Published in final edited form as:

Stat Methods Med Res. 2017 April ; 26(2): 528–541. doi:10.1177/0962280214551190.

Jackknife variance of the partial area under the empirical ROC curve

Andriy I. Bandos^{1,*}, Ben Guo¹, and David Gur²

¹Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, U.S.A

²Department of Radiology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15216, U.S.A

Abstract

Receiver Operating Characteristic (ROC) analysis provides an important methodology for assessing traditional (e.g., imaging technologies and clinical practices) and new (e.g., genomic studies, biomarker development) diagnostic problems. The area under the clinically/practically relevant part of the ROC curve (partial area or pAUC) is an important performance index summarizing diagnostic accuracy at multiple operating points (decision thresholds) that are relevant to actual clinical practice. A robust estimate of the pAUC is provided by the area under the corresponding part of the empirical ROC curve. We derive a closed-form expression for the jackknife variance of the partial area under the empirical ROC curve. Using the derived analytical expression we investigate the differences between the jackknife variances and a conventional variance estimator. The relative properties in finite samples are demonstrated in a simulation study. The developed formula enables an easy way to estimate the variance of the empirical pAUC, thereby substantially reducing the computation burden, and provides important insight into the structure of the variability. We demonstrate that when compared with the conventional approach, the jackknife variance has negligible bias, and leads to a more appropriate type I error rate of the Wald-type test. The use of the jackknife variance is illustrated in the analysis of a dataset from a diagnostic imaging study.

Keywords

jackknife; closed-form; nonparametric estimation; partial AUC; ROC analysis

1 Introduction

Receiver Operating Characteristic (ROC) analysis is a commonly accepted approach for evaluating the accuracy of diagnostic tests. A large number of statistical inferences based on ROC indices (e.g., area under the ROC curve, or AUC) rely on jackknife approaches [1–3], which proved to have excellent properties for a broad range of practical problems. One of the most widely known nonparametric approaches for statistical inferences regarding the

* anb61@pitt.edu; phone: +1 412 383 573; fax: +1 412 624 2183.

AUC [4] is equivalent to the Wald-test with a variance estimated using a *two-sample jackknife* approach [5]. The conventional *one-sample jackknife* approach leads to a very similar, although a somewhat more conservative, result [6]. Some extensions of conventional ROC analyses also successfully utilize the jackknife approach for statistical inferences [7].

The area under a clinically/practically relevant portion of the ROC curve (called partial AUC, or pAUC) is one of the most recognized ROC indices for practically relevant inferences [8–10]. Several approaches for statistical inferences for partial AUC have been developed over the last three decades [11–13]. Recent investigations also demonstrated that in a number of practical scenarios the partial AUC may enable more efficient statistical inferences than the area under the entire ROC curve [14].

A nonparametric estimator of the partial AUC can be obtained in a straightforward manner as the area under the empirical ROC curve over the range of interest (which is often specified in terms of the acceptable levels of 1-specificity, or False Positive Fraction, FPF). In the presence of continuous data, empirical partial AUC over the empirical points in the range of interest provides a consistent estimator with an approximate normal sampling distribution for large sample sizes [11]. Estimation of the variability of the pAUC estimator is not straightforward and several variance estimation approaches were proposed for this purpose [e.g., 11–13].

In this work we derive closed-form expressions of the two-sample jackknife variance of the partial area under the empirical ROC curve over the range of interest $f \in (0, e)$. We demonstrate the differences in formulations between jackknife variance and a conventional variance estimator [13,10] and we conduct a simulation study of the relative properties of these variance estimators. Advantages of the jackknife variance are illustrated on a sample dataset from an actual diagnostic imaging observer study. We then summarize our findings and discuss the practical implications in the Discussion section.

2 Methods

Let $\{x_j\}$, $j=1, \dots, n_x$, denote diagnostic results for a sample “non-diseased” subjects and $\{y_j\}$, $j=1, \dots, n_y$, denote diagnostic results for a sample of “diseased” subjects. In the target population, diagnostic results follow distributions with survival functions of $S_X(x) = P(X > x)$, and $S_Y(y) = P(Y > y)$, respectively. The corresponding Receiver Operating Characteristic (ROC) curve can then be written as $ROC(f) = S_Y(S_X^{-1}(f))$, where f swaps all values of the False Positive Fraction (FPF) from 0 to 1. And the partial area under the ROC curve (pAUC) over the range $f \in (0, e)$ is

$$A(e) = \int_0^e ROC(f) df$$

The empirical ROC curve can be constructed using the empirical (linearly interpolated) survival functions, which is equivalent to connecting adjacent empirical points $(\widehat{FPF}, \widehat{TPF})$

with straight lines, where $\widehat{FPF}(c) = \sum_{i=1}^{n_x} I(x_i > c)/n_x$ and $\widehat{TPF}(c) = \sum_{j=1}^{n_y} I(y_j > c)/n_y$. The partial area under the empirical ROC curve, i.e., the empirical pAUC, can be written using AUC-like formulation with x 's constrained to the region of values corresponding to the range of FPF's that is of interest [e.g., 12,13]. For the purposes of deriving the variance estimator it is however convenient to express the empirical partial AUC in terms of the order statistics (circumventing the region of test results), i.e.:

$$\hat{A}(e) = \underbrace{\frac{1}{n_x n_y} \sum_{i=n_x - \lfloor en_x \rfloor + 1}^{n_x} \sum_{j=1}^{n_y} I(\chi_i < y_j)}_{\hat{\theta}_e} + \frac{1}{n_y} \left(\sum_{j=1}^{n_y} I(\chi_{n_x - \lfloor en_x \rfloor} < y_j) \right) \times \left(e - \frac{1}{n_x} \lfloor en_x \rfloor \right) \quad (1)$$

where χ_i denotes an i^{th} order statistic of ratings for non-diseased subjects and $\lfloor \bullet \rfloor$ denotes an integer part of the argument. The first summand in equation (1),

$\hat{\theta}_e = \frac{1}{n_x n_y} \sum_{i=n_x - \lfloor en_x \rfloor + 1}^{n_x} \sum_{j=1}^{n_y} I(\chi_i < y_j)$, is the area under the empirical ROC curve up to the last empirical point in the range (0,e). For continuous data this is also a consistent estimator for the partial AUC [11,13]. The second summand is the area under the empirical ROC curve over the range ($\lfloor en_x \rfloor/n_x$, e), which is asymptotically negligible (For data with ties, however, the second term may be substantial even for large sample sizes, and the formulation for the empirical pAUC becomes more cumbersome.) As was demonstrated by Wieand et al. [11], for large sample size $\hat{\theta}_e$ is approximately normally distributed with mean $A(e)$ and variance σ^2 , where

$$\sigma^2 = \frac{1}{n_x} \int_0^e \int_0^e [ROC'(p)ROC'(q) \{ \min(p, q) - pq \}] dpdq + \frac{1}{n_y} \int_0^e \int_0^e [ROC(\max(p, q)) \{ 1 - ROC(\min(p, q)) \}] dpdq \quad (2)$$

where $ROC'(f)$ is a derivative of the ROC curve at point $(f, ROC(f))$. Consistent estimate of σ^2 can then be obtained by substituting in equation (2) the values of sensitivity and slopes at the points on the empirical ROC curve [11].

equation (1) provides a convenient way for deriving the formulation for traditional jackknife pseudo-values for the empirical pAUC since all, but possibly one, summands are undisturbed by elimination of a single observation. The jackknife variance can then be estimated as a sample variance of all pseudovalues ("one-sample" jackknife) or, recognizing the existence of two independent samples [5], as a sum of sample variances for pseudovalues based on diseased and non-diseased subjects ("two-sample" jackknife variance). Both approaches allow for a closed-form formulation, but the two-sample jackknife variance is preferable in general and in the case of empirical pAUC in particular. In fact, for the full

AUC, the two-sample jackknife variance is equivalent to the variance estimator proposed by DeLong et al. [Ref. 4, p. 840].

The derivation of a closed-form expression for the jackknife variance is best illustrated on the jackknife variance for the partial area up to the last empirical point in the range (i.e., $\hat{\theta}_e$) under the scenario in which the range size e is such that $e < en_x - \lfloor en_x \rfloor$. This condition ensures that the number of empirical points within the range of interest, $t = \lfloor en_x \rfloor$ is undisturbed by deletion of a single observation (i.e., $\lfloor e(n_x - 1) \rfloor = \lfloor en_x \rfloor$). We note that the derivation of the formulation for larger e follows the same steps as outlined below albeit, details are omitted for brevity. For this scenario all changes imposed by the deletion of a single observation can be grouped into following three categories related to:

1. deletion of a non-diseased subject with one of the $n_x - t$ lowest ratings ($t = \lfloor en_x \rfloor$),
2. deletion of a non-diseased subject with one of the t highest ratings, and
3. deletion of a diseased subject.

The first scenario, namely, exclusion of the k^{th} low-scored non-diseased subject ($k=1, \dots, n_x - t$), leads to the jackknifed pAUC (i.e., $\hat{\theta}_e$ computed on $n_x - 1$ x's and n_y y's) that can be formulated as follows:

$$\hat{\theta}_{(-k, \bullet)} = \frac{1}{(n_x - 1)n_y} \sum_{i=n_x-t+1}^{n_x} \sum_{j=1}^{n_y} I(\chi_i < y_j) = \frac{n_x}{n_x - 1} \hat{\theta} \quad (3)$$

The second scenario, namely, exclusion of the k^{th} high-scored non-diseased subject ($k=n_x - t + 1, \dots, n_x$) leads to the following expression:

$$\hat{\theta}_{(-k, \bullet)} = \frac{\sum_{i=n_x-t+1}^{n_x} \sum_{j=1}^{n_y} I(\chi_i < y_j)}{(n_x - 1)n_y} = \frac{n_x}{n_x - 1} \hat{\theta} + \frac{1}{(n_x - 1)n_y} \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) - \frac{1}{(n_x - 1)n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j) \quad (4)$$

Finally, exclusion of the l^{th} diseased subject ($l=1, \dots, n_y$) leads to the following statistic:

$$\hat{\theta}_{(\bullet, -l)} = \frac{1}{n_x(n_y - 1)} \sum_{i=n_x-t+1}^{n_x} \sum_{\substack{j=1, \\ j \neq l}}^{n_y} I(\chi_i < y_j) = \frac{n_y}{n_y - 1} \hat{\theta} - \frac{1}{n_x(n_y - 1)} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l). \quad (5)$$

Computation of the two-sample jackknife variance requires separate averaging of the jackknifed pAUCs corresponding to the excluded diseased and non-diseased subjects. The average corresponding to non-diseased subjects assumes the following expression:

$$\hat{\theta}_{(-\bullet, \bullet)} = \hat{\theta} + \frac{t}{n_x(n_x - 1)} \times \frac{\sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j)}{n_y} \quad (6)$$

and the average corresponding to diseased subjects becomes:

$$\hat{\theta}_{(\bullet, -\bullet)} = \frac{\sum_{l=1}^{n_y} \hat{\theta}_{(\bullet, -l)}}{n_y} = \hat{\theta} \quad (7)$$

By definition, the two-sample jackknife variance can be computed as the sum of sample variances of the pseudovalues corresponding to diseased and non-diseased subjects [5]. However, each sample variance of pseudovalues can be re-written as a scaled sample variance of the jackknifed statistics [e.g., 3], resulting in the following expression for the two-sample jackknife variance:

$$\hat{V}_{J_2}(\hat{\theta}) = \frac{n_x - 1}{n_x} \sum_{k=1}^{n_x} \left(\hat{\theta}_{(-k, \bullet)} - \hat{\theta}_{(-\bullet, \bullet)} \right)^2 + \frac{n_y - 1}{n_y} \sum_{l=1}^{n_y} \left(\hat{\theta}_{(\bullet, -l)} - \hat{\theta}_{(\bullet, -\bullet)} \right)^2 \quad (8)$$

Plugging-in expressions (3) to (7) into equation (8), leads to the following closed-form formulation of the two-sample jackknife variance of the nonparametric pAUC estimator:

$$\begin{aligned} \hat{V}_{J_2}(\hat{\theta}) = & \frac{n_x - t}{(n_x - 1)t} \left(\hat{\theta} - \frac{t}{n_y n_x} \times \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right)^2 \\ & + \frac{1}{(n_x - 1)n_x} \sum_{k=n_x-t+1}^{n_x} \left\{ \frac{n_x \hat{\theta}}{t} - \frac{1}{n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j) \right\}^2 + \frac{t^2}{(n_y - 1)n_x^2 n_y} \sum_{l=1}^{n_y} \left(\frac{n_x \hat{\theta}}{t} - \frac{1}{t} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l) \right)^2 \end{aligned} \quad (9)$$

The last two terms (with a slightly different coefficient in the first) determine the variance estimator proposed by He and Escobar [13]. The first term, however, is functionally independent and of the same order of magnitude as the last two terms; hence it should not be ignored even for large sample sizes. The first term involves the estimate of the level of sensitivity at the first operating point outside the range of interest (corresponding to threshold χ_{n_x-t}), which is a component of the slope of the empirical ROC curve [11] at the last point in the range of interest (t/n_x , $ROC(t/n_x)$). The slopes at all empirical ROC points within the range of interest determine the $1/n_x$ -order component of the asymptotic variance of $\hat{\theta}_e$ (equation 2). As we demonstrate below, the first term in equation (9) contributes substantially to the overall value of the jackknife variance, leading to a superior variance

estimator (due to better accounting for the variability at the last empirical point in the range of interest). In the next section we illustrate the properties of the jackknife variance in a simulation study.

3 Simulation study

3.1 Methods

We evaluated the finite-sample properties of the jackknife and the conventional variance estimators [10,13] of the partial AUC. Variance estimators were assessed in terms of their estimated relative bias and type I error rate of the corresponding Wald-type tests. Simulated datasets corresponded to binormal ROC curves [15] with parameters that were typical and generally expected in the medical diagnostic area.

“Binormal” ROC curves provide a flexible model for fitting various types of empirical data and are often used in practice [10]. Binormal ROC curves are conventionally parameterized with two parameters a and b ($b > 0$) as follows:

$$ROC(f) = \Phi \left(a + b \times \Phi^{-1}(f) \right)$$

where f is the False Positive Fraction (FPF), or equivalently $1 - \text{specificity}$, and $ROC(f)$ is the True Positive Fraction (TPF), or equivalently, sensitivity , corresponding to f ; $\Phi(\cdot)$ is a cumulative density function (c.d.f.) of the standard normal distribution, and $\Phi^{-1}(\cdot)$ is the corresponding percentile function (i.e., the inverse c.d.f.). The parameter b determines the shape of the ROC curve. In particular, the ROC curve is concave when $b=1$, and it has a non-concavity region (“hook”) in the right corner when $b < 1$, and in the left corner when $b > 1$. For a fixed b , higher values of the parameter a lead to a higher ROC curve, hence, to a larger area under the ROC curve (AUC). The AUC can then be computed from parameters a and b as:

$$AUC = \Phi \left(a / \sqrt{1+b^2} \right)$$

We considered binormal ROC curves with concave shape ($b=1$) and moderate improperness in the range of low specificity ($b=0.5$) and with full AUCs ranging from 0.6 to 0.9. Several of these ROC curves are illustrated in figure 1.

Simulation scenarios were determined by a specific ROC curve (i.e., fixed b and AUC), sample size, and the range of interest for pAUC (i.e., integration range). We present results for sample sizes corresponding to a relatively small study (with 50 non-diseased and 50 diseased subjects), a moderately large study with (100 non-diseased and 100 diseased subjects) where Wald-type tests are expected to have satisfactory properties, as well as large sample sizes (500 non-diseased and 500 diseased subjects) in order to illustrate the trends with increasing sample sizes. Considered integration regions for pAUC included ranges of possible clinical interest $fpf \in (0,0.05), (0,0.1), (0,0.2)$ as well as a wider range of interest $(0,0.5)$ for comparison with results in the literature.

Variance estimators were evaluated in terms of their estimated relative bias and type I error rate of the corresponding Wald-type tests. For each dataset the relative bias was estimated using the ratio of the computed variance estimator ($\hat{V}_\bullet$) to the Monte Carlo variance of the pAUC estimated from 10,000 datasets for each scenario, i.e., $RB = (\hat{V}_e - \hat{V}_{MC}) / \hat{V}_{MC}$. The type I error rate was estimated as the fraction of the 10,000 simulations in which the absolute values of the Wald type statistic was larger than 0.975th normal percentile, i.e.

$\left| (\hat{\theta}_e - A(e)) / \sqrt{\hat{V}_\bullet(\hat{\theta}_e)} \right| > z_{0.975}$, where $A(e)$ was determined by both the ROC curve and the integration range used in each simulation scenario.

3.2 Results

Table 1 summarizes the estimated relative bias and type I error rates corresponding to the proposed variance estimator, $\hat{V}_{J2}$, and the conventional variance estimator, $\hat{V}_{HE}$, for the same nonparametric estimators of the pAUC, $\hat{\theta}_e$. For inferences based on moderate sample sizes (100 diseased: 100 non-diseased subjects), the bias of the jackknife variance does not exceed 13%, whereas the bias of the conventional variance estimator ranges from -18% to -41% when a typical clinical range of interest of (0,0.1) is considered. The estimates of the type I error rate of the jackknife-based Wald test are somewhat elevated but overall are reasonably close to the nominal 0.05 level, whereas the Wald-test with the conventional variance estimator has type I error estimates as high as 0.15.

For the larger sample size (500:500), the bias of the jackknife variance becomes negligible (<3%), whereas bias of the conventional estimator remains consistently high in the range from -20% to -40%. The type I error rate for the jackknife-based Wald test is close to the nominal 0.05 level, whereas for the conventional variance estimator the type I error rate remains significantly elevated. Figure 2 illustrates the overall impact of increasing sample size on the relative bias for inferences based on the pAUC_(0,0.1) in the case of concave binormal ROC curves with a full AUC of 0.8.

We note that for smaller sample sizes the properties of both variance estimators deteriorate. This is clearly evident in the results corresponding to a very short integration range (e.g., (0,0.05)). Shortening the integration range reduces the effective sample size (i.e., the number of points used in computations) thereby degrading the properties of both the estimator of the partial AUC and its variance estimator. The empirical estimator of partial AUC is known to be biased [17], and our results demonstrate that for a narrow range (e.g., (0,0.05)) with small sample sizes this bias is rather substantial, leading to an elevated type I error rate of the Wald test. Thus, for small sample sizes and/or short integration ranges the properties of both the jackknife and the conventional approaches are expected to be suboptimal.

When considering a substantially wider range of interest (0, 0.5), which is not of interest in most medical diagnostic applications, but may be relevant for other fields, the properties of the conventional variance estimator improves somewhat. However, for a range as wide as (0, 0.5) the bias of the conventional variance estimator can still be as high as 30%. The approaches coincide for full AUC when the range of interest is (0, 1), since for the full range both variance estimators become equivalent to variance estimator by DeLong et al. [4].

4 Example

In this example we illustrate the advantages of using the jackknife variance for analysis of pilot data. The data for this example were taken from an observer study investigating the effect of luminance and spatial resolution during the interpretation of chest radiographs for the presence of interstitial disease [16]. We used a subset of data corresponding to a single diagnostic modality and single-reader evaluating 223 images non-depicting and 84 images depicting the presence interstitial lung disease. The empirical ROC curve is illustrated in Figure 3.

The data consist of the pseudo-continuous confidence ratings (0 to 100) reflecting the opinion of a radiologist regarding the presence (or absence) of interstitial lung disease on a given posterior-anterior chest radiograph. Actual presence of interstitial disease for each patient was known from independent assessments such as chest CT or subsequent imaging depicting disease presence and/or disease progression. Using specific threshold confidence ratings could be categorized (greater/less than the threshold) for providing recommendations for a particular clinical action (e.g., additional imaging based diagnostic work-up such as CT, or routine follow-up). The proportion of diseased subjects with ratings above the specific threshold defines the corresponding sensitivity, and the proportion of non-diseased subjects with rating below the threshold defines the corresponding specificity. The ROC curve allows for an assessment of the overall ability of the radiologist's confidence ratings to discriminate between diseased and non-diseased subjects, for all thresholds simultaneously. The area under the ROC curve (AUC) quantifies the overall discriminative ability of all confidence ratings, whereas the partial AUC over the range (0,e) corresponds to the discriminative ability of the high-value ratings.

We used this dataset as hypothetical pilot data for planning a future study of diagnostic performance under the imaging modality being evaluated. The estimated value of the partial AUC over the range (0, 0.1) was 0.027, which corresponded to the standardized partial AUC of 0.62. The standardized partial AUC (spAUC) was introduced by McClish [8] in order to represent values of partial AUC using values from 0.5 to 1 regardless of the integration region of interest, and is defined as follows:

$$\tilde{A}_e = \frac{1}{2} \left(1 + \frac{\hat{A}(e) - e^2/2}{e - e^2/2} \right) \quad (10)$$

Variance of the estimate of the standardized partial AUC is a rescaled variance of the corresponding partial AUC estimator, namely: $V(\tilde{A}_e) = V(\hat{A})/e^2(4-e)^2$.

In this example our goal was to use the estimated variance for projecting the sample size needed for 80% power of detecting a difference of observed $\tilde{A}_{0.1}$ from the null value h_0 : $\tilde{A}_{0.1} = 0.55$. (i.e., difference of ≈ 0.12). We estimated sample size n_x^*, n_y^* using the following standard formula [10]:

$$n_Y^* = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \times \tilde{V}_\bullet}{\Delta^2} n_X^* = n_Y^* \times k \quad (11)$$

where $k = n_X/n_Y$, $\tilde{V}_\bullet$ is the variance function estimated as $\tilde{V}_\bullet = n_Y \times \hat{V}_\bullet(\tilde{A}_e)$, z_p is the p^{th} percentile of the standard normal distribution, α is the significance level (here $\alpha=0.05$), β is type II error rate (1-power, here $\alpha=0.2$), Δ is the difference between the null and alternative values of the standardized partial AUC (in this example $\Delta=0.12$).

First, we estimated the variance of the empirical standardized pAUC in the original data set using the jackknife variance, $\hat{V}_{J2}$, and the conventional variance, $\hat{V}_{HE}$, by He and Escobar [13]. Then, based on equation (11), we projected the sample sizes for the two approaches. Finally, we used the estimated samples sizes in the bootstrap resampling scheme (based on 10,000 random resamples) to estimate the actual statistical power achieved for the predicted sample sizes.

Table 2 summarizes results for the estimated statistical power corresponding to the sample sizes projected based on the two variance estimators. As expected based on results outlined in previous sections, the conventional variance estimate (for the standardized partial AUC, $\tilde{A}_{0.1}$) has a much lower value (5.98×10^{-4}) than the jackknife variance (10.13×10^{-4}). As a result, the sample size predicted by the conventional estimates (260:98) is much lower than sample size predicted based on the jackknife variance (429:162). The underestimated sample size based on the conventional approach leads to a substantially lower than the planned level of statistical power (58%, compared with the nominal level of 80%), whereas the sample size predicted based on the jackknife variance results in a statistical power of 82%.

5 Discussion

We derived an analytical formula for computing the two-sample jackknife variance of the nonparametric estimator of the partial AUC over the range of high specificity, $fp \in (0, e)$. We demonstrated that the jackknife variance is substantially more accurate than a previously proposed closed-form variance estimator. Namely, the jackknife variance has substantially lower bias and leads to a more appropriate Wald-type test, in particular when the range of interest is narrow.

Clearly, the jackknife variance can be computed by straightforward resampling. However, the formulation we derived has two essential advantages. First, it provides an explicit structure of the variance estimator, thereby providing an insight into the essential components and laying the ground for future improvements of efficient sample size estimation approaches. Second, the analytical expression allows for a substantial reduction in computational burden. For example, in our simulations we observed approximately a 600-fold reduction in computation time, specifically, computing table 1 using resampling would take approximately 300 hours while the use of the analytical expression reduces the time to approximately 30 minutes (using SAS v.9.3 on Intel® Core i5 2.5Ghz, 6Gb).

In this paper we focused on continuous data in order to enable both easier formulation and a direct comparison to the conventional variance estimator [13]. The conventional variance for partial AUC is defined for $\hat{\theta}_e$ estimator of partial AUC, which is reasonable only for continuous data. For discrete data, estimator $\hat{\theta}_e$ (i.e., the area under the empirical ROC curve up to the last empirical ROC point in the range of interest) is not generally consistent and the area under the empirical ROC curve $\hat{A}(e)$ should be used instead. Formulation for the jackknife variance of $\hat{A}(e)$ in the general case can also be derived in closed-form, although the expression is more cumbersome than equation (9).

Properties of the jackknife variance estimator depend on the integration range for partial AUC and sample size. For narrow ranges of integration and/or for small sample sizes both the jackknife variance and the empirical estimator of the partial AUC could be substantially biased. In particular, we do not recommend using either the jackknife variance estimator or the Wald-type tests for very narrow ranges (e.g. 0,0.05) unless the sample size is greater than 100:100. For ranges wider than (0, 0.2) the jackknife variance would provide a good estimate even with sample sizes as low as 50:50.

In this paper we focused on partial AUC over the range (0,e), $A(e)$. This range of integration beginning at $\text{fpf}=0$ is frequently of great interest in clinical practice. The $A(e)$ could be used to compute pAUC over arbitrary ranges (e.g., $A(e_1, e_2) = A(e_2) - A(e_1)$). However, for statistical inferences about pAUC over range (e_1, e_2) , the variance formula should include additional components. Indeed, the variance estimator in (9) accounts only for the uncertainty related to the upper boundary of the range, e_2 , whereas an appropriate variance of $A(e_1, e_2)$ would have to account for uncertainty related to both the lower and upper bounds of the range of integration. As a result, the corresponding closed-form of jackknife variance would be substantially more cumbersome. However, simulation results based on the resampling-based computation of jackknife variance (omitted for brevity) indicate that the jackknife variance maintains similar advantages over the conventional variance estimator for $A(e_1, e_2)$ as well.

In summary, the development of an analytical formulation for the jackknife variance of the nonparametric estimator of the partial AUC enables an efficient computation and provides important insights into the structure of the variance components. In addition, our investigation highlights an important deficiency in the current methodology and resolves current problems with the conventional variance estimator by offering a closed-form estimator with smaller biases and better asymptotic properties.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number GM098253.

References

1. Quenouille MH. Notes on bias in estimation. *Biometrika*. 1956; 43:353–360.
2. Tukey JW. Bias and confidence in not-quite large samples. *Ann Math Stat*. 1958; 29:614.
3. Efron B, Stein C. The jackknife estimate of variance. *Ann Stat*. 1981; 9:586–596.
4. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
5. Arvesen JN. Jackknifing U-statistics. *Ann Math Statist*. 1969; 40:2076–2100.
6. Bandos, AI., Rockette, HE., Gur, D. Resampling methods for the area under the ROC curve. Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning; Pittsburgh, USA. 2006. p. 1-8.
7. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data. *Stat Med*. 2005; 24:1579–1607. [PubMed: 15685718]
8. McClish D. Analyzing a portion of the ROC curve. *Med Decis Making*. 1989; 9:190–195. [PubMed: 2668680]
9. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*. 1996; 201:745–750. [PubMed: 8939225]
10. Zhou, XH., Obuchowski, NA., McClish, DK. Statistical methods in diagnostic medicine. 2nd. New York: Wiley & Sons Inc; 2011.
11. Wieand HS, Gail M, James B, et al. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989; 76:585–592.
12. Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics*. 2003; 59:614–623. [PubMed: 14601762]
13. He Y, Escobar M. Nonparametric statistical inference method for paired areas under receiver operating characteristic curves, with application to genomic studies. *Stat Med*. 2008; 27:5991–5308. [PubMed: 18792086]
14. Ma H, Bandos AI, Rockette HE, et al. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat Med*. 2013; 32:3449–3458. [PubMed: 23508757]
15. Egan, JP. Signal detection theory and ROC analysis. New York: Academic Press; 1975.
16. Herron JM, Bender TM, Campbell WL, et al. Effects of luminance and resolution on observer performance with chest radiographs. *Radiology*. 2000; 215:169–174. [PubMed: 10751483]
17. Liu A, Schisterman EF, Wu C. Nonparametric estimation and hypothesis testing on the partial area under receiver operating characteristic curves. *Communications in Statistics- Theory and Methods*. 2005; 34:2077–2088.

Appendix: detailed derivation of the jackknife variance

Here we provide additional details on the derivations of equations (3) to (9). Exclusion of the k^{th} low-scored non-diseased subject ($k'=1, \dots, n_x-t$), immediately leads to equation (3). Exclusion of the k^{th} high-scored non-diseased subject ($k=n_x-t+1, \dots, n_x$) leads to equation (4) via the following algebraic manipulations:

$$\begin{aligned}\hat{\theta}_{(-k, \bullet)} &= \frac{\sum_{\substack{i=n_x-t, \\ i \neq k}}^{n_x} \sum_{j=1}^{n_y} I(\chi_i < y_j)}{(n_x - 1) n_y} = \frac{1}{(n_x - 1) n_y} \left[n_x n_y \hat{\theta} + \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) - \sum_{j=1}^{n_y} I(\chi_k < y_j) \right] \\ &= \frac{n_x}{n_x - 1} \hat{\theta} + \frac{1}{(n_x - 1) n_y} \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) - \frac{1}{(n_x - 1) n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j)\end{aligned}$$

Exclusion of the l^{th} diseased subject ($l=1, \dots, n_y$) leads to equation (5) because of the following:

$$\begin{aligned}\hat{\theta}_{(\bullet, -l)} &= \frac{1}{n_x(n_y - 1)} \sum_{i=n_x-t+1}^{n_x} \sum_{\substack{j=1, \\ j \neq l}}^{n_y} I(\chi_i < y_j) = \frac{1}{n_x(n_y - 1)} \left[n_x n_y \hat{\theta} - \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l) \right] \\ &= \frac{n_y}{n_y - 1} \hat{\theta} - \frac{1}{n_x(n_y - 1)} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l)\end{aligned}$$

The average corresponding to non-diseased subjects in equation (6) is obtained as follows:

$$\begin{aligned}\hat{\theta}_{(-\bullet, \bullet)} &= \frac{\sum_{k=1}^{n_x} \hat{\theta}_{(-k, \bullet)}}{n_x} = \frac{\sum_{k'=1}^{n_x-t} \hat{\theta}_{(-k', \bullet)}}{n_x} + \frac{\sum_{k=n_x-t+1}^{n_x} \hat{\theta}_{(-k, \bullet)}}{n_x} \\ &= \frac{n_x - t}{n_x - 1} \hat{\theta} + \frac{t}{n_x - 1} \hat{\theta} + \frac{t \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j)}{n_x(n_x - 1) n_y} - \frac{\sum_{k=n_x-t+1}^{n_x} \sum_{j=1}^{n_y} I(\chi_k < y_j)}{n_x(n_x - 1) n_y} \\ &= \frac{n_x}{n_x - 1} \hat{\theta} + \frac{t \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j)}{n_x(n_x - 1) n_y} - \frac{1}{n_x - 1} \hat{\theta} + \frac{t}{(n_x - 1) n_x} \times \frac{\sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j)}{n_y}\end{aligned}$$

Equation (7) for the average corresponding to diseased subjects immediately follows from equations (5) and (1).

Expression for the overall variance (9) is obtained from equations (3) to (7) combined according to equation (8) and transformed using the following steps:

$$\begin{aligned}
\hat{V}_{J_2}(\hat{\theta}) &= \frac{n_x - 1}{n_x} \sum_{k=1}^{n_x} \left(\hat{\theta}_{(-k, \bullet)} - \hat{\theta}_{(-\bullet, \bullet)} \right)^2 + \frac{n_y - 1}{n_y} \sum_{l=1}^{n_y} \left(\hat{\theta}_{(\bullet, -l)} - \hat{\theta}_{(\bullet, -\bullet)} \right)^2 = [\text{using (3) to (7)}] \\
&= \frac{n_x - 1}{n_x} \sum_{k=1}^{n_x-t} \left(\frac{n_x}{n_x - 1} \hat{\theta} - \hat{\theta} - \frac{t}{(n_x - 1)n_x} \times \frac{1}{n_y} \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right)^2 \\
&\quad + \frac{n_x - 1}{n_x} \sum_{k=n_x-t+1}^{n_x} \left(\frac{n_x}{n_x - 1} \hat{\theta} + \frac{1}{(n_x - 1)n_y} \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) - \frac{1}{(n_x - 1)n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j) - \hat{\theta} - \frac{t}{(n_x - 1)n_x n_y} \times \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right. \\
&\quad \left. + \frac{n_y - 1}{n_y} \sum_{l=1}^{n_y} \left(\frac{n_y}{n_y - 1} \hat{\theta} - \frac{1}{n_x(n_y - 1)} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l) - \hat{\theta} \right)^2 \right) \\
&= \frac{n_x - t}{(n_x - 1)n_x} \left(\hat{\theta} - \frac{t}{n_y n_x} \times \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right)^2 + \frac{1}{(n_x - 1)n_x} \sum_{k=n_x-t+1}^{n_x} \left\{ \hat{\theta} - \frac{1}{n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j) + \frac{n_x - t}{n_y n_x} \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right. \\
&\quad \left. + \frac{1}{(n_y - 1)n_y} \sum_{l=1}^{n_y} \left(\hat{\theta} - \frac{1}{n_x} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l) \right)^2 \right\} \\
&= \frac{n_x - t}{(n_x - 1)n_x} \left(\hat{\theta} - \frac{t}{n_y n_x} \times \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right)^2 \\
&\quad + \frac{1}{(n_x - 1)n_x} \sum_{k=n_x-t+1}^{n_x} \left\{ \frac{n_x}{t} \hat{\theta} - \frac{1}{n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j) - \frac{n_x - t}{t} \hat{\theta} + \frac{n_x - t}{n_y n_x} \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right\}^2 \\
&\quad + \frac{1}{(n_y - 1)n_y} \sum_{l=1}^{n_y} \left(\hat{\theta} - \frac{1}{n_x} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l) \right)^2 \\
&= \frac{n_x - t}{(n_x - 1)n_x} \left(\hat{\theta} - \frac{t}{n_y n_x} \times \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right)^2 + \frac{(n_x - t)^2}{t(n_x - 1)n_x} \left(\hat{\theta} - \frac{t}{n_y n_x} \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right)^2 \\
&\quad + \frac{1}{(n_x - 1)n_x} \sum_{k=n_x-t+1}^{n_x} \left\{ \frac{n_x}{t} \hat{\theta} - \frac{1}{n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j) \right\}^2 + \frac{1}{(n_y - 1)n_y} \sum_{l=1}^{n_y} \left(\hat{\theta} - \frac{1}{n_x} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l) \right)^2 \\
&= \frac{n_x - t}{(n_x - 1)t} \left(\hat{\theta} - \frac{t}{n_y n_x} \times \sum_{j=1}^{n_y} I(\chi_{n_x-t} < y_j) \right)^2 \\
&\quad + \frac{1}{(n_x - 1)n_x} \sum_{k=n_x-t+1}^{n_x} \left\{ \frac{n_x}{t} \hat{\theta} - \frac{1}{n_y} \sum_{j=1}^{n_y} I(\chi_k < y_j) \right\}^2 + \frac{1}{(n_y - 1)n_y} \sum_{l=1}^{n_y} \left(\hat{\theta} - \frac{1}{n_x} \sum_{i=n_x-t+1}^{n_x} I(\chi_i < y_l) \right)^2
\end{aligned}$$

Formulation (9) follows immediately from the last expression above.

Figure 1.
Binormal ROC curves with frequently encountered parameter values

Figure 2. Estimated relative bias (%) of the variance estimators as a function of sample size (estimates are based on 10,000 simulations).

Figure 3.
Empirical ROC curve for the data used in the example (full emp. AUC=0.76)

Table 1

Relative bias of the variance estimators of the nonparametric pAUC estimator and the type I error rate for the corresponding Wald-type tests (10,000 simulations).

Sample size (n_x, n_y)	ROC curve parameters		Range 0–0.05				Range 0–0.1				Range 0–0.2				Range 0–0.5					
			Relative Bias		Type I error rate		Relative Bias		Type I error rate		Relative Bias		Type I error rate		Relative Bias		Type I error rate			
	b	a	AUC*	J2	He	J2	He	J2	He	J2	He	J2	He	J2	He	J2	He	J2	He	
500:500	1	0.4	0.6	0.05	−0.43	0.06	0.14	0.02	−0.41	0.06	0.13	0.02	−0.37	0.05	0.12	0.02	−0.28	0.05	0.10	
		0.7	0.7	0.05	−0.47	0.06	0.16	0.03	−0.43	0.05	0.14	0.03	−0.38	0.05	0.12	0.01	−0.26	0.05	0.09	
		1.2	0.8	0.04	−0.49	0.06	0.17	0.02	−0.44	0.05	0.15	0.00	−0.38	0.05	0.13	0.00	−0.23	0.05	0.09	
		1.8	0.9	0.02	−0.48	0.06	0.17	0.00	−0.42	0.05	0.14	−0.01	−0.34	0.05	0.11	0.00	−0.17	0.05	0.07	
		0.5	0.3	0.6	0.02	−0.27	0.05	0.10	0.01	−0.21	0.05	0.09	0.00	−0.16	0.05	0.07	−0.01	0.05	0.07	
	0.5	0.6	0.7	0.01	−0.29	0.05	0.10	0.00	−0.23	0.05	0.09	−0.01	−0.17	0.05	0.08	0.00	−0.10	0.05	0.06	
		0.9	0.8	0.02	−0.28	0.05	0.10	0.01	−0.22	0.05	0.09	0.02	−0.15	0.05	0.07	0.02	−0.07	0.05	0.06	
		1.4	0.9	−0.01	−0.28	0.06	0.10	−0.01	−0.21	0.05	0.08	−0.01	−0.14	0.05	0.07	0.00	−0.06	0.05	0.06	
		1	0.4	0.6	0.23	−0.36	0.07	0.15	0.13	−0.37	0.06	0.14	0.07	−0.34	0.06	0.12	0.04	−0.27	0.05	0.10
		0.7	0.7	0.18	−0.42	0.08	0.16	0.09	−0.41	0.06	0.15	0.03	−0.38	0.06	0.13	0.01	−0.26	0.05	0.10	
100:100	1	1.2	0.8	0.16	−0.44	0.08	0.18	0.10	−0.41	0.06	0.15	0.04	−0.35	0.06	0.13	0.03	−0.21	0.05	0.09	
		1.8	0.9	0.12	−0.42	0.10	0.19	0.06	−0.39	0.07	0.15	0.01	−0.32	0.06	0.12	0.01	−0.16	0.05	0.08	
		0.5	0.3	0.6	0.10	−0.22	0.06	0.09	0.07	−0.18	0.05	0.08	0.04	−0.13	0.05	0.07	0.03	−0.08	0.05	0.06
		0.6	0.7	0.09	−0.25	0.06	0.11	0.05	−0.20	0.05	0.09	0.02	−0.15	0.05	0.08	0.00	−0.10	0.05	0.07	
		0.9	0.8	0.07	−0.25	0.07	0.11	0.04	−0.20	0.06	0.09	0.02	−0.15	0.05	0.07	0.02	−0.08	0.05	0.06	
	0.5	1.4	0.9	0.06	−0.23	0.07	0.11	0.03	−0.18	0.06	0.08	0.01	−0.13	0.05	0.07	0.01	−0.06	0.05	0.06	
		1	0.4	0.6	0.54	−0.25	0.14	0.21	0.22	−0.33	0.07	0.14	0.09	−0.34	0.07	0.13	−0.01	−0.29	0.06	0.12
		0.7	0.7	0.39	−0.34	0.14	0.27	0.17	−0.37	0.07	0.15	0.08	−0.35	0.06	0.13	0.02	−0.25	0.06	0.10	
		1.2	0.8	0.30	−0.38	0.15	0.30	0.13	−0.38	0.08	0.16	0.06	−0.34	0.07	0.13	0.03	−0.20	0.05	0.09	
		1.8	0.9	0.27	−0.37	0.15	0.32	0.11	−0.35	0.09	0.16	0.05	−0.28	0.07	0.12	0.02	−0.15	0.06	0.08	
50:50	0.5	0.3	0.6	0.20	−0.19	0.15	0.25	0.09	−0.16	0.06	0.09	0.04	−0.14	0.06	0.08	0.01	−0.09	0.06	0.07	
		0.6	0.7	0.23	−0.18	0.17	0.27	0.11	−0.15	0.06	0.09	0.06	−0.11	0.05	0.08	0.04	−0.05	0.05	0.06	
		0.9	0.8	0.15	−0.23	0.23	0.36	0.04	−0.20	0.07	0.09	0.00	−0.16	0.06	0.08	−0.02	−0.09	0.06	0.07	
		1.4	0.9	0.15	−0.20	0.46	0.60	0.06	−0.16	0.06	0.09	0.03	−0.11	0.05	0.07	0.00	−0.06	0.06	0.07	

$$AUC = \frac{1}{2} \left(1 + \frac{AUC}{\sqrt{1 + b^2}} \right)$$

For binormal ROC curves with parameters a and b

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Estimated variance of standardized pAUC (from the original sample of 223:94 subjects), predicted sample sizes, and the estimated statistical power achieved with the predicted sample sizes.

Estimated quantities	H&E	J2
$\hat{V}_*(\tilde{A}_{0.1})$	0.000598	0.001013
n^*_X	260	429
n^*_Y	98	162
<i>Actual power</i>	57.94%	81.97%