Revisiting sample size planning for receiver operating characteristic studies: A confidence interval approach with precision and assurance

Di Shu1,2,3 and Guangyong Zou4,5

Abstract
Estimation of areas under receiver operating characteristic curves and their differences is a key task in diagnostic studies. Here we develop closed-form sample size formulas for such studies with a focus on estimation rather than hypothesis testing, by explicitly incorporating pre-specified precision and assurance, with precision denoted by the lower limit of confidence interval and assurance denoted by the probability of achieving that lower limit. For sample size estimation purposes, we introduce a normality-based variance function for valid estimation allowing for unequal variances of observations in the disease and non-disease groups. Simulation results demonstrate that the proposed formulas produce empirical assurance probability close to the pre-specified assurance probability and empirical coverage probability close to the nominal level. Compared with a frequently used existing variance function, the proposed function provides more accurate and efficient sample size estimates. For an illustration of the proposed formulas, we present real-world worked examples. To facilitate implementation, we have developed an online calculator openly available at https://dishu.page/calculator/.

Keywords
Area under receiver operating characteristic curve, assurance probability, confidence interval, diagnostic study, sample size estimation

1 Introduction
Diagnostic tools (commonly referred to as tests) frequently generate data on an ordinal or continuous scale to detect the true disease status. In what to follow, we regard a higher observation as implying stronger evidence of a participant being diseased. Situations where smaller observations suggest stronger evidence can be handled analogously. The area under receiver operating characteristic curve (AUC) is commonly used to quantify the overall accuracy of a diagnostic tool.1–3 The receiver operating characteristic (ROC) curve is a plot of sensitivity against 1-specificity obtained by varying the cutoff point above which a participant will be considered test positive (diseased). The AUC is one useful summary measure of the entire ROC curve because it represents the probability of correctly identifying the participant condition from a random pair of diseased and non-diseased participants.1,4 The values of AUC typically range from 0.5 (no apparent ability) to 1 (perfect accuracy).
One common approach to the comparison of two diagnostic tools is through inference on the difference between their AUCs.5

As for any other studies, sample size estimation is usually the first task in the planning of a study evaluating the overall diagnostic ability of a marker, focusing on the estimation of single AUCs and/or their differences.6 Inadequate and overly planned sample sizes both can considerably waste research resources; too small or too big a study might even be considered unethical.7 Previous sample size formulas and software packages were primarily developed for achieving either a desired power in hypothesis testing, a desired expected confidence interval (CI) width, or a desired acceptance rule in an acceptance-sampling-theory framework.1,5,8–18 Here we focus on a different criterion, whether the estimates are of acceptable precision, as quantified by the lower confidence limit.19–21 Note that hypothesis testing focuses on a single value (the null hypothesis) while CI estimation focuses on a range of plausible values and thus is more informative than the former approach.22 Based on benchmarks23,24 such as an AUC between 0.9 and 1.0 reflecting “excellent” accuracy, 0.8 and 0.9 “good” accuracy, 0.7 and 0.8 “fair” accuracy, and below 0.6 “poor” accuracy, and below 0.6 “failed”, one might ask “how many participants should be recruited to assure with 90% chance that the lower confidence limit is above 0.8 (so that I feel comfortable to call the test good)’’?

The purpose of this report is to derive, evaluate, and implement simple sample size formulas for planning ROC studies in order to achieve acceptable precision in the estimation of single AUCs and their differences. Our approach aligns well with the trend toward using CIs for inferences,22,25 by incorporating a pre-specified probability, which is referred to as assurance, of achieving the desired lower confidence limit.

2 Methods

2.1 Proposed sample size formula: Estimation of a single AUC

We first derive a sample size formula for achieving a pre-specified lower bound of the two-sided 95% CI for the AUC with assurance probability 1 − β. Let θ and θ0 denote the pre-specified AUC and the lower bound, respectively. Here 1 − β is called the assurance probability because we aim to compute the required sample size such that $P(\hat{\theta}_L \geq \theta_0) \geq 1 - \beta$ for a given $\theta > \theta_0$, where $\theta_L$ denotes the lower confidence limit of $\theta$.

One may regard the assurance here as equivalent to power in hypothesis testing that focuses on a specific null value, but presenting as assurance probability is more coherent with the goal of estimation, which is to find the range of plausible values based on the data at hand.

Specifically, consider testing a null hypothesis $H_0$: $\text{AUC} = \theta_0$ against an alternative hypothesis $H_1$: $\text{AUC} = \theta > \theta_0$ with a rejection condition $\hat{\theta}_L \geq \theta_0$. This hypothesis test is closely related to the assurance approach because its power is $1 - \beta$. However, it should be noted that the resulting level of the test, $P(\hat{\theta}_L \geq \theta_0 \mid \text{AUC} = \theta_0)$, generally does not equal to $\alpha$, the significance level of the CI and the target level for a one-tail test. For example, when $\sqrt{n}(\hat{\theta} - \theta_0)$ exactly follows a normal distribution $N(0, \sigma^2)$ and that the CI is symmetric, the test level $P(\hat{\theta} - z_{a/2}\sigma / \sqrt{n} \geq \theta_0 \mid \text{AUC} = \theta_0) = \alpha / 2$, where $\hat{\sigma}$ is an estimate of $\sigma$ and $z_{a/2}$ is the upper $\alpha / 2$ quantile of the standard normal distribution. Such a test provides a less transparent interpretation of sample size estimation than the assurance probability.

Prior studies noted poor coverage from Wald-type intervals when the sample size is relatively small.2,26–28 Therefore, we adopt asymmetric CIs based on the logit transform which was found to have better coverage.2,26–28

With derivations presented in Appendix A, we propose the following sample size formula:

$$n = \left[ \frac{z_\beta + z_{a/2}}{\logit(\theta) - \logit(\theta_0)} \right]^2 \frac{f(\theta)}{\theta^2(1-\theta)^2} \pi \frac{\pi}{3}$$

where the logit transformation is defined as $\logit(\theta) = \log(\theta / (1 - \theta))$, $z_\beta$ is the upper $\beta$ quantile of the standard normal distribution, $1 - \alpha$ is the confidence level (conventionally, $\alpha = 0.05$), and $f(\theta)$ is the kernel of the variance for the estimated AUC (factoring out the sample size), that is, $f(\theta) / n = \text{var}(\hat{\theta})$. When $\beta = 0.5$, the sample size reduces to that for achieving an average (or expected) CI width. In other words, sample size based on average width can only have 50% assurance.19

Note that $\pi/3$ in the equation for $n$, the reciprocal for the well-known nonparametric statistic relative efficiency,29 serves as an inflation factor in sample size calculation, given that the nonparametric variance estimation by DeLong et al.30 is commonly used in the stage of data analysis regardless of distribution assumptions made in the stage of sample size planning.
Under a commonly used assumption that data are normally distributed, a variance function has previously been derived\(^3,^8,^9\) with \(f(\theta)\) given by

\[
0.0099e^{-\phi^{-1}(\theta)^2}\left[10\phi^{-1}(\theta)^2 + 8 + \frac{2\phi^{-1}(\theta)^2}{r}ight](r + 1)
\]

where \(\phi^{-1}(\cdot)\) is the inverse of the cumulative distribution function of the standard normal distribution and \(r\) is the sample size ratio of the control group to the disease group. This function, together with a less frequently used version that additionally involves a standard deviation ratio parameter, tends to provide conservative sample size estimates\(^3,^8,^21\); they were originally derived to improve on the variance based on an exponential model,\(^1\) which can underestimate the sample size when the disease and non-disease groups have unequal variances of observations.\(^8\) While a conservative estimate is useful to reduce sample size underestimation, being conservative comes from an overestimation of the true variance. More efficient (and still valid) sample size planning can be accomplished with a more accurate variance function.

Here we propose a new variance function by combining the delta method and the results in Bonett\(^31\) or Reiser and Guttman\(^32\):

\[
f(\theta) = \frac{1}{2}\phi(\phi^{-1}(\theta))^2 \left[\frac{(\phi^{-1}(\theta))^2}{1 + B^2} \left\{ r + 1 + \frac{(r + 1)B^4}{r} \right\} + \frac{2(r + 1) + 2(r + 1)B^2}{r(1 + B^2)} \right]
\]

where \(\phi^{-1}(\cdot)\) is the inverse of the cumulative distribution function of the standard normal distribution, \(\phi(\cdot)\) is the probability density function of the standard normal distribution, \(r\) is the sample size ratio of the control group to the disease group, and \(B\) is the ratio of control participant standard deviation to diseased participant standard deviation. Compared with the conservative option, our variance function needs to additionally specify the value of \(B\), but is intended to give more efficient sample size estimates.

### 2.2 Proposed sample size formula: Estimation of a difference between two AUCs

We now present the sample size required for achieving a prespecified lower bound of the two-sided 95% CI for the difference of two correlated AUCs with assurance probability of \(1 - \beta\). Let \(\theta_1, \theta_2,\) and \(\Delta_0\) denote the prespecified first AUC, second AUC, and the lower bound, respectively. Define \(\theta^*=(\theta_2 - \theta_1 + 1)/2\) and \(\theta_0^*=(\Delta_0 + 1)/2\). Applying the logit transformation to \(\theta^*\) results in constructing CI for a difference between two AUCs (\(\Delta\)) based on a transformation of \(\log((1 + \Delta)/(1 - \Delta))\), which has been applied to a case of a difference between two proportions.\(^33\)

With derivations placed in Appendix B, we arrive at the sample size formula below:

\[
n = \left[\frac{\varphi(\phi^{-1}(\theta_1))}{(1 + B^2)^2} \left\{ r + 1 + \frac{(r + 1)B^4}{r} \right\} + \frac{2(r + 1) + 2(r + 1)B^2}{r(1 + B^2)} \right]
\]

where

\[
f(\theta^*) = \frac{1}{4} \left[f^{(1)}(\theta_1) + f^{(2)}(\theta_2) - 2\rho\sqrt{f^{(1)}(\theta_1)}\sqrt{f^{(2)}(\theta_2)}\right]
\]

\[
f^{(1)}(\theta_1) = \left[\frac{\varphi(\phi^{-1}(\theta_1))}{(1 + B^2)^2} \left\{ r + 1 + \frac{(r + 1)B^4}{r} \right\} + \frac{2(r + 1) + 2(r + 1)B^2}{r(1 + B^2)} \right]
\]

and

\[
f^{(2)}(\theta_2) = \left[\frac{\varphi(\phi^{-1}(\theta_2))}{(1 + B^2)^2} \left\{ r + 1 + \frac{(r + 1)B^4}{r} \right\} + \frac{2(r + 1) + 2(r + 1)B^2}{r(1 + B^2)} \right]
\]

with \(\rho\) being the correlation between two estimated AUCs (from two diagnostic tests), \(r\) the sample size ratio of the control group to the disease group as before, and \(B_t\) the ratio of the standard deviation of the control group to the standard deviation of the disease group for test \(t (t = 1, 2)\). Since two estimated AUCs are calculated from the same group of participants, \(\rho\) is likely to be positive and should be taken into account in statistical procedures.\(^5\)

### 2.3 Evaluation

We conducted an extensive series of simulations to evaluate the performance of the proposed sample size formulas. Two criteria used are (i) the empirical assurance probability (in percent), defined as the percentage of the lower confidence limits
in 10,000 simulation runs that exceeded the pre-specified lower bound and (ii) the empirical coverage (in percent), defined as the percentage of 95% CIs in 10,000 simulation runs that covered the true value of parameters (i.e., single AUC for one-sample setting and the difference of two AUCs for two-sample setting). Valid and efficient sample size planning is expected to have an empirical assurance probability close to the pre-specified assurance probability and empirical coverage close to the nominal 95%. The simulation setup is described below.

**Configurations:** We applied the proposed formulas to compute the required sample size under various configurations: a pre-specified confidence level at 0.95, an assurance probability of 50% or 80%, and a distance of 0.05 or 0.1 between the estimand and the lower bound, denoted by $d$ (i.e., $d = \theta - \theta_0$ in a one-sample case with $\theta = 0.7, 0.8$ or 0.9 and $d = \theta_2 - \theta_1 - \Delta_0$ in a two-sample case with $\theta_2 - \theta_1 = 0.2$). We also considered various combinations of the group size ratio $r$ and the standard deviation ratio $B$, each equal to 1 or 2; in a two-sample setting, we tested a reasonable situation where $B_1$ and $B_2$ are equal and denoted by $B$.

**Data generation process:** Once the sample size was obtained using the formula proposed, we simulated 10,000 data sets each consisting of test values for the disease group from the standard normal distribution and test values for the non-disease group from a normal distribution whose mean and standard deviation were determined by underlying configuration. For a two-sample setting, to allow for a weak, moderate, or strong correlation between two estimated AUCs, we specified a correlation between the rating data of two tests as 0.2, 0.5, or 0.8. This correlation, once specified, can be used to obtain the correlation between two estimated AUCs. Specifically, we generated a sample of size of 5000 with specified rating data correlation, calculated its empirical correlation between two estimated AUCs, repeated the procedure 5000 times, and computed the average value. This yielded approximated between-AUC correlation of 0.15, 0.42, or 0.71 when correlation, calculated its empirical correlation between two estimated AUCs, repeated the procedure 5000 times, and computed the average value. This yielded approximated between-AUC correlation of 0.15, 0.42, or 0.71 when $B = 1$ and 0.13, 0.37, or 0.63 otherwise. While such transformation between two correlations was needed in simulations in order to obtain the between-AUC correlation, in practice, users can specify the anticipated between-AUC correlation directly. We additionally conducted simulations with the between-AUC correlation misspecified as the true value multiplied by 1.1 or 0.9. This sensitivity analysis setting allows for investigating how misspecification of the correlation impacts the performance of the method.

**Analyzing simulated data:** For each simulated data set, we constructed the 95% CI using the DeLong nonparametric method with the logit transformation applied to $\theta$ for a single AUC and $\theta^*$ for the setting of a difference between two AUCs.

For further illustration of the proposed method, we shall present worked examples based on a prior ROC study that compared multiple tracer kinetic analysis methods in terms of ability to provide a good diagnosis of myocardial ischemia.

The simulation study was performed using R Version 3.6.1 software, with the pROC package used to implement DeLong nonparametric variance estimation.

### 3 Results

#### 3.1 Performance of the proposed method

We reported the empirical assurance probability (in percent) and the empirical coverage (in percent) for evaluation. Results in Table 1 suggest for a one-sample setting, the empirical assurance probability is close to the nominal levels in virtually all scenarios considered. For example, the formula predicts that if the true AUC = 0.9, a study with a total sample size of 412 with a 1:1 control-to-disease ratio and common variance would have a lower limit to be above 0.85 with 80% probability, comparable with 83.44% as estimated based on 10,000 simulation runs.

A similar conclusion can be drawn for a two-sample setting from the results in Table 2. For example, if the two AUCs are 0.9 and 0.7, the formula predicts a total sample of 446 with equal group size and common variance would provide a lower limit to be above 0.15 with 80% assurance, comparable with the estimated value of 82.34% based on 10,000 simulation runs. There are a few cases where the estimated assurance probabilities are lower than the nominal level. Specifically, when two tests were strongly correlated, pre-specified assurance probability = 50%, $\theta_2 - \theta_1 - \Delta_0 = 0.1$. This suboptimal performance corresponds to the smallest estimated total sample size of approximately 60. Promising results on assurance probability were observed for sample sizes as small as 66 in a one-sample setting and 92 in a two-sample setting. These results showed the proposed sample size formulas performed reasonably well under a wide range of conditions.

The empirical coverage results were quite close to the nominal 95% in all cases. These results demonstrated that combining the DeLong nonparametric variance estimation with the logit transformation produced accurate CIs even when the total sample size is as small as 63 in the one-sample setting and 56 in the two-sample setting.

Tables 3 and 4 summarize simulation results under the sensitivity analysis setting with between-AUC correlation misspecified as 10% larger than the true value and 10% lower than the true value, respectively. Inflation of the correlation by 10% led to an underestimation of sample size and worse assurance probabilities, and the assurance performance became
more sensitive to a relative change in correlation as the correlation increased. Misspecifying the correlation as 10% lower than the true value resulted in a more conservative sample size and assurance results, as anticipated. The coverage results remained good in both tables, again, demonstrating the reliable performance of the DeLong nonparametric variance estimation when combined with the logit transformation.

### 3.2 Worked examples

As an illustration, consider a study conducted by Biglands et al.\(^{34}\) who compared four tracer kinetic analysis methods for a diagnosis of myocardial ischemia. The study involved 50 participants, 31 without ischemia and 19 with ischemia. Among all methods, the Fermi method gave the highest AUC estimate of 0.92 and the one-compartment model produced the lowest AUC estimate of 0.80, both using myocardial perfusion reserve as the continuous measure in ROC analyses. These two methods were found significantly different with an estimated AUC difference of 0.12 (CI 0.02, 0.21).

Suppose we are planning a future study similar to the above one. Thus it is reasonable to specify the group size ratio as 1.6 (close to 31/19), the standard deviation ratio for the Fermi method as 1.6 (close to 1.02/0.93), and the standard deviation ratio for the one-compartment model as 1.2 (close to 1.31/1.11), where the standard deviations 1.02, 0.93, 1.31, and 1.11 were reported in Biglands et al.\(^{34}\)

Assume the true AUC is 0.92 for the Fermi method. What is the required sample size to achieve a lower bound of 0.8? With \(\theta = 0.92\), \(r = 1.6\) and \(B = 1.1\), we calculate \(f(\theta)\) as \(f(0.92) = 0.0679\). With 80% assurance, the required sample size for the disease and control groups is given by

\[
\begin{align*}
\alpha_D & = \frac{n}{r + 1} = 35.5 \\
\alpha_C & = nr/(r + 1) = 56.9
\end{align*}
\]

Since the sample size is an integer, we require 36 participants with ischemia and 57 without ischemia, adding up to a total of 93 participants. Achieving assurance of 90% requires 125 participants with 48 ischemia cases and 77 controls. When the sample size is 50 as was used in the original study,\(^{34}\) the corresponding assurance probability is 53%.

| \(\theta\) | \(\theta_0\) | \(B\) | \(r\) | \(n\) | ECP | EAP | \(n\) | ECP | EAP |
|---|---|---|---|---|---|---|---|---|---|
| 0.9 | 0.85 | 1 | 1 | 202 | 95.13 | 51.81 | 412 | 95.08 | 83.44 |
| | | 2 | 228 | 94.80 | 51.91 | 464 | 94.83 | 84.25 |
| | | 2 | 224 | 94.84 | 48.56 | 456 | 95.08 | 80.79 |
| | | 2 | 194 | 94.81 | 49.12 | 393 | 94.94 | 81.23 |
| 0.8 | 0.75 | 1 | 1 | 66 | 95.15 | 50.68 | 136 | 95.11 | 86.88 |
| | | 2 | 75 | 94.96 | 52.08 | 152 | 95.18 | 86.77 |
| | | 2 | 74 | 94.54 | 49.02 | 150 | 94.88 | 84.11 |
| | | 2 | 63 | 95.25 | 47.52 | 129 | 95.48 | 84.25 |
| 0.7 | 0.65 | 1 | 1 | 352 | 95.23 | 50.96 | 716 | 95.34 | 82.81 |
| | | 2 | 395 | 94.99 | 51.89 | 806 | 94.82 | 82.85 |
| | | 2 | 370 | 95.46 | 49.37 | 756 | 95.15 | 80.66 |
| | | 2 | 326 | 95.14 | 49.74 | 665 | 95.24 | 80.42 |
| 0.6 | 0.5 | 1 | 1 | 100 | 95.70 | 51.44 | 204 | 95.36 | 84.01 |
| | | 2 | 113 | 95.17 | 51.65 | 230 | 95.04 | 84.02 |
| | | 2 | 106 | 95.15 | 48.05 | 216 | 95.47 | 81.34 |
| | | 2 | 93 | 95.27 | 48.66 | 191 | 95.38 | 82.09 |
| 0.6 | 0.5 | 1 | 1 | 454 | 95.09 | 51.22 | 926 | 94.80 | 81.26 |
| | | 2 | 510 | 95.06 | 50.60 | 1041 | 95.35 | 80.97 |
| | | 2 | 464 | 95.47 | 48.10 | 946 | 94.92 | 79.06 |
| | | 2 | 413 | 95.40 | 48.64 | 843 | 94.50 | 78.63 |
| 0.6 | 0.65 | 1 | 1 | 122 | 95.30 | 51.23 | 248 | 95.42 | 81.36 |
| | | 2 | 137 | 95.07 | 50.31 | 279 | 95.54 | 82.07 |
| | | 2 | 124 | 95.11 | 47.04 | 254 | 95.25 | 79.17 |
| | | 2 | 111 | 95.10 | 48.23 | 225 | 94.88 | 79.94 |

n: estimated sample size; ECP: empirical coverage percent, estimated by the percentage of times that the 2-sided 95% confidence intervals contain the true value across 10,000 simulated data sets; EAP: empirical assurance probability, estimated by the percentage of times that the lower limits of 2-sided 95% confidence intervals exceed the pre-specified value across 10,000 simulated data sets.
Assume the true AUCs for the one-compartment model and the Fermi method are 0.80 and 0.92, respectively. What is the required sample size to achieve a user-specified lower bound of the AUC difference?

\[ \theta_1 = 0.80, \theta_2 = 0.92, r = 1.6, B_1 = 1.2, B_2 = 1.1, \]

we calculated \( f(\theta_1) \) and \( f(\theta_2) \) as

\[ f(\theta_1) = 0.1865 \text{ and } f(\theta_2) = 0.0679, \]

respectively. While the correlation between two AUCs is not available from Biglands et al., the authors noted high correlations (≥0.88) between tracer kinetic analysis methods in terms of myocardial blood flow measures.

Assuming similar correlations for myocardial perfusion reserve measures, it is reasonable to specify \( \rho = 0.8 \) and calculate \( f(\theta^*) \) as

\[ f \left( \frac{0.92 - 0.80 + 1}{2} \right) = \frac{1}{4} \left( 0.1865 + 0.0679 - 2 \times 0.8 \sqrt{0.1865 \times 0.0679} \right) = 0.0186. \]

Assume a lower confidence limit of 0.02 as reported by Biglands et al. Then \( \theta^* = (0.02 + 1)/2 = 0.51. \) For 80% assurance, \( n_D = 23.9 \) and \( n_C = 38.3. \) Thus, the study requires 24 participants with ischemia and 39 without ischemia. For 90% assurance, 85 participants are required, including 33 with ischemia and 52 without.

For 80% assurance, if we wish to reach a larger lower bound at 0.05, 127 participants are required. Finally, ignoring the correlation and specifying \( \rho = 0 \) leads to a perhaps overly conservative sample size of 434.

4 Discussion

We have proposed simple formulas for computing sample sizes required for achieving pre-specified lower bounds of the confidence intervals for single AUCs and their differences. Our method allows for unequal variances between...
test values of the disease and non-disease groups and takes into account the discrepancy in variance estimation comparing sample size planning and data analysis phases (parametric versus nonparametric). While binormal data were assumed in derivations, the proposed method allows for any test data that can be monotonically transformed into binormal (e.g., log normal data) – users would supply input values based on information of the transformed data. Simulation results demonstrated satisfactory performance of the proposed formulas in various practical settings in terms of assurance probability and coverage, even when the total sample size was as small as 70.

Although the variance function in Obuchowski3,8,9 also assumes a binormal distribution, it led to conservative sample size estimates with empirical assurance probability notably higher than the pre-specified value (Table 5 vs. Table 1). For example, this existing method estimates a required total sample size of 650 to achieve a lower limit of 0.85 with 80% probability, assuming the true AUC = 0.9, a 1:1 control-to-disease ratio, and common variance. In comparison, our method returns a more efficient result of 412. When non-normal data distributions are deemed appropriate, \( f(\theta) \) can be replaced with the corresponding variance functions such as equations (6.4) and (6.6) in Zhou et al.3

When specifying correlation (\( \rho \)) in the formula for a difference between two AUCs, a plausible value of the correlation may be obtained with guidance from subject matter knowledge such as prior, similar work in the literature.30 Pilot data, if available, can also be used to estimate the correlation between the two areas.8 When information on the correlation between observations of the test (i.e., raw data) is available, that correlation can be used to obtain the correlation between two AUCs.5 When there is doubt or uncertainty in an available correlation estimate, conservatively specifying a value slightly lower than that estimate might be reassuring especially when the expected correlation is strong, as shown in our simulations. When there is only a rough, qualitative idea about the correlation, that is, weak, moderate, or strong, then a simple

### Table 3

| Correlation | \( \Delta_0 \) | \( B \) | \( r \) | \( n \) | ECP | EAP | \( n \) | ECP | EAP |
|-------------|----------------|------|------|------|-----|-----|------|-----|-----|
| Strong      | 0.15           | 1    | 1    | 184  | 95.02 | 40.92 | 374  | 94.82 | 74.68 |
|             |                | 2    | 1    | 206  | 94.53 | 41.35 | 420  | 95.09 | 75.58 |
|             |                | 2    | 2    | 230  | 94.43 | 41.36 | 468  | 94.90 | 75.40 |
|             | 0.1            | 1    | 1    | 48   | 94.41 | 35.68 | 96   | 94.52 | 75.38 |
|             |                | 2    | 1    | 53   | 94.52 | 36.03 | 107  | 94.36 | 75.59 |
|             |                | 2    | 2    | 60   | 94.21 | 38.60 | 120  | 93.86 | 75.89 |
|             | 0.15           | 1    | 1    | 340  | 94.75 | 46.88 | 694  | 94.73 | 79.00 |
|             |                | 2    | 1    | 383  | 94.51 | 48.34 | 780  | 94.82 | 79.73 |
|             |                | 2    | 2    | 380  | 94.82 | 45.66 | 774  | 94.91 | 77.72 |
|             | 0.1            | 1    | 1    | 88   | 94.66 | 46.19 | 176  | 94.66 | 80.24 |
|             |                | 2    | 1    | 98   | 94.30 | 45.26 | 198  | 94.38 | 79.81 |
|             |                | 2    | 2    | 86   | 94.51 | 44.53 | 176  | 94.76 | 79.39 |
| Moderate    | 0.15           | 1    | 1    | 486  | 94.82 | 50.44 | 992  | 95.15 | 80.91 |
|             |                | 2    | 1    | 546  | 94.99 | 51.76 | 1116 | 95.02 | 80.73 |
|             |                | 2    | 2    | 518  | 95.53 | 48.21 | 1056 | 95.21 | 78.30 |
|             | 0.1            | 1    | 1    | 124  | 94.85 | 49.89 | 252  | 95.25 | 82.22 |
|             |                | 2    | 1    | 140  | 94.67 | 50.68 | 284  | 95.13 | 82.28 |
|             |                | 2    | 2    | 132  | 94.63 | 47.39 | 268  | 95.01 | 79.24 |
| Weak        | 0.15           | 1    | 1    | 486  | 94.82 | 50.44 | 992  | 95.15 | 80.91 |
|             |                | 2    | 1    | 546  | 94.99 | 51.76 | 1116 | 95.02 | 80.73 |
|             |                | 2    | 2    | 518  | 95.53 | 48.21 | 1056 | 95.21 | 78.30 |
|             | 0.1            | 1    | 1    | 124  | 94.85 | 49.89 | 252  | 95.25 | 82.22 |
|             |                | 2    | 1    | 140  | 94.67 | 50.68 | 284  | 95.13 | 82.28 |
|             |                | 2    | 2    | 132  | 94.63 | 47.39 | 268  | 95.01 | 79.24 |

n: estimated sample size; ECP, empirical coverage percent, estimated by the percentage of times that the 2-sided 95% confidence intervals contain the true value across 10,000 simulated data sets; EAP: empirical assurance probability, estimated by the percentage of times that the lower limits of 2-sided 95% confidence intervals exceed the pre-specified value across 10,000 simulated data sets.
Table 4. When between-AUC correlation was misspecified as 10% lower than the true value: performance of the proposed sample size formula for estimating a difference between two AUCs ($\theta_1 = 0.7$ vs $\theta_2 = 0.9$) with pre-specified lower limit ($\Delta_0$) and assurance probability of 50% or 80%, under various degrees of correlation and various combinations of standard deviation ratio ($B$) and sample size ratio ($r$) of the control group to the disease group.

| Correlation | $\Delta_0$ | $B$ | $r$ | $n$ | ECP | EAP | $n$ | ECP | EAP |
|-------------|------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Strong      | 0.15       | 1   | 1   | 254 | 95.06 | 56.74 | 516 | 95.24 | 88.21 |
|             |            | 2   | 285 | 94.87 | 56.71 | 581 | 94.86 | 88.26 |
|             |            | 2   | 296 | 94.80 | 53.30 | 604 | 94.29 | 85.36 |
|             |            | 2   | 263 | 94.34 | 54.89 | 537 | 95.25 | 87.32 |
|             | 0.1        | 1   | 1   | 66  | 94.29 | 53.71 | 132 | 94.69 | 89.06 |
|             |            | 2   | 74  | 94.25 | 53.94 | 149 | 94.72 | 89.73 |
|             |            | 2   | 76  | 94.14 | 51.02 | 154 | 94.76 | 87.42 |
|             |            | 2   | 68  | 94.48 | 52.31 | 137 | 94.59 | 86.61 |
| Moderate    | 0.15       | 1   | 1   | 382 | 95.11 | 52.35 | 778 | 94.94 | 83.90 |
|             |            | 2   | 429 | 95.04 | 52.58 | 876 | 94.97 | 84.24 |
|             |            | 2   | 418 | 94.58 | 50.70 | 854 | 95.13 | 81.28 |
|             |            | 2   | 371 | 95.07 | 52.33 | 756 | 95.15 | 82.50 |
|             | 0.1        | 1   | 1   | 98  | 94.94 | 51.99 | 198 | 94.72 | 84.82 |
|             |            | 2   | 110 | 94.81 | 53.34 | 224 | 94.63 | 85.18 |
|             |            | 2   | 106 | 94.53 | 50.06 | 218 | 95.20 | 83.19 |
|             |            | 2   | 95  | 94.76 | 49.26 | 192 | 95.50 | 83.37 |
| Weak        | 0.15       | 1   | 1   | 500 | 95.50 | 51.38 | 1022 | 95.07 | 82.12 |
|             |            | 2   | 563 | 94.87 | 51.97 | 1151 | 94.99 | 82.18 |
|             |            | 2   | 530 | 94.84 | 48.00 | 1084 | 94.50 | 79.28 |
|             | 0.1        | 1   | 1   | 128 | 95.23 | 51.29 | 260 | 94.85 | 83.53 |
|             |            | 2   | 144 | 95.01 | 52.22 | 293 | 94.96 | 83.34 |
|             |            | 2   | 136 | 94.61 | 48.90 | 276 | 94.72 | 80.72 |

$n$: estimated sample size; ECP: empirical coverage percent, estimated by the percentage of times that the 2-sided 95% confidence intervals contain the true value across 10,000 simulated data sets; EAP: empirical assurance probability, estimated by the percentage of times that the lower limits of 2-sided 95% confidence intervals exceed the pre-specified value across 10,000 simulated data sets.

and intended to be reasonably conservative strategy is to specify the correlation as 0, 0.2, or 0.5. When little is known, one might specify $\rho = 0$ given that $\rho \geq 0$ in most practical settings.

Sometimes the investigator decides to recruit participants by taking a random sample from a large population in which the disease prevalence $p_D$ is known a priori. In this case, the group size ratio $r$ can be specified as $(1 - p_D)/p_D$, because $p_D = n_D/(n_D + n_C) = 1/(1 + r)$.

We emphasize that it is always a good idea to investigate the sensitivity of sample size estimates under a wide range of values for input parameters for the formulas.

The proposed approach is consistent with the trend toward using confidence intervals for inference. In a similar spirit, it is useful to develop sample size formulas for achieving a pre-specified CI width; this direction was previously investigated in reliability studies motivated by inadequate sample size from methods based on expected interval widths. Extending the work presented here to accommodate partial AUC is another future topic, as we have limited our discussion to the area under the entire ROC curve. However, while sometimes it may be clinically meaningful to examine partial AUC over some range of specificity, a partial area is used less frequently than a full area in practice.

In conclusion, closed-from sample size formulas were proposed to help plan ROC studies focusing on estimation of single AUCs and their differences. This approach achieves a desired chance that the lower confidence limit will be above a pre-specified bound. The formulas can be easily computed with a free online calculator at https://dishu.page/calculator/.
Acknowledgements

The authors thank two anonymous reviewers and Dr. Benjamin Reiser for their constructive comments which led to an improved version of this paper.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Dr. G. Zou is partially supported by the Individual Discovery Grant program of the Natural Sciences and Engineering Research Council of Canada.

ORCID iD

Di Shu https://orcid.org/0000-0001-7564-5186

References

1. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29–36.
2. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press, 2003.
3. Zhou XH, Obuchowski NA and McClish DK. Statistical methods in diagnostic medicine. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2011.

Table 5. When using the existing variance function in Obuchowski\textsuperscript{8,9}: performance of the proposed sample size formula for estimating a single AUC (\(\theta\)) with pre-specified lower limit (\(\theta_0\)) and assurance probability 50\% or 80\%, under various standard deviation ratios of the control group to the disease group (\(B\)) and sample size ratios of the control group to the disease group (\(r\)).

| \(\theta\) | \(\theta_0\) | \(B\) | \(r\) | \(n\) | ECP | EAP | \(n\) | ECP | EAP |
|---|---|---|---|---|---|---|---|---|---|
| 0.9 | 0.8 | 1 | 1 | 318 | 95.06 | 73.54 | 650 | 95.37 | 96.51 |
| | | | 2 | 402 | 95.12 | 78.55 | 821 | 94.74 | 97.92 |
| | | | 2 | 402 | 95.35 | 82.37 | 821 | 95.01 | 98.85 |
| 0.8 | 0.75 | 1 | 1 | 104 | 95.29 | 74.62 | 212 | 95.13 | 97.35 |
| | | | 2 | 132 | 95.39 | 80.57 | 267 | 95.65 | 98.43 |
| 0.7 | 0.7 | 1 | 1 | 130 | 95.59 | 63.53 | 266 | 95.46 | 92.31 |
| | | | 2 | 158 | 94.92 | 67.13 | 321 | 95.33 | 93.93 |
| 0.6 | 0.65 | 1 | 1 | 510 | 95.17 | 56.43 | 1040 | 94.86 | 85.34 |
| | | | 2 | 594 | 95.23 | 57.99 | 1214 | 94.75 | 86.44 |
| 0.5 | 0.6 | 1 | 1 | 136 | 95.38 | 55.33 | 278 | 95.09 | 86.16 |
| | | | 2 | 159 | 95.00 | 57.25 | 324 | 95.02 | 87.46 |
| | | | 1 | 136 | 95.46 | 51.59 | 278 | 94.82 | 82.44 |
| | | | 2 | 159 | 95.22 | 63.46 | 324 | 95.15 | 92.33 |

\(n\): estimated sample size; ECP: empirical coverage percent, estimated by the percentage of times that the 2-sided 95\% confidence intervals contain the true value across 10,000 simulated data sets; EAP: empirical assurance probability, estimated by the percentage of times that the lower limits of 2-sided 95\% confidence intervals exceed the pre-specified value across 10,000 simulated data sets.
4. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol* 1975; **12**: 387–415.
5. Hanley JA and McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; **148**: 839–843.
6. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; **351**: h5527.
7. Halpern SD, Karlawish JH and Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002; **288**: 358–362.
8. Obuchowski NA. Computing sample size for receiver operating characteristic studies. *Investig Radiol* 1994; **29**: 238–243.
9. Obuchowski NA, Lieber ML and Wians Jr FH. ROC curves in clinical chemistry: Uses, misuses, and possible solutions. *Clin Chem* 2004; **50**: 1118–1125.
10. MedCalc: Sample size calculation for area under ROC curve. https://www.medcalc.org/manual/sample-size-ROC-AUC.php (accessed 7 July 2022).
11. MedCalc: Sample size calculation for comparison of two ROC curves. https://www.medcalc.org/manual/sample-size-ROC-comparison.php (accessed 7 July 2022).
12. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform* 2011; **12**: 77.
13. PASS 2022 power analysis and sample size software. Kaysville, UT, USA: NCSS, LLC, 2022. https://www.ncss.com/software/pass
14. Goksuluk D, Korkmaz S, Zararsiz G, et al. easyROC: An interactive web-tool for ROC curve analysis using R language environment. *R J* 2016; **8**: 213–230.
15. Nahm FS. Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean J Anesthesiol* 2022; **75**: 25–36.
16. Mathews P. *Sample size calculations: Practical methods for engineers and scientists*. Fairport Harbor, OH: Mathews Malnar and Bailey, 2010.
17. Reiser B and Guttmann I. Sample size choice for reliability verification in strength-stress models. *Can J Stat* 1989; **17**: 253–259.
18. Reiser B, Faraggi D and Guttmann I. Choice of sample size for testing the P(X > Y). *Commun Stat Theory Methods* 1992; **21**: 559–569.
19. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Stat Med* 2012; **31**: 3972–3981.
20. Flahault A, Cadilhac M and Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005; **58**: 859–862.
21. La G. Sample size formulas for estimating areas under the receiver operating characteristic curves with precision and assurance. Electronic Thesis and Dissertation Repository. 8045, 2021. https://ir.lib.uwo.ca/etd/8045
22. Gardner MJ and Altman DG. Confidence intervals rather than P values: Estimation rather than hypothesis testing. *BMJ* 1986; **292**: 746–750.
23. El Khouli RH, Macura KJ, Barker PB, et al. Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. *J Magn Reson Imaging* 2009; **30**: 999–1004.
24. Muller MP, Tomlinson G, Marrie TJ, et al. Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia? *Clin Infect Dis* 2005; **40**: 1079–1086.
25. Greenland S. On sample-size and power calculations for studies using confidence intervals. *Stat Med* 2006; **25**: 559–573.
26. Newcombe RG. Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: Asymptotic methods and evaluation. *Stat Med* 2006; **25**: 559–573.
27. Zou GY and Yue L. Using confidence intervals to compare several correlated areas under the receiver operating characteristic curves. *Stat Med* 2013; **32**: 5077–5090.
28. Perme MP and Manevski D. Confidence intervals for the Mann–Whitney test. *Stat Methods Med Res* 2019; **28**: 3755–3768.
29. Lehmann EL and Romano JP. *Testing statistical hypotheses*. 3rd ed. New York: Springer, 2005.
30. DeLong ER, DeLong DM and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; **44**: 837–845.
31. Bonett DG. Confidence intervals for standardized linear contrasts of means. *Psychol Methods* 2008; **13**: 99–109.
32. Reiser B and Guttmann I. Statistical inference for Pr (Y < X): The normal case. *Technometrics* 1986; **28**: 253–257.
33. Zou G and Donner A. A simple alternative confidence interval for the difference between two proportions. *Control Clin Trials* 2004; **25**: 3–12.
34. Biglands JD, Magee DR, Sourbron SP, et al. Comparison of the diagnostic performance of four quantitative myocardial perfusion estimation methods used in cardiac MR imaging: CE-MARC substudy. *Radiology* 2015; **275**: 393–402.
35. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. http://www.R-project.org/
36. Beal SL. Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* 1989; **45**: 969–977.
37. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989; **9**: 190–195.
38. Dodd ZE and Pepe MS. Partial AUC estimation and regression. *Biometrics* 2003; **59**: 614–623.
Appendix A: Deriving sample size formula for estimation of a single AUC

Application of the delta method\(^{19,21}\) with a logit transformation for AUC yields a general sample size formula as

\[
n = \left\{ \frac{z_\alpha + z_{\alpha/2}}{\logit(\theta) - \logit(\theta_0)} \right\}^2 f(\theta) / \theta^2 (1 - \theta)^2
\]

(1)

where the logit transformation gives \(\logit(\theta) = \log(\theta / (1 - \theta))\), \(z_\alpha\) is the upper \(\alpha\) quantile of the standard normal distribution, \(1 - \alpha\) is the confidence level, and \(f(\theta)\) is the kernel of the variance for the estimated AUC (factoring out the sample size), that is, \(f(\theta) / n = \text{var}(\hat{\theta})\).

To obtain \(f(\theta)\), a parametric distribution of data is typically specified so that the variance can be described by a few parameters. We make a common assumption that the test values are normally distributed. Let random samples \(X_i (i = 1, \ldots, n_C) \sim N(\mu_C, \sigma_C^2)\) and \(Y_j (j = 1, \ldots, n_D) \sim N(\mu_D, \sigma_D^2)\) be test values for participants without disease (control) and for participants with disease, respectively, where \(n_D\) denotes the size of the disease group, \(n_C\) the size of the control group, and \(r\) the sample size ratio of the control group to the disease group. Define \(B\) as the standard deviation ratio of the control group to the disease group, that is, \(B = \sigma_C / \sigma_D\).

We first work under the assumption that a larger test value is indicative of stronger evidence of a participant being diseased, which gives \(\text{AUC} = P(Y > X) + 0.5P(Y = X)\), where \(P(Y = X) = 0\) for continuous test values.\(^4\) With binormal test values, the \(\text{AUC}\) can be further written as \(\phi(\eta^2 / \sqrt{2})\) where \(\eta = (\mu_D - \mu_C) / (\sqrt{\sigma_D^2 + \sigma_C^2}) / 2\) and \(\phi(\cdot)\) is the cumulative distribution function of the standard normal distribution, leading to an \(\text{AUC}\) estimator

\[
\hat{\theta} = \phi\left(\frac{\hat{\eta}}{\sqrt{2}}\right)
\]

where

\[
\hat{\eta} = \frac{(1 / n_D) \sum_{j=1}^{n_D} Y_j - (1 / n_C) \sum_{i=1}^{n_C} X_i}{\sqrt{(\sigma_D^2 + \sigma_C^2) / 2}}
\]

By the delta method, we approximate \(\text{var}(\hat{\theta})\) by \((1 / 2)[\phi(\eta / \sqrt{2})]^2 \text{var}(\hat{\eta})\), where \(\phi(\cdot)\) is the derivative of \(\phi(\cdot)\) and hence the probability density function of the standard normal distribution, and \(\text{var}(\hat{\eta})\) is readily approximated by results available in Bonett,\(^{31}\) which are equivalent to results in Reiser and Gutman.\(^{32}\) Specifically,

\[
\text{var}(\hat{\eta}) \approx \frac{\eta^2}{2(\sigma_D^2 + \sigma_C^2)^2} \left( \frac{\sigma_D^2}{n_D} + \frac{\sigma_C^2}{n_C} \right) + \frac{2\sigma_D^2}{(\sigma_D^2 + \sigma_C^2)n_D} + \frac{2\sigma_C^2}{(\sigma_D^2 + \sigma_C^2)n_C}
\]

\[
= \frac{\eta^2}{2(1 + B^2)^2} \left\{ \frac{r + 1}{n} + \frac{(r + 1)B^2}{rn} \right\} + \frac{2(r + 1)}{1 + B^2} \frac{(r + 1)}{n} + \frac{2(r + 1)B^2}{1 + B^2} \frac{(r + 1)}{rn}
\]

where \(n = n_D + n_C\) is the total sample size. Factoring \(n\) out and substituting \(\eta = \sqrt{2}\phi^{-1}(\theta)\) yields

\[
f(\theta) = \frac{1}{2}[\phi(\phi^{-1}(\theta))]^2 \left\{ \frac{\phi^{-1}(\theta)}{1 + B^2} \right\}^2 \left( \frac{r + 1 + (r + 1)B^2}{r} \right) + \frac{2(r + 1)}{1 + B^2} + \frac{2(r + 1)B^2}{r(1 + B^2)}
\]

Under a different assumption that a smaller test value, instead of a larger test value, indicates stronger evidence of a participant being diseased, \(\text{AUC} = P(Y < X)\). In this case, \(\hat{\theta} = \phi(-\hat{\eta} / \sqrt{2})\) and \(\text{var}(\hat{\theta})\) can be approximated by \((1 / 2)[\phi(-\eta / \sqrt{2})]^2 \text{var}(\hat{\eta}) = (1 / 2)[\phi(\phi^{-1}(\theta))]^2 \text{var}(\hat{\eta})\), implying that the same above \(f(\theta)\) is applicable to both assumptions.

While sample size now can be computed by substituting \(f(\theta)\) into equation (1), the resulting sample size may still be inadequate. The reason is that the DeLong nonparametric approach\(^{10}\) to the analysis of a single AUC or areas under two correlated ROC curves has been commonly used in practice, with different efficiency compared to a parametric approach based on which the sample size is planned. Ignoring this discrepancy might lead to inadequate sample size, because nonparametric estimation is often slightly less efficient than the parametric counterpart.
To deal with this discrepancy, we further inflate the sample size by $\pi / 3$, the reciprocal for the well-known nonparametric statistic relative efficiency,$^{39}$ and arrive at the proposed sample size formula:

$$
n = \left\{ \frac{z_{\beta} + z_{\alpha/2}}{\logit(\theta) - \logit(\theta_0)} \right\}^2 \times \frac{\pi f(\theta^*)}{\logit(\theta) - \logit(\theta_0)} \left[ \frac{\phi^{-1}(\theta) - \phi^{-1}(\theta_0)}{(1 + B^2_{1})} \right]^2 \left[ r + 1 + \frac{(r + 1)B^2_{1}}{r} \right] \left[ \frac{2(r + 1)B^2_{2}}{r(1 + B^2_{1})} \right] \left[ 6\theta^*(1 - \theta^*) \right]$$

Notably, it can be shown that $(r, B) = (a, b)$ and $(r, B) = (1/a, 1/b)$ produce the same estimated total sample size.

Appendix B: Deriving sample size formula for estimation of a difference between two AUCs

We begin by introducing the notations. For test $t$, where $t = 1, 2$, let random samples $X^{(t)}_i (i = 1, \ldots, n_D) \sim N(\mu_{Ct}, \sigma^2_{Ct})$ and $Y^{(t)}_j (j = 1, \ldots, n_D) \sim N(\mu_{Dt}, \sigma^2_{Dt})$ be normally distributed test values for participants without disease (control) and for participants with disease, respectively. Define $B_t$ as the standard deviation ratio, that is, $B_t = \sigma_{Ct} / \sigma_{Dt}$, for $t = 1, 2$.

We derive the sample size formula through $(\theta_2 - \theta_1 + 1)/2$, which ranges from 0 to 1, in order to leverage formula (1) based on the logit transformation. Noticing that the lower bound of the 95% CI for $(\theta_2 - \theta_1 + 1)/2$ is $(\Delta_0 + 1)/2$, we define $\theta^* = (\theta_2 - \theta_1 + 1)/2$ and $\theta^*_0 = (\Delta_0 + 1)/2$. Note that using $\theta^*$ is equivalent to constructing a CI for the difference of AUCs ($\Delta$) based on a transformation $\log((1 + \Delta)/(1 - \Delta))$, a strategy known to perform well for a difference between two proportions.33

Let $\theta^* = (\theta_2 - \theta_1 + 1)/2$. By equation (1), we obtain a general sample size formula for a two-sample case (after applying the efficiency adjustment factor $\pi / 3$):

$$
n = \left\{ \frac{z_{\beta} + z_{\alpha/2}}{\logit(\theta) - \logit(\theta_0)} \right\}^2 \times \frac{\pi f(\theta^*)}{\logit(\theta) - \logit(\theta_0)} \left[ \frac{\phi^{-1}(\theta) - \phi^{-1}(\theta_0)}{(1 + B^2_{1})} \right]^2 \left[ r + 1 + \frac{(r + 1)B^2_{1}}{r} \right] \left[ \frac{2(r + 1)B^2_{2}}{r(1 + B^2_{1})} \right] \left[ 6\theta^*(1 - \theta^*) \right]$$

where $f(\theta^*)$ is the kernel of $\text{var}(\hat{\theta}^*)$ given by

$$
\text{var}(\hat{\theta}^*) = \frac{1}{4} \left\{ \text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2) - 2\rho \sqrt{\text{var}(\hat{\theta}_1) \text{var}(\hat{\theta}_2)} \right\}.
$$

Here $\rho$ is the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$, the estimated AUCs under two tests.

Factoring $n$ out yields

$$
f(\theta^*) = \frac{1}{4} \left\{ f^{(1)}(\theta_1) + f^{(2)}(\theta_2) - 2\rho \sqrt{f^{(1)}(\theta_1) f^{(2)}(\theta_2)} \right\}
$$

where

$$
f^{(1)}(\theta_1) = \frac{1}{2} \left[ \phi^{-1}(\theta_1) \right]^2 \left[ \frac{2(r + 1)B^2_{1} + 2(r + 1)B^2_{2}}{r(1 + B^2_{1})} \right]
$$

is the kernel of $\text{var}(\hat{\theta}_1)$ and

$$
f^{(2)}(\theta_2) = \frac{1}{2} \left[ \phi^{-1}(\theta_2) \right]^2 \left[ \frac{2(r + 1)B^2_{1} + 2(r + 1)B^2_{2}}{r(1 + B^2_{1})} \right]
$$

is the kernel of $\text{var}(\hat{\theta}_2)$.

We arrive at the proposed sample size by substituting $f(\theta^*)$ in equation (4) into equation (3).