Bamber, 1975). The empirical estimator of the AUC, $\widehat{\theta}$, is

$$\widehat{\theta} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij} \tag{1}$$

where $Z_{ij} = I(X_i < Y_j) + \frac{1}{2}I(X_i = Y_j)$ and $I(\cdot)$ is an indicator function. If CDFs F and G are continuous then $E[\widehat{\theta}] = \theta$ and

$$Var[\widehat{\theta}] = \frac{\theta(1-\theta)}{mn}(1 + (n-1)\rho_y + (m-1)\rho_x) \tag{2}$$

where $\rho_y = [P(X < \min(Y_i, Y_j)) - \theta^2]/\theta(1-\theta)$ and $\rho_x = [P(\max(X_i, X_j) < Y) - \theta^2]/\theta(1-\theta)$ are the correlation coefficients between two indicators variables based on the same y (or x) observation. The major contributors to the variance are ρ_y/m and ρ_x/n; the first term is $O(1/mn)$. Both ρ_y and ρ_x take different forms depending on the underlying model, but both can be estimated non-parametrically if data are available (Mee, 1990; Hanley and McNeil, 1982; Bamber, 1975).

Ties reduce the variance of $\widehat{\theta}$ (Putter, 1955). Assume that F and G have common discontinuities (the only ones that allow ties of importance), finite in number and denoted by ξ_k, $k = 1, \ldots, K$. Define $p_k = P(X_i = \xi_k)$ and $q_k = P(Y_j = \xi_k)$ then $E[\widehat{\theta}] = \theta^* = P(X < Y) + \frac{1}{2}P(X = Y)$ and the variance of $\widehat{\theta}$ is now

$$Var[\widehat{\theta}] = \frac{\theta^*(1-\theta^*)}{mn}(1 + (n-1)\rho_y^* + (m-1)\rho_x^*) - \frac{1}{12mn}\sum_k p_k q_k[(n-1)q_k + (m-1)p_k + 3] \tag{3}$$

where ρ_y^* and ρ_x^* are now the corresponding correlation coefficients. The correction term is often negligible, but can, under certain circumstances, be of the same order of magnitude as the initial term in the variance (Lehmann, 1998, p. 20).

In the sections that follow, we will use the fact that the empirical AUC estimator, $\widehat{\theta}$, and difference between two such estimators for different tests, say $\widehat{\theta}_A - \widehat{\theta}_B$, are both approximately normally distributed in large samples (Lehmann, 1998, 1999). Hence, an approximate $(1-\alpha)100\%$ confidence interval for θ is

$$\widehat{\theta} \pm Z_{\alpha/2}\sqrt{V(\widehat{\theta})} \tag{4}$$

where $V(\widehat{\theta})$ is given by (2). Likewise a hypothesis test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1 > \theta_0$ based on the test statistic $(\widehat{\theta} - \theta_0)/\sqrt{V_0(\widehat{\theta})}$ has Type II error probability:

$$\beta = \Phi\left[Z_{\alpha/2}\frac{\sqrt{V_0(\widehat{\theta})}}{\sqrt{V_1(\widehat{\theta})}} + \frac{\theta_0 - \theta_1}{\sqrt{V_1(\widehat{\theta})}}\right] \tag{5}$$

Here we ignored the lower tail of the test, which contributes very little in any case. Eqs. (4) and (5) readily generalize when the object of inference is $\theta_A - \theta_B$ and they are often used for sample size projections with the understanding that mis-specification of ρ_y and ρ_x can be costly (Delong et al., 1988; Hanley and Hajian-Tilaki, 1997).

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

Although ρ_y and ρ_x are often intractable, there are two important exceptions where an analytical form for $V(\widehat{\theta})$ can be obtained (Basu, 1981; Hajian-Tilaki et al., 1997). The first is the 'biexponential' model where $X_i \sim \text{Exponential}(\lambda_X)$ and $Y_j \sim \text{Exponential}(\lambda_Y)$. Direct calculation yields $P(\max(X_i, X_j) < Y) = 2\theta^2/(1+\theta)$ and $P(X < \min(Y_i, Y_j)) = \theta/(2-\theta)$ with correlation parameters $\rho_x = \theta/(1+\theta)$ and $\rho_y = (1-\theta)/(2-\theta)$ (Basu, 1981; Hanley and McNeil, 1982). Because the correlation parameters depend on the AUC in a simple fashion, this model is routinely used for sample size calculations (Hanley and McNeil, 1982; Obuchowski, 1995). However, this can (and often does) lead to sample sizes that are too small for many applications (Obuchowski, 1994; Walsh, 1997).

The second noteworthy case is an equal variance 'binormal' model where a Taylor series expansion of the AUC variance is free of correlation parameters. The binormal assumption is that some monotonic transformation, say $h(\cdot)$, can be found such that $h(X_i) \sim \text{N}(\mu_x, \sigma_x^2)$ and $h(Y_j) \sim \text{N}(\mu_y, \sigma_y^2)$ (Metz, 1978, 1986; Egan, 1975; Swets et al., 1961). Without loss of generality let $\mu_x < \mu_y$. The area under that ROC curve is $P(X < Y) = P(h(X) < h(Y)) = \theta_{\text{nor}} = \Phi[(\mu_y - \mu_x)/\sqrt{\sigma_x^2 + \sigma_y^2}]$ where $\Phi[\cdot]$ is the standard normal CDF. θ_{nor} is a monotone increasing function of $\delta = (\mu_y - \mu_x)/\sqrt{\sigma_x^2 + \sigma_y^2}$ and inference on δ leads to inference on θ by a simple transformation (Reiser and Guttman, 1986; Enis and Geisser, 1971; Owen et al., 1964). A Taylor series argument yields the following approximation for the standard error when $\sigma_x = \sigma_y$:

$$Var[\widehat{\theta}_{\text{nor}}] \approx (0.0099 \times \exp\{-A^2/2\}) \times \left(\frac{5A^2 + 8}{n} + \frac{A^2 + 8}{m} \right) \tag{6}$$

where $A = 1.414 \times \Phi^{-1}[\theta]$ (Obuchowski, 1994, 1998). See also Obuchowski and McClish (1997). We will use these two cases for comparison against the sample size bounds.

3. Bounding the variance of the empirical AUC estimator

When F and G are both continuous, Birnbaum and Klose (1957) show that

$$\frac{\theta(1-\theta)}{mn} B_i(m, n, \theta) \leqslant Var[\widehat{\theta}] \leqslant \frac{\theta(1-\theta)}{mn} \max(m, n) = \frac{\theta(1-\theta)}{\min(m, n)} \tag{7}$$

where $i = 1$ if $(\min(m, n) - 1)/(\max(m, n) - 1) \leqslant 2\min(\theta, 1-\theta)$, $i = 2$ otherwise, and

$$B_1(m, n, \theta) = \left[\min(m, n) - \frac{(\min(m, n) - 1)^2}{12(\max(m, n) - 1)\theta(1-\theta)} \right] \tag{8}$$

$$B_2(m, n, \theta) = \left[1 - (m + n - 2)\frac{\min(\theta, 1-\theta)}{\max(\theta, 1-\theta)} + \frac{4}{3\max(\theta, 1-\theta)} \sqrt{2\min(\theta, 1-\theta)(m-1)(n-1)} \right] \tag{9}$$

These bounds are achievable when the underlying distributions are continuous. The confidence interval, $\widehat{\theta} \pm Z_{\alpha/2}\sqrt{\theta(1-\theta)/\min(m, n)}$, is robust in that it maintains at least $(1-\alpha)100\%$ coverage probability under model failure (see Birnbaum, 1956). We will make use of the fact that (7) implies

$$B_i(m, n, \theta) \leqslant (1 + (n-1)\rho_y + (m-1)\rho_x) \leqslant \max(m, n) \tag{10}$$

where $i = 1, 2$ depending on $(\min(m, n) - 1)/(\max(m, n) - 1)$ as noted above. The middle term of (10) is the inflation factor that corrects the variance for the dependence between indicator variables.

When ties are present, Eq. (3) indicates that the upper bound on the variance still holds. In fact, Lemma 5.1 of Lehmann (1951) shows that these bounds extend immediately by simply subtracting $(1/12mn)\sum_k p_k q_k[(n-1)q_k + (m-1)p_k + 3]$ from each of the upper and lower bounds and replacing θ with θ^*. Note that this correction term, which requires specification of the probability mass on points of common discontinuities, is often quite small. Hence ignoring ties, which we do for the remainder of this paper, is at worse conservative.

Now consider two distinct diagnostic tests, A and B. For test 'A', let $X_{A1}, X_{A2}, \ldots, X_{Am}$ and $Y_{A1}, Y_{A2}, \ldots, Y_{An}$ have CDFs F_A and G_A, respectively. The AUC is $\theta_A = P(X_A < Y_A)$ and the variance of $\widehat{\theta}_A$ depends on ρ_{Ay} and ρ_{Ax}, the cluster correlations under modality 'A'. Let a similar notational structure exist for test 'B'. The difference in AUCs, $\theta_A - \theta_B$, is of interest. The variance of the estimated difference is

$$Var[\widehat{\theta}_A - \widehat{\theta}_B] = Var[\widehat{\theta}_A] + Var[\widehat{\theta}_B] - 2\rho\sqrt{Var[\widehat{\theta}_A]Var[\widehat{\theta}_B]} \tag{11}$$

where $\rho = \text{Corr}[\widehat{\theta}_A, \widehat{\theta}_B]$. If different participants are evaluated on each modality then $\rho = 0$. However, we gain efficiency by evaluating the same cases under both modalities because ρ is now positive. Hanley and McNeil (1983) provide a useful discussion and tabulation of ρ. Assuming ρ is fixed and nonnegative, Eq. (11) can be bounded by substituting the upper and lower bounds

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

for the variance as follows:

$$Var[\widehat{\theta}_A - \widehat{\theta}_B] < \frac{\theta_A(1-\theta_A)+\theta_B(1-\theta_B)}{\min(m,n)} - \frac{2\rho\sqrt{\theta_A(1-\theta_A)B_i(m,n,\theta_A)\theta_B(1-\theta_B)B_i(m,n,\theta_B)}}{mn} \tag{12}$$

$$Var[\widehat{\theta}_A - \widehat{\theta}_B] > \frac{\theta_A(1-\theta_A)B_i(m,n,\theta_A)+\theta_B(1-\theta_B)B_i(m,n,\theta_B)}{mn} - \frac{2\rho\sqrt{\theta_A(1-\theta_A)\theta_B(1-\theta_B)}}{\min(m,n)} \tag{13}$$

where $i=1$ if $(\min(m,n)-1)/(\max(m,n)-1) \leqslant 2\min(\theta,1-\theta)$, $i=2$ otherwise. Define the lower bound to be zero whenever the RHS of (13) is negative. These bounds are achievable only when $\rho=0$. Nevertheless, they are useful reference points.

If the cluster correlations are the same across modalities (i.e., $\rho_{Ax}=\rho_{Bx}=\rho_x$ and $\rho_{Ay}=\rho_{By}=\rho_y$), then a set of tighter, achievable bounds exist. Factoring out the now common over-dispersion term $(1+(n-1)\rho_y+(m-1)\rho_x)$ in Eq. (11) and bounding by (10) yields

$$Var[\widehat{\theta}_A - \widehat{\theta}_B] \leqslant \frac{1}{\min(m,n)}\left[\theta_A(1-\theta_A)+\theta_B(1-\theta_B)-2\rho\sqrt{\theta_A(1-\theta_A)\theta_B(1-\theta_B)}\right] \tag{14}$$

$$Var[\widehat{\theta}_A - \widehat{\theta}_B] \geqslant \frac{\max(B_i(m,n,\theta_A),B_i(m,n,\theta_B))}{mn} \times \left[\theta_A(1-\theta_A)+\theta_B(1-\theta_B)-2\rho\sqrt{\theta_A(1-\theta_A)\theta_B(1-\theta_B)}\right] \tag{15}$$

The maximum of the B_i's is required in the first term of the RHS of Eq. (15) because (10) must hold for both modalities. These bounds are achievable, but hold only when the cluster correlations are equal across modalities. Upper bounds (12) and (14) are equal when $\rho=0$.

4. Sample size projections for a single AUC

To bound the projected sample size, we substitute the maximum and minimum variance from (7) for $V(\widehat{\theta})$ in (4) and (5) and solve. The idea is to find the smallest sample size that will meet the statistical planning criteria in every model, call it $n_{\max}$, and the smallest sample size that will meet the statistical criteria for some model, call it $n_{\min}$. Any projected sample size must then be less than or equal to $n_{\max}$ and greater than or equal to $n_{\min}$. These limits reveal the range of potential sample sizes, which we can use to address the projection's sensitivity to violations of assumptions.

Substituting the maximum variance for $V(\widehat{\theta})$ yields the smallest sample size that achieves a $(1-\alpha)100\%$ CI of at most length L in every model. This upper bound for projected sample sizes is

$$n_{\max} = \min(m,n) = 4(Z_{\alpha/2})^2\frac{\theta(1-\theta)}{L^2} \tag{16}$$

The smaller of the normal and abnormal groups must be at least $n_{\max}$ to guarantee that a $(1-\alpha)100\%$ CI will have length L or less. There is no model under which $n_{\max}$ would result in a $(1-\alpha)100\%$ CI with length greater than L.

Substituting the minimum variance for $V(\widehat{\theta})$ yields the smallest sample size that could possibly achieve a $(1-\alpha)100\%$ CI of length L. The lower bound for projected sample sizes is the solution to

$$\frac{nm}{B_i(m,n,\theta)} = 4(Z_{\alpha/2})^2\frac{\theta(1-\theta)}{L^2} \tag{17}$$

where $i=1,2$ depending on $(\min(m,n)-1)/(\max(m,n)-1)$ as noted earlier. This equation must be solved numerically for $n_{\min}$. A useful device is to fix the ratio of group sizes, say k, so that $m=kn$ and then solve for n. Here $k=m/n$ is the ratio of normals to abnormals; typically $k>1$. The solution, $n_{\min}$, is the smallest sample size that could possibly yield a $(1-\alpha)100\%$ CI of length L. There is some model for which $n_{\min}$ would result in a $(1-\alpha)100\%$ CI with length L.

The same logic applies for power calculations: just substitute the maximum and minimum variance from (7) for $V(\widehat{\theta})$ in (5). Numerical solutions are again needed to solve for $n_{\min}$, but an analytical expression exists for $n_{\max}$:

$$n_{\max} = \min(m,n) = \frac{(Z_\beta\sqrt{\theta_1(1-\theta_1)}+Z_{\alpha/2}\sqrt{\theta_0(1-\theta_0)})^2}{(\theta_1-\theta_0)^2} \tag{18}$$

The smaller of the normal and abnormal groups must be at least $n_{\max}$ to guarantee at least $(1-\beta)$ power to detect θ_1 under any model. With $n_{\min}$, there is at least one model under which there is $(1-\beta)$ power to detect θ_1. Simply put, at least $n_{\min}$ observations will be required to achieve $(1-\beta)$ power, but up to $n_{\max}$ observations may be needed.

Example 1a. An investigator guesses that the AUC for a new diagnostic assay will be 85% and he would like to estimate that AUC within a margin of error of 5% (i.e., $L=0.1$) from a 95% CI. Equal numbers of disease and non-diseased participants will be recruited. The maximum projection is 196 participants in both groups and the lower projection is at least 101 participants per group. (Here, $i=2$ and $B_2(101,101,0.85)=51.62$ so that the LHS of (17) is 197.61, which is the closest to 195.91 = RHS without going below). Under the equal variance binormal model we project 151 participants per group and under the biexponential

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

model we need 117 per group. The biexponential model appears quite optimistic. The maximum projections require 30% more participants than the binormal and 68% more than the biexponential.

Example 1b. An investigator wants to test the hypothesis that the accuracy of a new diagnostic assay is 0.8. She believes recent improvements will increase the accuracy to 0.9 and she wants to have 80% power to detect this difference with 5% Type I error. If the ratio of normals to abnormals is 1:1 (i.e., $k = 1$), then she requires at least 58 abnormal participants, but no more than 108. The binormal model suggests 81 abnormal and the biexponential model suggests 66. But if two normals are enrolled for every abnormal (i.e., $k = 2$), then the binormal model suggests only 66 abnormal and the biexponential only 57. Now the projection bounds indicate that at least 39 abnormals are required, but no more than 108 are needed. Remember that the maximum bound depends only on the smaller of the two groups. (Here, $i = 2$ for all cases. When $k = 1$, $B_2(58, 58, 0.8) = 22.06$, $\sqrt{V(0.8)} = 0.0393$, $B_2(58, 58, 0.9) = 17.73$, $\sqrt{V(0.9)} = 0.0264$; when $k = 2$, $B_2(78, 39, 0.8) = 29.27$, $B_2(78, 39, 0.9) = 24.06$, $\sqrt{V(0.8)} = 0.0392$ and $\sqrt{V(0.9)} = 0.0267$. Substituting these into (5) gives $1 - \beta \approx 0.807$ in both cases.)

4.1. Multiple readers and their average AUC

Sometimes the outcome of a diagnostic test must be interpreted by a 'reader'. For example, X-rays and MRI scans require the interpretation of a reader. But two readers may score the same test result, in this case an image, differently. This disagreement ultimately results in different ROC curves for each reader and introduces a source of variability that is not due to the diagnostic test itself. (This framework also applies to multicenter trials when the outcome of an 'objective' diagnostic test for the same case may vary as it is evaluated at each site.) The primary interest in a multireader design such as this is no longer each reader's AUC, but instead the average reader AUC.

A sample of test outcomes for the rth reader ($r = 1, \ldots, R$ readers) will be represented as $X_{r1}, X_{r2}, \ldots, X_{rm}$ and $Y_{r1}, Y_{r2}, \ldots, Y_{rn}$ with CDFs F_r and G_r. Each reader's AUC is $\theta_r = P(X_r < Y_r)$. Its empirical estimate is $\widehat{\theta}_r = (1/mn)\sum_{i=1}^{m}\sum_{j=1}^{n} Z_{rij}$ where $Z_{rij} = I(X_{ri} < Y_{rj})$, $i = 1, \ldots, m, j = 1, \ldots, n$ and the variance of $\widehat{\theta}_r$ depends on θ_r, ρ_{ry} and ρ_{rx}.

In the design stage, it is not unreasonable to assume that readers are exchangeable in the sense that they have the same true accuracy and cluster correlation parameters. That is, $\theta_r = \theta$, $\rho_{ry} = \rho_y$ and $\rho_{rx} = \rho_x$ for all r. Then, the obvious empirical estimate of θ is $\widehat{\theta}_\cdot = \frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r$ and $Var[\widehat{\theta}_r]$ does not depend on r by assumption. Then

$$Var[\widehat{\theta}_\cdot] = \frac{1}{R^2} Cov\left[\sum_{r=1}^{R}\widehat{\theta}_r, \sum_{s=1}^{R}\widehat{\theta}_s\right] = Var[\widehat{\theta}]\left[\frac{1}{R} + \frac{R-1}{R}\rho_{\mathrm{dr,sm}}\right] \tag{19}$$

where $\rho_{\mathrm{dr,sm}}$ is the correlation between *d*ifferent *r*eaders on the *s*ame *m*odality (typically $\rho_{\mathrm{dr,sm}} > 0$). Note that $Var[\widehat{\theta}]$ is the term we have been bounding and estimating under different models.

Eq. (19) shows that including multiple readers deflates the variance by $[1/R + \rho_{\mathrm{dr,sm}}(R-1)/R]$. Therefore, accounting for multiple readers amounts to nothing more than deflating the projected sample sizes (and bounds) by the same factor. That is, the new upper bound is $n'_{\max} = n_{\max} \times [1/R + \rho_{\mathrm{dr,sm}}(R-1)/R]$ and the new lower bound is $n'_{\min} = n_{\min} \times [1/R + \rho_{\mathrm{dr,sm}}(R-1)/R]$. Likewise for projections under the binormal and biexponential models. For example, with four readers and $\rho_{\mathrm{dr,sm}} = 0.3$, the projections are reduced by a factor of 0.475. In Example 1a, this means the maximum reduces to 94, the lower to 48, the binormal to 72, and biexponential to 56. The same applies for the power projections in Example 1b.

4.2. Efficiency considerations

The loss in efficiency incurred when using the maximum sample size projection depends on the true underlying model. The maximum potential increase in sample size, over all models, is $(n_{\max}/n_{\min} - 1)100\%$. In Example 1a, the potential increase in sample size is 94.1%. However, this maximum is only achieved if the true underlying model happens to yield the smallest possible variance for the empirical AUC estimator. This is unlikely. A more realistic approach is to compare the maximum sample size to that obtained under the two most commonly assumed models: the biexponential and binormal. In Example 1a, we saw that the increase in sample size was 68% and 30%.

To illustrate, we assume $k = m/n \geqslant 1$, which is most common in practice and compare the sample sizes required to provide a confidence interval of length L. The ratios of projected sample sizes for the maximum versus binormal and maximum versus biexponential are

$$\frac{n_{\max}}{n_{\mathrm{nor}}} = \frac{\theta(1-\theta)}{(0.0099 \times \exp\{-A^2/2\}) \times \left(5A^2 + 8 + \dfrac{A^2+8}{k}\right)} \tag{20}$$

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

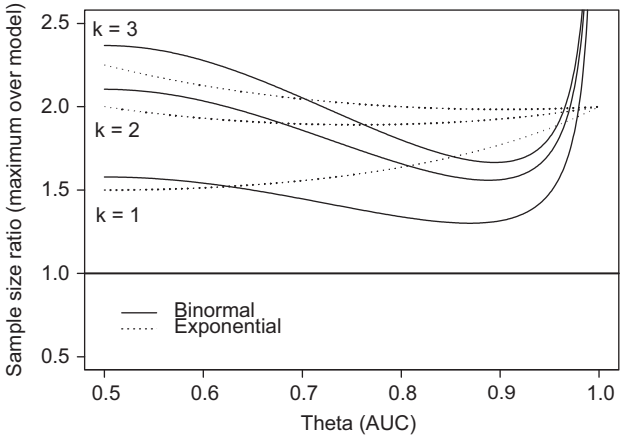

Fig. 1. Sample Size ratios.

Table 1
Comparison of multireader standard errors on $\widehat{\theta}_A$

$\rho_{dr,sm} = 0.33$			Multireader Standard errors of $\widehat{\theta}_A$			
θ_A	Raters	(m, n)	Lower bound	Exponential	Binormal	Upper bound
0.70	3	(50, 100)	0.0306	0.0320	0.0340	0.0482
		(100, 100)	0.0269	0.0274	0.0283	0.0341
		(200, 300)	0.0169	0.0172	0.0181	0.0241
0.70	6	(50, 100)	0.0274	0.0286	0.0304	0.0431
		(100, 100)	0.0240	0.0245	0.0253	0.0305
		(200, 300)	0.0151	0.0153	0.0162	0.0215
0.70	9	(50, 100)	0.0262	0.0273	0.0291	0.0412
		(100, 100)	0.0230	0.0234	0.0242	0.0291
		(200, 300)	0.0144	0.0147	0.0155	0.0206
0.80	3	(50, 100)	0.0257	0.0264	0.0303	0.0421
		(100, 100)	0.0222	0.0233	0.0257	0.0298
		(200, 300)	0.0140	0.0143	0.0162	0.0210
0.80	6	(50, 100)	0.0230	0.0236	0.0270	0.0376
		(100, 100)	0.0199	0.0208	0.0230	0.0266
		(200, 300)	0.0125	0.0128	0.0145	0.0188
0.80	9	(50, 100)	0.0220	0.0225	0.0259	0.0360
		(100, 100)	0.0190	0.0199	0.0220	0.0254
		(200, 300)	0.0120	0.0123	0.0139	0.0180
0.90	3	(50, 100)	0.0175	0.0182	0.0223	0.0316
		(100, 100)	0.0149	0.0168	0.0195	0.0223
		(200, 300)	0.0094	0.0101	0.0121	0.0158
0.90	6	(50, 100)	0.0156	0.0162	0.0199	0.0282
		(100, 100)	0.0133	0.0150	0.0174	0.0199
		(200, 300)	0.0084	0.0090	0.0108	0.0141
0.90	9	(50, 100)	0.0149	0.0155	0.0191	0.0270
		(100, 100)	0.0127	0.0144	0.0166	0.0191
		(200, 300)	0.0081	0.0086	0.0103	0.0135

and

$$\frac{n_{max}}{n_{exp}} = \frac{k}{\frac{1-\theta}{2-\theta} + k\frac{\theta}{1+\theta}} \tag{21}$$

For the biexponential model we use an approximation to the variance to obtain a closed form expression for the sample size ratio. Namely $Var[\widehat{\theta}] \approx \theta(1-\theta)(\rho_y + k\rho_x)/nk$, which just assumes that $1/kn^2$ is zero.

Fig. 1 displays these ratios as a function of the area under the curve, $\theta \in [0.5, 1]$, and $k \in \{1, 2, 3\}$. For large values of the AUC the sample size ratio for the binormal model approaches infinity. This is an artifact of the Taylor series approximation (Obuchowski and McClish, 1997); under both models the variance of the AUC estimator converges to zero as θ approaches 1, but the binormal

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

convergence is (exponentially) faster. Hence the appearance of greatly increased efficiency at extremely high AUCs. The figure indicates that if the underlying model is exponential then the maximum sample size represents an increase ranging from 50% to 100% when $k = 1$. Under the equal variance binormal model the smallest sample size ratio is approximately 1.3. It is known that the biexponential model tends to give sample sizes that are too small (Obuchowski, 1994; Walsh, 1997) and Fig. 1 indicates that those sample size projections can be quite optimistic (i.e., half of the maximum possible). However, using a binormal model that assumes equal variance may not be much better.

While the inclusion of multiple readers does not change these considerations, the resulting standard errors will be much smaller and consideration of their absolute magnitude is very important. Table 1 provides some examples of standard errors for comparison. The absolute loss in efficiency is quite small and the difference can be mediated by adding readers.

5. Sample size projections for comparing two AUCs

The goal of many studies is to compare the accuracy of two tests by comparing their AUCs. The approach for this case is similar to that already outlined concerning a single AUC; we substitute for $Var[\widehat{\theta}_A - \widehat{\theta}_B]$ its bounds in the two-sample versions of Eqs. (4) and (5) to derive the bounds on the projected sample size. The only caveat is that there are two sets of bounds to choose from: the strict (and sometimes unachievable) bounds or the (achievable) bounds that assume equal cluster correlations across modalities. We assume throughout the same ratio of abnormal to normals for each modality (e.g., when both tests are run on each subject or when both tests drawn from the same global population).

Most of these calculations must be done numerically, substituting the bounds for the actual variance. However, if we are willing to assume equal cluster correlations, the upper bound on the projected sample size for a CI is

$$n_{\max} = \min(m, n) = 4\frac{(Z_{\alpha/2})^2}{L^2}[\theta_A(1-\theta_A) + \theta_B(1-\theta_B) - 2\rho\sqrt{\theta_A(1-\theta_A)\theta_B(1-\theta_B)}] \tag{22}$$

Thus, a minimum of $n_{\max}$ observations in each group will guarantee a $(1-\alpha)100\%$ CI of length L as long as cluster correlations are equal across modalities. Likewise the sample size that provides $1-\beta$ power for a α-sized hypothesis test of $H_0 : \Delta_0 = \theta_{A0} - \theta_{B0}$ versus $H_1 : \Delta_1 = \theta_{A1} - \theta_{B1}$ is given by the solution to

$$\beta = \Phi\left[Z_{\alpha/2}\frac{\sqrt{V_0}}{\sqrt{V_1}} + \frac{\Delta_0 - \Delta_1}{\sqrt{V_1}}\right] \tag{23}$$

where $V_i = Var[\widehat{\theta}_{Ai} - \widehat{\theta}_{Bi}]$. The upper bound on the sample size projection is the solution to this equation when V_i is replaced with (12) (or (14) if we are willing to assume equal correlations), while the lower bound is the solution when V_i is replaced with (13) (or (15) if we are willing to assume equal correlations). The binormal or biexponential models can also be used calculate V_i via Eqs. (11) and (2).

Example 2a. Two assays are to be compared. Equal numbers of normal and abnormal participants are available (i.e., $k = 1$) and each participant will undergo both assays (say with $\rho = 0.3$). An investigator guesses that the AUCs for assays A and B are 0.7 and 0.8 and would like a 95% CI on their difference to have a margin of error of 0.075 (i.e., length of 0.15). The maximum sample size is 209 abnormals or 178 if equal cluster correlations can be assumed. The minimum sample size is 76 abnormals or 111 if equal cluster correlations can be assumed. Projections under the biexponential and binormal model give 113 and 127 abnormals, respectively. Compared with the bounds these projections appear optimistic.

Example 2b. Same situation as Example 2a, but the investigator wishes to conduct a hypothesis test against the null hypothesis that the two assays have equal accuracy of 0.7. With 80% power and 5% type I error, the maximum sample size is 260 abnormals or 223 if equal cluster correlations can be assumed. The minimum sample size is 101 abnormals or 139 if equal cluster correlations can be assumed. Projections under the biexponential and binormal model give 143 and 156, respectively. Here as well the binormal and biexponential models appear optimistic.

A practical approach to balancing efficiency and robustness concerns might be to choose the sample size given from the lower upper bound (i.e., assume equal cluster correlations) or to select the midpoint of the sample size range (143 in Example 2a and 181 in Example 2b).

5.1. Multiple readers and the average difference in AUCs

The arguments of Section 4.1 apply here directly. The empirical difference in average reader AUCs is $\widehat{\theta}_{A\cdot} - \widehat{\theta}_{B\cdot} = (1/R)\sum_{r=1}^{R}(\widehat{\theta}_{Ar} - \widehat{\theta}_{Br}) = (1/R)\sum_{r=1}^{R}\widehat{\theta}_{dr}$. Note the $Var[\widehat{\theta}_{dr}]$ does not depend on r because of our common accuracy and common cluster correlation

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

Table 2
Comparison of multireader standard errors of $\widehat{\theta}_A - \widehat{\theta}_B$ with 5 readers and $\sigma_b^2 = 0.00104$, $\sigma_w^2 = 0.00012$

$\rho_1 = 0.44$, $\rho_2 = 0.33$, $\rho_3 = 0.29$				Multireader standard errors of $\widehat{\theta}_A - \widehat{\theta}_B$ for 5 readers				
θ_A	θ_B	(m, n)	$\rho_{dr,diff}$	Lower bound	LB[a]	ANOVA	UB[a]	Upper bound
0.60	0.50	(50, 100)	0.072	0.0000	0.0243	0.0284	0.0376	0.0454
		(100, 100)	0.072	0.0168	0.0217	0.0242	0.0266	0.0299
		(400, 500)	0.072	0.0067	0.0102	0.0148	0.0133	0.0153
0.60	0.55	(50, 100)	0.072	0.0000	0.0242	0.0282	0.0375	0.0453
		(100, 100)	0.072	0.0167	0.0216	0.0241	0.0265	0.0298
		(400, 500)	0.072	0.0067	0.0101	0.0148	0.0133	0.0153
0.75	0.70	(50, 100)	0.073	0.0000	0.0216	0.0246	0.0339	0.0412
		(100, 100)	0.072	0.0132	0.0189	0.0219	0.0240	0.0274
		(400, 500)	0.073	0.0050	0.0089	0.0139	0.0120	0.0140
0.85	0.80	(50, 100)	0.077	0.0000	0.0178	0.0207	0.0291	0.0356
		(100, 100)	0.075	0.0087	0.0153	0.0191	0.0205	0.0239
		(400, 500)	0.076	0.0027	0.0072	0.0130	0.0103	0.0122
0.90	0.80	(50, 100)	0.100	0.0000	0.0174	0.0195	0.0284	0.0346
		(100, 100)	0.094	0.0081	0.0149	0.0182	0.0199	0.0233
		(400, 500)	0.096	0.0025	0.0070	0.0127	0.0100	0.0119
0.90	0.85	(50, 100)	0.082	0.0000	0.0151	0.0183	0.0256	0.0315
		(100, 100)	0.080	0.0054	0.0129	0.0173	0.0180	0.0213
		(400, 500)	0.081	0.0000	0.0061	0.0124	0.0090	0.0108

[a]Bound assumes equal correlation across modalities.

assumption. Then we have

$$\begin{aligned} Var[\widehat{\theta}_{A\cdot} - \widehat{\theta}_{B\cdot}] &= \frac{1}{R^2} Cov\left[\sum_{r=1}^{R} \widehat{\theta}_{dr}, \sum_{s=1}^{R} \widehat{\theta}_{ds}\right] \\ &= \frac{1}{R}\left[Var[\widehat{\theta}_{dr}] + (R-1)\mathrm{Corr}[\widehat{\theta}_{dr}, \widehat{\theta}_{ds}]\sqrt{Var[\widehat{\theta}_{dr}]Var[\widehat{\theta}_{ds}]}\right] \\ &= Var[\widehat{\theta}_A - \widehat{\theta}_B]\left[\frac{1}{R} + \frac{R-1}{R}\rho_{dr,diff}\right]. \end{aligned} \tag{24}$$

Here $\rho_{dr,diff}$ represents the correlation between *d*ifferent *r*eaders on the *diff*erence in AUC estimates. That is, it is the correlation of the empirical AUC differences among the readers. Also, the quantity ρ from Section 5 will now be denoted by $\rho_{sr,dm}$ (the correlation between the *s*ame *r*eader on *d*ifferent *m*odalities). Just like before, accounting for multiple readers amounts to nothing more than deflating the projected sample sizes and bounds. That is, the new upper bound is $n'_{max} = n_{max} \times [1/R + \rho_{dr,diff}(R-1)/R]$ and the new lower bound is $n'_{min} = n_{min} \times [1/R + \rho_{dr,diff}(R-1)/R]$. Likewise for projections under the binormal and biexponential models.

5.2. Relationship to other multireader projections

Multireader designs typically specify three different correlation parameters: ρ_1 ($\rho_{sr,dm}$ in our notation), the correlation between AUCs estimated by the same reader but in two different modalities; ρ_2 ($\rho_{dr,sm}$ in our notation), the correlation between AUCs estimated by different readers in the same modality; and ρ_3 ($\rho_{dr,dm}$ in our notation), is the correlation between AUCs estimated by different readers in different modalities (Obuchowski, 1995; Roe and Metz, 1997). So far, we have only encountered $\rho_{sr,dm}$ and $\rho_{dr,diff}$, but all are related. Expanding $Cov[\widehat{\theta}_{dr}, \widehat{\theta}_{ds}]$ shows

$$\rho_{dr,diff} = \frac{\left[Var(\widehat{\theta}_{Ar}) + Var(\widehat{\theta}_{Br})\right]\rho_{dr,sm} - 2\rho_{dr,dm}\sqrt{Var(\widehat{\theta}_{Ar})Var(\widehat{\theta}_{Br})}}{Var(\widehat{\theta}_{Ar}) + Var(\widehat{\theta}_{Br}) - 2\rho_{sr,dm}\sqrt{Var(\widehat{\theta}_{Ar})Var(\widehat{\theta}_{Br})}} \tag{25}$$

This relationship depends on both group sample sizes, (m, n), through $Var(\widehat{\theta})$, although this expression is $O_p(1)$. When the true variances are in doubt (as is often the case) this relationship cannot be calculated. However, the biexponential model can be used to calculate the variances for comparison purposes.

Traditional approaches to multireader studies are based on a random effects ANOVA model that works with the reader specific AUC estimates (see Thompson and Zucchini, 1989; Obuchowski, 1995; Roe and Metz, 1997). While these ANOVA models may apply in principle to any large sample situation, they fail to incorporate distributional assumptions that enter through the variance.

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

Not unexpectedly, these projections are highly sensitive to assumptions about the components of variance. Obuchowski (1995) shows that the ANOVA variance for the difference between two AUC estimates is

$$Var(\widehat{\theta}_{A\cdot} - \widehat{\theta}_{B\cdot}) = \frac{2}{R}\left[\sigma_b^2(1-\rho_b) + \frac{\sigma_w^2}{J} + \sigma_c^2[1-\rho_1+(R-1)(\rho_2-\rho_3)]\right] \tag{26}$$

where σ_b^2 is the between reader variability; σ_w^2 is the within reader variability; σ_c^2 is variability due to cases; ρ_b is the correlation between AUCs estimated by a set of readers under two different modalities; the correlation parameters ρ_1, ρ_2, ρ_3, discussed earlier; and J is the number of replications (almost always $J=1$). Technically, this ANOVA variance is unconstrained and can produce standard errors much larger (or smaller) than even the absolute upper (or lower) bound established earlier. However, one vetted set of estimates from the radiology literature is: $\sigma_b^2 = 0.00104$, $\sigma_w^2 = 0.0001$, $\rho_b = 0.82$, $\rho_1 = 0.44$, $\rho_2 = 0.33$, $\rho_3 = 0.29$ (Obuchowski, 1995). The last component of variance σ_c^2 is estimated as $(Var[\widehat{\theta}_A] + Var[\widehat{\theta}_B])/2$ assuming a biexponential model (Obuchowski, 1995).

For illustrative purposes, Table 2 displays multireader standard errors under the ANOVA model with the components of variance identified above and our bounds, under a variety of conditions and sample sizes with five readers. The ANOVA standard errors under this set of variance components fair well for moderate sample sizes. They are close to the sharp lower bound for smaller sample sizes and in some cases exceed the upper bounds (see Table 2). Of course, small changes in the assumptions about the variance components (e.g., say $\sigma_w^2 = 0.001$ or $\sigma_b^2 = 0.0001$) often cause relatively large changes in the ANOVA standard errors. While these results are not generalizable, they most likely represent the best case scenario for the ANOVA method (because we have chosen vetted estimates of the variance components from the literature, etc.). The point is that even if the variance components and correlations are chosen wisely, the standard errors could still be too large or too small.

6. Considerations for pooled AUCs

It might be of interest to assess the accuracy of the entire population of readers, taken as a single unit. In this case, we are interested in how the group of readers performs as a whole and so the focus is on the AUC that results from pooling data across all readers. This is interesting when there is little variability from reader to reader, e.g., when the 'readers' are actually an objective test, such as a blood test performed at different centers.

A sample of test outcomes for the rth reader ($r = 1, \ldots, R$ readers), under modality A, will be represented as $X_{Ar1}, X_{Ar2}, \ldots, X_{Arm}$ and $Y_{Ar1}, Y_{Ar2}, \ldots, Y_{Arn}$ with CDFs F_{Ar} and G_{Ar}, respectively, and AUC $\theta_{Ar} = P(X_{Ar} < Y_{Ar})$. An additional level is added to account for the population of readers. For outcomes from the 'non-diseased' population let $F_{Ar}(X) = F_A(X; \alpha_r)$ be the reader specific CDF where α_r has CDF $H_A(\alpha_r)$. Then the individual scores, X_A, have CDF $F_A(X) = \int_{-\infty}^{X} \int f_A(z; \alpha_r)\, dH_A(\alpha_r)\, dz$. The key insight here is that this model considers all of the X's as being generated from a single population of 'non-diseased' scores comprised from all readers. The one caveat is that this model assumes that all readers operate on the same scoring scale, while the analysis of average reader AUCs in the previous section does not.

For the 'diseased' population, we have Y_{Ar} with CDF $G_{Ar}(Y) = G_A(Y; \beta_r)$ where β_r has CDF $W_A(\beta_r)$. Because $H_A(\alpha_r)$ and $G_{Ar}(Y) = G_A(Y; \beta_r)$ are not necessarily equal, reader variability may differ depending on the population being scored, even though the readers remain constant. The resulting convolution is $G_A(Y) = \int_{-\infty}^{Y} \int g_A(z; \beta_r)\, dW_A(\beta_r)\, dz$. The AUC for modality A is $\theta_A = P(X_{Ak} < Y_{Al})$ (without reader subscript). As expected, the population AUC is a weighted average of the reader specific AUCs, but the weights do not add to one (see the Appendix for details and an example).

The pooled AUC is estimated by $\widehat{\theta}_A = \sum_{k=1}^{Rm}\sum_{l=1}^{Rn} Z_{Akl}/R^2mn$ where $Z_{Akl} = I(X_{Ak} < Y_{Al})$ makes no distinction between readers. The variance of $\widehat{\theta}_A$ follows from Eq. (2). Now the bounds on sample size projections for comparing two AUCs (see Section 5) apply here by simply replacing m and n with mR and nR, respectively. This is easily adapted to designs that are not fully factorial by replacing mR by $\sum_r m_i$ and nR by $\sum_r n_i$, where r indexes the readers and m_i and n_i represent the number of non-diseased and diseased cases evaluated by rater i.

7. Comments

This paper outlines a method for bounding the projected sample sizes in studies that seek to estimate or compare the area under ROC curves. These bounds provide an important benchmark for assessing the sensitivity of projections to underlying assumptions. The potential efficiency loss from designing a study with the maximum sample size projection must be balanced against the potential for not achieving the desired statistical objective. In multireader studies, this efficiency loss can be offset by adding readers. In other situations, the midpoint of the sample size range may represent an acceptable balance between efficiency and robustness. Small studies or studies collecting an unequal number of disease and non-diseased cases could substantially benefit from making some sort of distributional assumption, if that assumption is indeed correct. However, in multireader studies, the absolute efficiency difference is often quite small (because the standard errors themselves are already very small), so there appears to be little advantage to making these distributional assumptions. Overall, these bounds provide a nice complement to any model-based approach.

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

Acknowledgements

This work was partially funded by ACRIN NCI U01-CA79778. The author wishes to thank Constantine Gatsonis and Benjamin Herman for their comments.

Appendix

The population AUC, θ, that is derived from pooling data across readers is a weighted average of the reader specific AUCs, $\theta_r(\alpha_r, \beta_r)$.

$$\begin{aligned}\theta &= \int_S G(S)f(S)\,\mathrm{d}S \\ &= \int_S \int_{-\infty}^{S} \int_{\beta_r} g(Y_S;\beta_r)w(\beta_r)\,\mathrm{d}\beta_r\,\mathrm{d}Y_S \int_{\alpha_r} f(S;\alpha_r)h(\alpha_r)\,\mathrm{d}\alpha_r\,\mathrm{d}S \\ &= \int_{\alpha_r} h(\alpha_r) \int_S \int_{-\infty}^{S} \int_{\beta_r} g(Y_S;\beta_r)w(\beta_r)\,\mathrm{d}\beta_r\,\mathrm{d}Y_S f(S;\alpha_r)\,\mathrm{d}S\,\mathrm{d}\alpha_r \\ &= \int_{\alpha_r}\int_{\beta_r} h(\alpha_r)w(\beta_r) \int_S \int_{-\infty}^{S} g(Y_S;\beta_r)\,\mathrm{d}Y_S f(S;\alpha_r)\,\mathrm{d}S\,\mathrm{d}\alpha_r\,\mathrm{d}\beta_r \\ &= \int_{\alpha_r}\int_{\beta_r} h(\alpha_r)w(\beta_r)\theta_r(\alpha_r,\beta_r)\,\mathrm{d}\alpha_r\,\mathrm{d}\beta_r\end{aligned}$$

While $\int_{\alpha_r} h(\alpha_r)\,\mathrm{d}\alpha_r = \int_{\beta_r} w(\beta_r)\,\mathrm{d}\beta_r = 1$, in general, $\int_{\alpha_r}\int_{\beta_r} h(\alpha_r)w(\beta_r)\,\mathrm{d}\alpha_r\,\mathrm{d}\beta_r \neq 1$. Hence θ is a weighted average of the reader specific AUCs, but the weights do not necessarily add to one. This allows θ to take values beyond the range of the reader specific AUCs. A similar situation is encountered in Sukhatme and Beam (1994).

For an example of this model structure, consider the rth reader under a Binormal model. We have that $X_{ri}|\mu_{xr} \sim \mathrm{N}(\mu_{xr}, \sigma_x^2)$ and $Y_{rj}|\mu_{yr} \sim \mathrm{N}(\mu_{yr}, \sigma_y^2)$. Variability in the population of readers is characterized by $\mu_{xr} \sim \mathrm{N}(\mu_x, \tau_x^2)$ and $\mu_{yr} \sim \mathrm{N}(\mu_y, \tau_y^2)$, resulting in a model that implies $X_k \sim \mathrm{N}(\mu_x, \sigma_x^2 + \tau_x^2)$ where $k = 1, \ldots, Rm$ and $Y_l \sim \mathrm{N}(\mu_y, \sigma_y^2 + \tau_y^2)$ where $l = 1, \ldots, Rn$. The accuracy for this group of readers is

$$P(X_k < Y_l) = \Phi\left[\frac{\mu_y - \mu_x}{\sqrt{(\sigma_y^2 + \tau_y^2) + (\sigma_x^2 + \tau_x^2)}}\right]$$

Thus the accuracy of the group may be significantly less than any individual reader depending on the magnitude of $(\tau_x^2 + \tau_y^2)$.

References

Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating graph. J. Math. Psychol. 12, 387–415.
Basu, A.P., 1981. The estimation of $P(X < Y)$ for distributions useful in life testing. Naval Res. Logist. Quart. 28 (3), 383–392.
Birnbaum, Z.W., 1956. On a use of the Mann–Whitney statistic. in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, CA. pp. 13–17.
Birnbaum, Z.W., Klose, O.M., 1957. Bounds for the variance of the Mann–Whitney statistic. Ann. Math. Stat. 28 (4), 933–945.
Delong, E.R., Delong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.
Egan, J.P., 1975. Signal Detection Theory and ROC Analysis. Academic Press, New York.
Enis, P., Geisser, S., 1971. Estimation of the probability that $X < Y$. J. Amer. Stat. Assoc. 66, 162–168.
Hajian-Tilaki, K.O., Hanley, J.A., Joseph, L., Collet, J., 1997. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. Med. Decision Making 17, 94–102.
Hanley, J.A., Hajian-Tilaki, K.O., 1997. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. Acad. Radiol. 4, 49–58.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.
Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843.
Hanley, J.A., McNeil, B.J., 1984. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med. Decision Making 4 (2), 137–150.
Janes, H., Pepe, M., 2006. The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker. Biostatistics 7 (3), 465–468.
Lehmann, E.L., 1951. Consistency and unbiasedness of certain nonparametric tests. Ann. Math. Statist. 22, 165–179.
Lehmann, E.L., 1998. Nonparametrics: Statistical Methods Based on Ranks. Prentice-Hall, Englkewood Cliffs, NJ.
Lehmann, E.L., 1999. Elements of Large-Sample Theory. Springer, New York.
Mee, R.W., 1990. Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann–Whitney statistic. J. Amer. Statist. Assoc. 85, 793–800.
Metz, C.E., 1978. Basic principles of ROC analysis. Sem. Nuclear Med. VIII (4), 283–298.
Metz, C.E., 1986. ROC Methodology in radiologic imaging. Invest. Radiol. 21, 720–733.
Obuchowski, N.A., 1994. Computing sample size for receiver operating characteristic studies. Investigative Radiol. 29 (2), 238–243.
Obuchowski, N.A., 1995. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. Acad. Radiol. 2, s22–s29.
Obuchowski, N.A., 1998. Sample size calculations in studies of test accuracy. Statist. Methods in Med. Res. 7, 371–392.

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015

Obuchowski, N.A., McClish, D.K., 1997. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Statist. Med. 16, 1529–1542.
Owen, D.B., Cradwell, K.J., Hanson, D.L., 1964. Sample size calculations in studies of test accuracy. J. Amer. Statist. Assoc. 59, 906–924.
Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York.
Putter, J., 1955. The treatment of ties in some nonparametric tests. Ann. of Math. Statist. 26 (3), 368–386.
Reiser, B., Guttman, I., 1986. Statistical inference for $P(X < Y)$: the normal case. Technometrics 28 (3), 253–257.
Roe, C.A., Metz, C.E., 1997. Variance-component modeling in the analysis of receiver operating characteristic index estimates. Acad. Radiol. 4, 587–600.
Sukhatme, S., Beam, C., 1994. Stratification in nonparametric ROC studies. Biometrics 50, 149–163.
Swets, J.A., Tanner Jr., W.P., Birdsall, T.G., 1961. Decision processes in perception. Psychol? Rev. 16, 669–679.
Thompson, M.L., Zucchini, W., 1989. On the statistical analysis of ROC curves. Statist. Med. 8, 1277–1290.
Walsh, S.J., 1997. Limitations to the robustness of binormal ROC curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power. Statist. Med. 16, 669–679.
Zhou, X.H., Obuchowski, N.A., McClish, D., 2002. Statistical Methods in Diagnostic Medicine. Wiley, New York.

Please cite this article as: Blume, J.D., Bounding sample size projections for the area under a ROC curve. J. Statist. Plann. Inference (2008), doi: 10.1016/j.jspi.2007.09.015