PERMUTATION TESTS AND CONFIDENCE INTERVALS FOR THE AREA UNDER THE ROC-CURVE

MARKUS PAULY, THOMAS ASENDORF, AND FRANK KONIETSCHKE

Abstract. We investigate rank-based studentized permutation methods for the nonparametric Behrens-Fisher problem, i.e. inference methods for the area under the ROC-curve (AUC). We hereby prove that the studentized permutation distribution of the Brunner-Munzel rank statistic is asymptotically standard normal, even under the alternative. This does not only imply consistency of the corresponding permutation test, but also that confidence intervals for the underlying treatment effects can be computed. The result further implies that the Neubert and Brunner studentized permutation test can be inverted for the computation of confidence intervals. In addition, we derive permutation-based range-preserving confidence intervals. Extensive simulation studies show that the permutation based confidence intervals appear to maintain the pre-assigned coverage probability quite accurately (even for rather small sample sizes). For a convenient application of the proposed methods, a freely available software package for the statistical software R has been developed. A real data example illustrates the application.

1. Introduction

The aim of a diagnostic test is the investigation of the ability to distinguish between diseased and non-diseased subjects (ensured by a gold standard) of a certain diagnostic modality. Hereby its accuracy can be quantified by various measures. In case of binary end-points, e.g. pregnant or non-pregnant, the sensitivity and the specificity are usually assessed. They are defined as the probabilities of the test correctly identifying the diseased and non-diseased subjects, respectively. When the end-points of the diagnostic test are measured on a metric or an ordinal scale, a cut-off value k has to be chosen in order to compute sensitivity and specificity. Each decision limit k may yield a different 2×2-table of dichotomized test results versus the true disease status. Thus, sensitivity and specificity can be estimated from each decision limit k. However, as k decreases, sensitivity increases while specificity decreases, and vice versa. Thus, there is a trade-off between sensitivity and specificity as the decision limit varies, see e.g. Kaufmann et al. (2005). A summary measure of this discriminatory accuracy is the so-called *receiver operating*

Date: April 17, 2014.
2000 *Mathematics Subject Classification.* 62G05, 62G09.
Key words and phrases. Confidence Intervals; Permutation Tests; ROC-Curve; Studentized statistics.

characteristic (ROC)-curve, which is a plot of the sensitivity versus 1− specificity for varying thresholds k. The upper left corner of the graph represents perfect discrimination, while the diagonal line represents a discrimination which is not better than chance. Kaufmann et al. (2005) point out that the ROC curve of a diagnostic test is invariant with respect to any monotone transformation of the test measurement scale for both diseased and non-diseased subjects, while being independent of the prevalence of disease in the sample. Therefore, the ROC-curve is an adequate measure for comparing diagnostic tests on different scales. In particular, the *area under the ROC-curve* (AUC) represents an accuracy measure, which is independent from the chosen cut-off value, and which is invariant under any monotone transformation of the data. If the AUC is 1, then the diagnostic test will be referred to be a *perfect* test, while an AUC of 1/2 represents a diagnostic test being as good as chance. Figure 1 presents the corresponding ROC-curves for a perfect, imperfect and an realistic diagnostic test. Therefore, *no diagnostic accuracy* can be expressed as $AUC = 1/2$. Lange and Brunner (2012) have further shown that the analysis of sensitivity, specificity and the AUC can be unified, i.e. sensitivity and specificity are areas under certain ROC-curves. In diagnostic trials, particularly in imaging

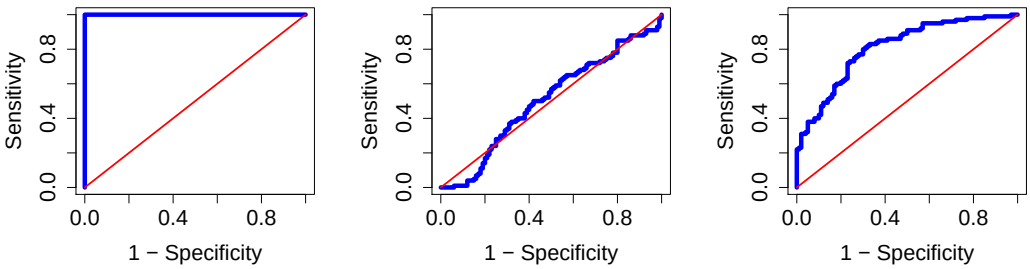

FIGURE 1. ROC-curves and their AUC's: Perfect diagnose (left; AUC=1), Imperfect diagnose (middle; AUC=1/2), Suitable diagnose (right; AUC=0.8)

studies, ordered categorical data scores are usually used to assess the severity of a disease. Therefore, parametric approaches in terms of mean based procedures are inadequate for testing the null hypothesis

$$H_0 : AUC = 1/2.$$

Brunner and Munzel (2000), Kaufmann et al. (2005) and Neubert and Brunner (2007) propose rank methods for the analysis of diagnostic tests. In particular, Neubert and Brunner (2007) propose a studentized permutation test in the Brunner-Munzel rank statistic for H_0 and have heuristically shown that their test is asymptotically valid under the null, i.e. keeps the type-I-error level for large sample sizes. In this paper we give a rigorous proof of this result and also analyze the behaviour of

their studentized permutation test under the alternative. To this end we prove that the permutation distribution of the Brunner-Munzel rank statistic is asymptotically standard normal for any underlying value of the AUC, i.e. not only under H_0 but also for $AUC \neq 1/2$. This is not only important for deducing the consistency of their test but it has the additional advantage that this permutation technique can also be applied for constructing adequate confidence intervals.

Note, that nonparametric ranking procedures are invariant under any monotone transformation of the data and are preferred for making statistical inferences. In diagnostics, however, p-value based approaches are not sufficient for the analysis of diagnostic tests. We particularly favor confidence intervals since they allow both a proof of hazard and a proof of safety, and we focus on the interpretation of an effect size instead of a probability (p-value), according to the EMEA recommendation for the evaluation of diagnostic agents:"*... Ideally, the impact on diagnostic thinking should be presented numerically; the rate of cases where diagnostic uncertainty with a new agent has been decreased (as compared to pre-test diagnosis established by mean of a conventional work including the comparator) should be analysed and reported (percentage, and confidence intervals)...*" (EMEA 2008, chapter 5.2.3, p. 13).
Brunner and Munzel (2000) propose asymptotic as well as approximate confidence intervals for the AUC. Kaufmann et al. (2005) state that these confidence intervals may not be range preserving, i.e., the lower / upper bounds may be smaller than 0/1, respectively. Range preserving confidence intervals are achieved by using the delta method (Kaufmann et al., 2005). Simulation studies indicate, however, that all of these procedures tend to be quite liberal or conservative when the true AUC is large (e.g. $AUC \geq 0.7$) and sample sizes are rather small. Small sample sizes, however, occur frequently in practice, e.g. in cancer diagnostic trials. It is the aim of the present paper to improve the existing procedures for small sample sizes. Hereby, we propose an unified studentized permutation approach by investigating the conditional studentized permutation distribution of Brunner and Munzel's (2000) linear rank statistic for the nonparametric Behrens-Fisher problem. Janssen (1997, 2005) proposes studentized permutation tests for the parametric Behrens-Fisher problem, whereas Fay and Proschan (2010) discuss test recommendations in both Behrens-Fisher problems. Moreover, Janssen (1999), Pauly (2011b), Konietschke and Pauly (2012, 2014), Omelka and Pauly (2012) as well as Pauly et al. (2014) support the use of studentized permutation tests for other testing problems.
Permutation and randomization based procedures for comparing ROC curves or for testing $H_0 : AUC = 1/2$ were proposed by Venkatraman and Begg (1996), Venkatraman (2000), Bandos et al. (2005, 2006), Neubert and Brunner (2007) as well as Braun and Alonzo (2008). Recently, Jin and Lu (2009) have also applied a permutation test based on the Mann-Whitney statistic estimate of the AUC to test for non-inferiority. More details about permutation and randomization tests can be found in the monographs of Basso et al. (2009), Good (2005) as well as Pesarin and

Salmaso (2010).
Note that all of the above mentioned randomization tests for the ROC curve or AUC are only reasoned by extensive simulation studies, heuristics and/or under invariance properties (as exchangeability) of the data. In this paper we will extend these results to non-exchangeable data. In particular, we prove that rank-based tests and confidence intervals derived from studentized permutation statistics are even asymptotically exact if the data is not exchangeable (e.g. heteroscedastic) and retain the finite exactness property under exchangeability. These nice features will then be demonstrated by simulation studies, where we will see that our new approach is more accurate than its competitors. Finally we illustrate the procedures using a real data set from a sonography ultra-sound imaging study.
The paper is organized as follows. In Section 2 we explain the underlying model and introduce the AUC. The current state of the art for constructing tests and confidence intervals for this quantity is picked up in Section 3, whereas our improved proposal based on permutation is considered in Section 4. An extensive Simulation study as well as the practical data example is presented in Sections 5 and 6. Finally all proofs are given in the Appendix.

2. Statistical Model

We consider a diagnostic trial involving N independent subjects in total, which may be partitioned into two groups containing n_1 non-diseased subjects in group 1 and n_2 diseased subjects in group 2, respectively, as classified by a gold-standard. Let X_{ij} denote the jth replicate in the ith group, $i = 1, 2$. Without loss of generality we assume that lower values of the outcome are associated with non-diseased subjects. To allow for continuous and discontinuous data in a unified way, let $F_i(x) = P(X_{i1} < x) + 1/2P(X_{i1} = x)$ denote the normalized version of the distribution function, which is the average of the right continuous version $F_i^+(x) = P(X_{i1} \leq x)$ and the left continuous version $F_i^-(x) = P(X_{i1} < x)$ of the distribution function, respectively. In the context of nonparametric models, the normalized version of the distribution function $F_i(x)$ was first mentioned by Lévy (1925). Later on, it was used by Ruymgaart (1980), Akritas, Arnold and Brunner (1997), Munzel (1999), Gao et al. (2008), among others, to derive asymptotic results for rank statistics including the case of ties in a unified way. We note that the F_i may be arbitrary distributions, with the exception of the trivial case of one-point distributions. The general model specifies only that

$$X_{ij} \sim F_i, \qquad i = 1, 2;\ j = 1, .., n_i, \tag{2.1}$$

and does not require that the distributions are related in any parametric way.
The distributions F_1 and F_2 can now be used for the definition of accuracy measures. The sensitivity of the diagnostic test is defined by $SE(k) = 1 - F_2(k)$, which is the probability of correctly classifying a diseased subject at a certain threshold value k. Similarly, the specificity of the diagnostic agent is defined by $SP(k) = F_1(k)$, i.e. the probability of correctly classifying a non-diseased subject. Finally, the ROC

curve is obtained as the plot of the sensitivity $SE(k) = 1 - F_2(k)$ on the vertical axis versus $1 - SP(k) = 1 - F_1(k)$ on the horizontal axis as the cut-off value k varies from $-\infty$ to ∞ (see Figure 1). Therefore, the area under the ROC-curve (AUC) is given by

$$p = AUC = \int_{-\infty}^{\infty} F_1(k) dF_2(k) = P(X_{11} < X_{21}) + \frac{1}{2} P(X_{11} = X_{21}), \tag{2.2}$$

which is also known as the *relative effect* in the literature (see, e.g. Brunner and Munzel 2000; Neubert and Brunner 2007). For the special case of ordinal data, p is also known as *ordinal effect size measure* (Ryu 2009; Ryu and Agresti 2008). Note that p is also estimated by the Mann-Whitney statistic. The intuitive interpretation of the AUC is as follows: if the observations coming from $F_1(x)$, i.e. from the non-diseased subjects, tend to be smaller than those coming from $F_2(x)$, i.e. from the diseased subjects, then $p > 1/2$. This means that this probability measures a separation of the two populations of diseased and non-diseased subjects: the larger the deviation from 1/2 (which means no discrimination), the larger is the separation of the two populations. Therefore the AUC is used as a summary index for the accuracy of a diagnostic test, see, e.g., Kaufmann et al. (2005) for further details.

2.1. **Point estimators and their limiting distribution.** Rank estimators of the AUC p defined in (2.2) are derived by replacing the unknown distribution functions $F_1(x)$ and $F_2(x)$ by their empirical counterparts

$$\widehat{F}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} h(x - X_{ij}), \tag{2.3}$$

where $h(x) = 0, \frac{1}{2}, 1$ according as $x < 0, x = 0, x > 0$, respectively (Ruymgaart 1980). The unbiased estimator

$$\widehat{p} = \int \widehat{F}_1 d\widehat{F}_2 = \frac{1}{n_1} \left(\overline{R}_{2\cdot} - \frac{n_2 + 1}{2} \right) \tag{2.4}$$

can be easily computed with the ranks R_{ij} of X_{ij} among all N observations. Here, $\overline{R}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} R_{ij}$ denotes the mean of the ranks in group i, $i = 1, 2$. Brunner and Munzel (2000) show that the standardized statistic $\sqrt{N}(\hat{p} - p)/\sigma_N$ follows, asymptotically, as $\min(n_1, n_2) \to \infty$, a standard normal distribution, where

$$\sigma_N^2 = \frac{N}{n_1 n_2} (n_1 \sigma_2^2 + n_2 \sigma_1^2) \tag{2.5}$$

with $\sigma_1^2 = Var(F_2(X_{11}))$ and $\sigma_2^2 = Var(F_1(X_{21}))$. Here and throughout the paper we assume that $\sigma_N^2 > 0$ holds. The unknown variance can be estimated by

$$\hat{\sigma}_N^2 = \frac{N}{n_1 n_2} (n_1 \widehat{\sigma}_2^2 + n_2 \widehat{\sigma}_1^2), \tag{2.6}$$

where $\widehat{\sigma}_i^2 = S_i^2/(N-n_i)^2$ and

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(R_{ij} - R_{ij}^{(i)} - \overline{R}_{i\cdot} + \frac{n_i + 1}{2} \right)^2 . \tag{2.7}$$

Here, $R_{ij}^{(i)}$ denotes the rank of X_{ij} among all n_i observations in group i, $i = 1, 2$. The asymptotic distribution of $\sqrt{N}(\widehat{p} - p)$ can now be used for the derivation of confidence intervals for the AUC p. This will be explained in the next section.

3. Tests and confidence intervals for the AUC

3.1. **Asymptotic procedures.** Based on the asymptotic normality of $\sqrt{N}(\widehat{p} - p)$ it follows by Slutsky's theorem that

$$T_{id} := \frac{\sqrt{N}(\hat{p} - p)}{\hat{\sigma}_N} \dot{\sim} N(0, 1). \tag{3.1}$$

Hence a two-sided asymptotical level α test for H_0 is given by $\varphi = \mathbf{1}\{|T_{id}| \geq z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile from the standard normal distribution. Moreover, by using the pivotal method, asymptotic $(1-\alpha)$-confidence intervals for p are given by

$$CI_{id} = \left[\widehat{p} \pm z_{1-\alpha/2} / \sqrt{N} \widehat{\sigma}_N \right]. \tag{3.2}$$

However, simulation studies indicate that these tests and confidence intervals tend to be very liberal, even with moderate sample sizes $n_i \equiv 20$. Small sample approximations of the distribution of T_{id} are discussed below.

3.2. **Small sample approximations.** Motivated by the Satterthwaite-Welch t-Test (Welch 1947), Brunner and Munzel (2000) propose to approximate the distribution of T_{id} as given in (3.1) by a central $t_{\hat{f}}$-distribution with

$$\hat{f} = \frac{\left(\sum_{i=1}^{2} S_i^2/(N-n_i) \right)^2}{\sum_{i=1}^{2} \left(S_i^2/(N-n_i) \right)^2 / (n_i - 1)} \tag{3.3}$$

degrees of freedom. Thus, approximate $(1-\alpha)$-confidence intervals for p are given by

$$CI_t = \left[\widehat{p} \pm t_{1-\alpha/2,\hat{f}} / \sqrt{N} \widehat{\sigma}_N \right], \tag{3.4}$$

where $t_{1-\alpha/2,\hat{f}}$ denotes the $(1-\alpha/2)$-quantile from the central $t_{\hat{f}}$-distribution with $\hat{f}$ degrees of freedom as given in (3.3). Similarly, the Brunner and Munzel (2000) test is obtained by substituting $z_{1-\alpha/2}$ in φ with $t_{1-\alpha/2,\hat{f}}$.
Moreover, we note that both the confidence intervals CI_{id} and CI_t defined in (3.2) and (3.4) may not be range preserving. Kaufmann et al. (2005) therefore propose

range preserving confidence intervals. They are calculated by using the so called δ-method with a differentiable function $g : (0,1) \to (-\infty, \infty)$ by

$$T_g := \frac{\sqrt{N}\,(g(\hat{p}) - g(p))}{g'(\hat{p})\hat{\sigma}_N} \dot{\sim} N(0,1), \tag{3.5}$$

where $g'(x)$ denotes the first derivative of $g(x)$ which is assumed to be non-zero valued around p. In particular we will be using the *logit* and *probit* transformation, which are defined as $g(x) = \text{logit}(x) = \log(x/(1-x))$ with $g'(t) = 1/(x - x^2)$ and $g(x) = \text{probit}(x) = \Phi^{-1}(x)$ with $g'(t) = 1/\varphi(\Phi^{-1}(t))$, where Φ denotes the distribution function and φ the density of the standard normal distribution. The inverse are given by $\text{logit}^{-1}(x) = \exp(x)/(1+\exp(x))$ and $\text{probit}^{-1}(x) = \Phi(x)$. Hence, range preserving $(1-\alpha)$-confidence intervals are given by

- $g = \text{logit}$:

$$CI_{logit} = \left[\text{logit}^{-1}\left(\text{logit}(\hat{p}) \pm \frac{z_{1-\alpha/2}\hat{\sigma}_N}{\hat{p}(1-\hat{p})\sqrt{N}}\right)\right] \tag{3.6}$$

- $g = \text{probit}$:

$$CI_{probit} = \left[\Phi\left(\Phi^{-1}(\hat{p}) \pm \frac{z_{1-\alpha/2}\sqrt{2\pi}\hat{\sigma}_N}{\exp\{-1/2(\Phi^{-1}(\hat{p}))^2\}}\right)\right]. \tag{3.7}$$

The coverage probabilities of both the confidence intervals CI_t in (3.4) as well as the range preserving confidence intervals CI_g in (3.6) and (3.7), respectively, can be improved by not using the critical values $z_{1-\alpha/2}$ or $t_{1-\alpha/2,\hat{f}}$, but by estimating the critical values from the so-called studentized permutation distributions of T_{id} in (3.1) and T_g in (3.5). This will be explained in the next section.

4. A studentized permutation approach

As already mentioned above Brunner and Munzel (2000) have suggested to approximate the distribution of T_{id} by a student-t distribution with according degrees of freedom. With that choice the accuracy of the corresponding confidence intervals (as well as the exactness of the corresponding tests) increases with small sample sizes, see Section 5 below. However, for unbalanced designs and small sample sizes its accuracy is still not satisfactorily. Therefore we propose the usage of a data-dependent studentized permutation method. Let us shortly recall the main idea, where the case $g = id$ corresponds to the test given in Neubert and Brunner (2007). We pool the data of both groups, randomly permute it and calculate the empirical effect, say $\widehat{p}^{\tau}$, of the permuted data. Heuristically each permuted observations should have the same effect $p = 1/2$ and the corresponding permuted test statistics, say T_g^{τ}, may be used to approximate the finite sample distribution of T_g and to construct corresponding confidence intervals. Below we will characterize this procedure

in detail and prove that it leads to asymptotic exact confidence intervals.
To describe our approach theoretically set

$$\mathbf{X} = (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) =: (Y_1, \ldots, Y_N)$$

for the pooled data and let τ be a random variable that is uniformly distributed on the symmetric group $\mathcal{S}_N$, i.e. the set of all permutation of the numbers $1, \ldots, N$, being independent from the data $\mathbf{X}$. For given $\mathbf{X}$ we then define the permuted pooled sample as $\mathbf{X}^\tau = (Y_{\tau(1)}, \ldots, Y_{\tau(N)})$ and calculate

$$T_{id}^\tau = \frac{\sqrt{N}(\widehat{p}^\tau - \frac{1}{2})}{\widehat{\sigma}_N^\tau}, \tag{4.1}$$

where $\widehat{p}^\tau = \widehat{p}(\mathbf{X}^\tau)$ and $\widehat{\sigma}_N^\tau = \widehat{\sigma}_N(\mathbf{X}^\tau)$ are the estimated treatment effect and the variance estimator (2.6) calculated from the permuted sample $\mathbf{X}^\tau$ instead of the original pooled sample $\mathbf{X}$. In practice this means that we randomly permute the pooled data and determine their (new) ranks to calculate the estimators $\widehat{p}^\tau$ and $\widehat{\sigma}^\tau$ with the help of their rank expressions (2.4) - (2.7). Note, that the observed data is permuted both between, as well as within the groups, making it possible that data which was observed in the first group is permuted to be in the second group and vice versa.
Our main result now states, that the conditional distribution of T_{id}^τ given the data $\mathbf{X}$ always (i.e. for any underlying AUC) approximates the null distribution of T_{id}.

Theorem 4.1. *Let T_{id}^τ as given in (4.1) and assume that $n_1/N \to \kappa \in (0,1)$. Then we have convergence for arbitrary p*

$$\sup_{x \in \mathbb{R}} \mid P(T_{id}^\tau \leq x \mid \mathbf{X}) - P_{H_0}(T_{id} \leq x) \mid \to 0 \quad \textit{in probability, as } N \to \infty. \tag{4.2}$$

Moreover, the corresponding quantiles converges as well, i.e. if z_α^τ denotes the α-quantile of the conditional distribution of T_{id}^τ we have

$$z_\alpha^\tau \to u_\alpha \quad \textit{in probability as } \min(n_1, n_2) \to \infty$$

for all $\alpha \in (0,1)$.

We note that this result proves that the Neubert and Brunner (2007) permutation test $\varphi^\tau = \mathbf{1}\{|T_{id}| \geq z_{1-\alpha/2}^\tau\}$ is i.) asymptotically exact under H_0 and ii.) consistent, i.e. it has asymptotic power of 1 for an arbitrary, but fixed $AUC \neq 1/2$.

The above theorem also states that the limiting distribution of T_{id}^τ is standard normal and does not depend on the distribution of the data. In particular, it is achieved for any underlying treatment effect p. Moreover, another application of the delta-method for a differentiable function g as above proves that the conditional distribution of

$$T_g^\tau = \frac{\sqrt{N}(g(\widehat{p}^\tau) - g(\frac{1}{2}))}{g'(\widehat{p}^\tau)\widehat{\sigma}_N^\tau}, \tag{4.3}$$

is asymptotically standard normal, i.e. (4.2) holds with T_{id}^τ replaced by T_g^τ.

Hence we can use the data-dependent $(1-\alpha/2)$ quantile $z^{\tau}_{1-\alpha/2}$ of the conditional permutation distribution of T^{τ}_g given the data $\mathbf{X}$ to calculate confidence intervals. The theorem then guarantees that these permutation based confidence intervals are asymptotically of level α.
The numerical algorithm for the computation of the permutation based confidence intervals as well as the p-value for $H_0 : p = 1/2$ is as follows:

(1) Given the data $\mathbf{X}$, compute $\widehat{p}$, $\widehat{\sigma}_N$ and T_g as in (3.5).
(2) For a sufficiently large number of random permutations n_{perm} (e.g. $n_{perm} = 10,000$), permute the data randomly, compute T^{τ}_g given in (4.3) for each permutation and save these values in $A_1, \ldots, A_{n_{perm}}$.
(3) Estimate $z^{\tau}_{1-\alpha/2}$ and $z^{\tau}_{\alpha/2}$ by the empirical $(1-\alpha/2)$- and $(\alpha/2)$-quantile from $A_1, \ldots, A_{n_{perm}}$.

Finally, the permutation based confidence intervals are given by

$$CI^{\tau}_g = \left[g^{-1}\left(g(\widehat{p}) - \frac{z^{\tau}_{1-\alpha/2}}{\sqrt{N}} g'(\widehat{p})\widehat{\sigma}_N \right); g^{-1}\left(g(\widehat{p}) - \frac{z^{\tau}_{\alpha/2}}{\sqrt{N}} g'(\widehat{p})\widehat{\sigma}_N \right) \right], \tag{4.4}$$

where $g \in \{id, logit, probit\}$. Estimate the two-sided p-value by

$$\text{p-value} = \min\{2p_1, 2-2p_1\}, \text{ where } p_1 = \frac{1}{n_{perm}} \sum_{\ell=1}^{n_{perm}} \mathbf{1}\{T_g \leq A_\ell\}.$$

5. Simulation Results

Neubert and Brunner (2007) have already investigated in extensive simulations that the studentized permutation test φ^{τ} controls the pre-assigned type-I-error level α accurately under the null, even for very small sample sizes and variuos underlying distributions. Therefore we restrict the following simulation study to investigate the small sample properties of the permutation based confidence intervals for the AUC. We compare the confidence intervals CI_{id}, CI_{logit} as well as CI_{probit} given in (3.2), (3.6) and (3.7), respectively, by using the standard normal quantiles $z_{1-\alpha/2}$, $t-$quantiles $t_{1-\alpha/2,\widehat{f}}$ or the permutation based quantiles $z^{\tau}_{1-\alpha/2}$ as given in (4.4). Thus, 9 different computation methods will be compared. The simulations vary in:

- The sample size of the diseased group of subjects ($n_1 = 5, 10, 20, 50$)
- The sample size of the non-diseased group of subjects ($n_2 = 5, 10, 20, 50$)
- The area under the curve ($AUC = 0.50, 0.60, 0.70, 0.80$)
- The distribution of the data (normal, logarithmic normal, exponential, uniform)

Due to the abundance of simulation results, we will merely present selected results for sample size settings $(n_1, n_2) = (5,5), (5,10), (10,10), (10,20)$. All further results are given in the supplementary material. All simulations were run using R (R Development Core Team, 2010) with $n_{sim} = 10,000$ simulation runs and $n_{perm} = 10,000$ random permutations. For an clear arrangement of the simulation results, we compare the asymptotic and approximate confidence intervals with the permutation

based confidence intervals with and without using a transformation function in the next subsections.

5.1. **Simulation results for untransformed confidence intervals.** First we consider the confidence intervals based on $g = id$. We investigate the behavior of the confidence intervals CI_{id} as given in (3.2), CI_t given in (3.4) and of the permutation based confidence intervals CI_{id}^{Perm} as given in (4.4) with regard to maintaning the pre-assigned coverage probability (95%) for varying true AUC's, sample sizes and data distributions. The simulation results are displayed in Figure 2. It can be readily seen that the standard confidence intervals CI_{id} and CI_t tend to be highly liberal when sample sizes are rather small ($n_i \leq 10$). On the contrary, the permutation based confidence intervals maintain the pre-assigned coverage probability at best, uniformly for all underlying data distributions (normal, uniform, log-normal, and exponential). Even when sample sizes are extremely small ($n_i = 5$) and in the case of unbalanced data, the permutation based confidence intervals greatly improve the standard intervals CI_{id} and CI_t. However, it can also be readily seen that all investigated intervals tend to highly liberal decisions when the true AUC is large. This effect is intuitively clear, since the variance of the estimator tends to close to be zero. Hence, very large sample sizes are necessary to obtain accurate results when $AUC \geq 80\%$. Next, the behavior of the range preserving confidence intervals will be explored.

5.2. **Simulation results for range-preserving confidence intervals.** We investigate the behavior of the logit-type range-preserving confidence intervals CI_{Logit}^{Normal} as given in (3.6) using standard normal quantiles, CI_{Logit}^{t} given in (3.6) using $t_{1-\alpha/2,\widehat{f}}$ quantiles and of the permutation based confidence intervals CI_{Logit}^{Perm} as given in (4.4) with regard to maintaning the pre-assigned coverage probability (95%) for varying true AUC's, sample sizes and data distributions. The simulation results are displayed in Figure 3. It turns out that both the confidence intervals CI_{Logit}^{Normal} and CI_{Logit}^{t} tend to result in rather conservative conclusions, particularly for small sample sizes $n_i \leq 10$. When sample sizes are extremely small ($n_i = 5$), their empirical coverage probabilities are very close to 1, a rather inappropriate property. This behavior can be detected for all kinds of investigated distributions. It can further be seen that the permutation based range-preserving confidence intervals reduce the amount of conservativity dramatically and maintain the pre-assigned coverage probability at best. However, a slightly conservative behavior can be detected. The same conclusions can be drawn using the probit-transformation function in Figure 4.

6. Example - Sonography Ultra-Sound Imaging Study

We reconsider the sonography ultra-sound imaging diagnostic study (Kaufmann et al. 2005) which was performed to assess leg or pelvic thrombosis in patients. The aim of this trial is to compare the diagnostic accuracy of the contrast medium Levovist, used in color-coded Doppler sonography, with non-enhanced sonography

(Baseline). Sonography with and without Levovist was performed for each of the $n_1 = 84$ non-diseased and $n_2 = 111$ diseased subjects as ensured by phlebography. Two blinded readers R1 and R2 interpreted the images from different modalities, sonography at baseline and Levovist-enhanced. Thus, the design of this trial is a typical paired-case paired-reader design (Kaufmann et al. 2005). The response variable is an ordered categorical score ranging from 1="thrombosis definitely no" through 5="thrombosis definitely yes".
We will estimate the AUC for each reader $\times$ modality combination seperately and compute 95%-confidence intervals with the studentized permutation approach. Figure 5 displays the ROC-curves for each reader $\times$ modality combination. The estimated AUC's as well as 95%-confidence intervals for each individual AUC are displayed in Table 1.

TABLE 1. AUC estimates and permutation-based 95%-confidence intervals for the sonography ultra-sound imaging diagnostic study.

Reader	Modality	AUC Estimate	Transformation	Lower	Upper
1	Baseline	0.50	Logit	0.36	0.65
			Probit	0.35	0.66
	Levovist	0.82	Logit	0.69	0.90
			Probit	0.69	0.90
2	Baseline	0.44	Logit	0.30	0.60
			Probit	0.29	0.60
	Levovist	0.67	Logit	0.50	0.80
			Probit	0.50	0.80

Table 1 demonstrates that the contrast medium Levovist enhances the diagnostic accuracy of sonography for both readers. Test results for the null hypotheses ,,no reader effect", ,,no modality effect" and ,,no interaction between reader and modality" are given in Kaufmann et al. (2005).

7. Discussion and conclusions

In this paper, we have investigated rank-based studentized permutation methods for the nonparametric Behrens-Fisher problem, i.e. inference methods for the area under the ROC-curve. In particular, we have theoretically shown that the studentized permutation test proposed by Neubert and Brunner (2007) can be inverted for the computation of confidence intervals for the underlying treatment. This property is highly desirable for practical applications. Furthermore, we have shown that permutation methods can be applied for the computation of range-preserving confidence intervals. Extensive simulation studies show that the permutation based confidence

intervals greatly improve the standard confidence intervals. However, all investigated confidence intervals tend to be highly liberal when the true AUC is very large and sample sizes occur. This liberality, however, decreases with larger sample sizes.

All investigated confidence intervals are implemented in the free R software package *nparcomp* within its function *npar.t.test* (Konietschke et al. 2014). The use of the function is as follows:

```
> data
Goldstandard Response
0             x
0             x
...           ...
0             x
1             x
1             x
...           ...
1             x
>library(nparcomp)
>npar.t.test(Response ~ Goldstandard, data=data, permu=TRUE,
 asy.method="logit") # Logit-type
```

8. Appendix

In order to show that the permutation based confidence intervals are asymptotically exact we need to prove conditional central limit theorems for the permutation statistic given in Equation (4.1). Therefore we would like to point out that its studentized version can be rewritten as a studentized two-sample rank statistic

$$\frac{\sqrt{N}(\widehat{p}-\frac{1}{2})}{\widehat{\sigma}_N} = \frac{\sqrt{\frac{n_1 n_2}{N}}\frac{1}{N}(\overline{R}_{2\cdot}-\overline{R}_{1\cdot})}{V_N}, \tag{8.1}$$

where $V_N^2 = \frac{n_2\widehat{\sigma}_1^2}{N} + \frac{n_1\widehat{\sigma}_2^2}{N}$. We will use this representation throughout the proofs for technical reasons. Moreover, we also introduce the quantity $H = \kappa F_1 + (1-\kappa)F_2$ with its natural estimator $\widehat{H} = N^{-1}(n_1\widehat{F}_1 + n_2\widehat{F}_2)$.

Proof of Theorem 4.1

We will prove that the conditional distribution of the permutation version of the statistic (8.1) is asymptotically standard normal for any treatment effect p. Therefore we start by addressing the enumerator of (8.1). Note, that it can be rewritten as

$$E_N = E_N(\mathbf{X}) := \sum_{i=1}^{N} c_{N,i}\widehat{H}(X_{N,i}),$$

where

$$X_{N,i} = \begin{cases} X_{1i}, & 1 \leq i \leq n_1 \\ X_{2i}, & n_1 + 1 \leq i \leq N \end{cases} \quad c_{N,i} = \sqrt{\frac{n_1 n_2}{N}} \begin{cases} -1/n_1, & 1 \leq i \leq n_1 \\ 1/n_2, & n_1 + 1 \leq i \leq N \end{cases}$$

Let τ be uniformly distributed on the symmetric group $\mathcal{S}_N$, i.e. the set of all permutations of the numbers $1, \ldots, N$, which is independent from the data $\mathbf{X}$. Here we understand independence in terms of an underlying product space, see e.g. the notation in Janssen (2005), Pauly (2011a) or Omelka and Pauly (2012) for more details. Denote the randomly permuted data by $\mathbf{X}^\tau := (X_{N,\tau(1)}, \ldots, X_{N,\tau(N)})$. Since we have $\widehat{H}^\tau(t) := \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{X_{N,\tau(i)} \leq t\} = \widehat{H}(t)$ the permuted enumerator fulfills

$$\begin{aligned} E_N^\tau &:= E_N(\mathbf{X}^\tau) && (8.2) \\ &= \sqrt{N} \sum_{i=1}^N c_{n,i} \frac{\widehat{H}(X_{N,\tau(i)})}{\sqrt{N}} && (8.3) \\ &\stackrel{d}{=} \sqrt{N} \sum_{i=1}^N c_{n,\tau(i)} \frac{\widehat{H}(X_{N,i})}{\sqrt{N}}, && (8.4) \end{aligned}$$

where $\stackrel{d}{=}$ means equality in distribution. We will now apply Theorem 4.1 from Pauly (2011a), see also Theorem 2.1 in Janssen (2005) for a more general version of this conditional central limit theorem for real-valued arrays. Note that our permuted coefficients $(c_{N,\tau(i)})_i$ fulfill Assumptions (2.3)-(2.5) in his paper. Hence it remains to check his Conditions (4.1)-(4.2) in the case of dimension 1 (i.e. $p = 1$ in the notation of his paper). Set $Z_{N,i} := \frac{\widehat{H}(X_{N,i})}{\sqrt{N}}$ and note that we have convergence in probability

$$\max_{1 \leq i \leq N} |Z_{N,i}| \longrightarrow 0.$$

Moreover, we have by means of the Extended Glivenko-Cantelli theorem, see e.g. in Shorack and Wellner (1986, Theorem 1 p.106), that $\sum_{i=1}^N (Z_{N,i} - \overline{Z}_N))^2$ is asymptotically equivalent to

$$\frac{1}{N} \sum_{i=1}^N (H(X_{N,i}) - \frac{1}{N} \sum_{j=1}^N H(X_{N,j}))^2.$$

Hence we can obtain from the law of large numbers that

$$\sum_{i=1}^N (Z_{N,i} - \overline{Z}_N)^2 \longrightarrow \sigma_\tau^2$$

converges in probability as $N \to \infty$, where

$$\sigma_\tau^2 := \kappa E(H(X_{11})^2) + (1 - \kappa) E(H(X_{21})^2) - [\kappa E(H(X_{11})) + (1 - \kappa) E(H(X_{21}))]^2.$$

Note, that our assumptions imply $\sigma_\tau^2 > 0$. Thus Theorem 4.1 from Pauly (2011a) shows conditional convergence in distribution of the permuted enumerator to a normal distribution with variance σ_τ^2, i.e.

$$\sup_{x\in\mathbb{R}} |P(E_N^\tau \leq x|\mathbf{X}) - \Phi(x/\sigma_\tau)| \longrightarrow 0 \tag{8.5}$$

in probability as $N \to \infty$, where Φ denotes the distribution function of the standard normal distribution.
It now remains to study the denominator of T_{id} or its squared version

$$V_N^2 = \frac{n_2}{N}\widehat{\sigma}_1^2 + \frac{n_1}{N}\widehat{\sigma}_2^2. \tag{8.6}$$

We will start by treating $\widehat{\sigma}_1^2$. Note that it can be rewritten as

$$\widehat{\sigma}_1^2(\mathbf{X}) = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(\widehat{F}_2(X_{1k}) - \frac{1}{n_1}\sum_{j=1}^{n_1}\widehat{F}_2(X_{1j}))^2.$$

Now its randomly permuted version behaves asymptotically as

$$\frac{1}{n_1}\sum_{k=1}^{n_1}(\widehat{F}_2^\tau(X_{N,\tau(k)}) - \frac{1}{n_1}\sum_{j=1}^{n_1}\widehat{F}_2^\tau(X_{N,\tau(j)}))^2, \tag{8.7}$$

where now, different to the situation with $\widehat{H}$ above, the permuted version

$$\widehat{F}_2^\tau(t) := \frac{1}{2n_2}\sum_{j=n_1+1}^{N}\Big(\mathbf{1}\{X_{N,\tau(j)} < t\} + \mathbf{1}\{X_{N,\tau(j)} \leq t\}\Big)$$

of F_2 is not invariant under the permutation of the data. Hence the summands $\widehat{F}_2^\tau(X_{N,\tau(k)}), 1 \leq k \leq n_1$, in (8.7) are dependent given the data $\mathbf{X}$. However, using a result from van der Vaart and Wellner (1996) we will see that there is asymptotically no difference if we replace $\widehat{F}_2^\tau$ by $\widehat{H}$. To be concrete, recall that the classes $\mathcal{F}^+ := \{\mathbf{1}\{(-\infty, t]\} : t \in \mathbf{R}\}$ and $\mathcal{F}^- := \{\mathbf{1}\{(-\infty, t)\} : t \in \overline{\mathbf{R}}\}$ are Donsker for every underlying distribution, i.e. especially for F_1^+ and F_2^+. Moreover, note that every Donsker Theorem implies a Glivenko-Cantelli Theorem in probability. Hence Theorem 3.7.1. in van der Vaart and Wellner (1996) together with the triangular inequality imply that

$$\sup_{x\in\mathbb{R}} |\widehat{F}_2^\tau(x) - \widehat{H}(x)| \longrightarrow 0$$

given $\mathbf{X}$ in probability. Together with the observations above from the first part of the proof, this shows that $\widehat{\sigma}_1^2(\mathbf{X}^\tau)$ is asymptotically equivalent (in probability) to

$$\begin{aligned}
&\frac{1}{n_1}\sum_{k=1}^{n_1}(H(X_{N,\tau(k)}) - \frac{1}{n_1}\sum_{j=1}^{n_1}H(X_{N,\tau(j)}))^2 \\
&= \frac{1}{n_1}\sum_{k=1}^{n_1}H(X_{N,\tau(k)})^2 - (\frac{1}{n_1}\sum_{j=1}^{n_1}H(X_{N,\tau(j)}))^2
\end{aligned}$$

given the data $\mathbf{X}$. Since H is bounded by 1, another application of Theorem 3.7.1. in van der Vaart and Wellner (1996) on the simple class $\mathcal{F} := \{H, H^2\}$ together with the above considerations imply the convergence of

$$\widehat{\sigma}_1^2(\mathbf{X}^\tau) \longrightarrow \sigma_\tau^2$$

in probability. Since the same holds for $\widehat{\sigma}_2^2(\mathbf{X}^\tau)$, it follows from the decomposition (8.6) that $V_N^2(\mathbf{X}^\tau)$ converges in probability to σ_τ^2 as $N \to \infty$. Hence it follows from (8.5) and Slutsky's theorem that

$$\sup_{x \in \mathbb{R}} |P(T_{id}^\tau \leq x | \mathbf{X}) - \Phi(x)| \longrightarrow 0$$

converges in probability as $N \to \infty$ and the proof is completed.

Acknowledgement

This work was supported by the German Research Foundation projects DFG Br 655/16-1 and Ho 1687/9-1.

References

[1] A. Agresti and E. Ryu. Pseudo-score confidence intervals for parameters in discrete statistical models. *Biometrika*, 97:215–222, 2010.

[2] M. G. Akritas, S.F. Arnold, and E. Brunner. Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92:258–265, 1997.

[3] A. I. Bandos, H. E. Rockette, and D. Gur. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statistics in Medicine*, 24:2873 – 2893, 2005.

[4] A. I. Bandos, H. E. Rockette, and D. Gur. A permutation test for comparing ROC curves in multireader studies 1: A multi-reader ROC permutation test. *Academic Radiology*, 13:414 – 420, 2006.

[5] D. Basso, F. Pesarin, L. Salmaso, and A. Solari. *Permutation Tests for Stochastic Ordering and ANOVA*. Springer, New York, 2009.

[6] T. M. Braun and T. A. Alonzo. A modified sign test for comparing paired ROC curves. *Biostatistics*, 9:364 – 372, 2008.

[7] E. Brunner and U. Munzel. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 1:17 – 21, 2000.

[8] European Medicines Agency (EMEA). Evaluation of medicines for human use, EMEA/596881/2007. *London*, 2008.

[9] M.P. Fay and M.A. Proschan. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010.

[10] X. Gao, M. Alvo, J. Chen, and G. Li. Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. *Journal of Statistical Planning and Inference*, 138:2574–2591, 2008.

[11] P. Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer, New York, 2005.

[12] A. Janssen. Studentized permutation test for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics and Probability Letters*, 36:9 – 21, 1997.

[13] A. Janssen. Testing nonparametric statistical functionals with application to rank tests. *Journal of Statistical Planning and Inference*, 81:71 – 93, 1999, Erratum 92, 297, 2001.

[14] A. Janssen. Resampling student's t-type statistics. *Annals of the Institute of Statistical Mathematics*, 57:507 – 529, 2005.
[15] H. Jin and Y. Lu. Permutation test for non-inferiority of the linear to the optimal combination of multiple tests. *Statistics and Probability Letters*, 79:664 – 669, 2009.
[16] J. Kaufmann, C. Werner, and E. Brunner. Nonparametric methods for analysing the accuracy of diagnostic tests with multiple readers. *Statistical Methods in Medical Research*, 14:129 – 146, 2005.
[17] F. Konietschke and M. Pauly. A studentized permutation test for the non-parametric Behrens-Fisher problem in paired data. *Electronic Journal of Statistics*, 6:1358 – 1372, 2012.
[18] F. Konietschke and M. Pauly. Bootstrapping and permuting paired t-test type statistics. *Statistics and Computing*, 24:283 – 296, 2014.
[19] F. Konietschke, M. Placzek, F. Schaarschmidt, and L. A Hothorn. nparcomp: An r software package for nonparametric multiple comparisons and simultaneous confidence intervals. *Journal of Statistical Software*, page In Press., 2014.
[20] K. Lange and E. Brunner. Sensitivity, specificity and ROC-curves in multiple reader diagnostic trials - a unified, nonparametric approach. *Statistical Methodology*, 9:490 – 500, 2012.
[21] P. Lèvy. *Calcul des Probabilites*. Gauthier-Villars, Paris, 1925.
[22] U. Munzel. Linear rank score statistics when ties are present. *Statistics and Probability Letters*, 41:389–395, 1999.
[23] K. Neubert and E. Brunner. A studentized permutation test for the non-parametric Behrens-Fisher problem. *Computational Statistics and Data Analysis*, 51:5192 – 5204, 2007.
[24] M. Omelka and M. Pauly. Testing equality of correlation coefficients in two populations via permutation methods. *Journal of Statistical Planning and Inference*, 142:1396 – 1406, 2012.
[25] M. Pauly. Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics*, 5:41 – 52, 2011a.
[26] M. Pauly. Discussion about the quality of F-ratio resampling tests for comparing variances. *Test*, 20:163 – 179, 2011b.
[27] M. Pauly, E. Brunner, and F. Konietschke. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society - Series B*, page In Press., 2014.
[28] F. Pesarin and L. Salmaso. *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley, Chichester, 2010.
[29] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.
[30] F.H. Ruymgaart. A unified approach to the asymptotic distribution theory of certain midrank statistics. *Statistique non Parametrique Asymptotique, 118, J.P. Raoult. Lecture Notes on Mathematics*, 821, 1980.
[31] E. Ryu. Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. *Statistics in Medicine*, 28:3179 – 3188, 2009.
[32] E. Ryu and A. Agresti. Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, 27:1703–1717, 2008.
[33] G. S. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*. Wiley, New York, 1986.
[34] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.
[35] E. S. Venkatraman. A permutation test to compare receiver operating characteristic curves. *Biometrics*, 56:1134 – 1138, 2000.
[36] E. S. Venkatraman and Colin B. Begg. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83:835 – 848, 1996.

[37] B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

Institute of Mathematics, University of Duesseldorf, 40225 Duesseldorf, Universitaetsstrasse 1, Germany
E-mail address: pauly@math.uni-duesseldorf.de

Department of Medical Statistics, University of Goettingen, 37073 Goettingen, Humboldtallee 32, Germany
E-mail address: thomas.asendorf@stud.uni-goettingen.de

Department of Medical Statistics, University of Goettingen, 37073 Goettingen, Humboldtallee 32, Germany
E-mail address: fkoniet@gwdg.de

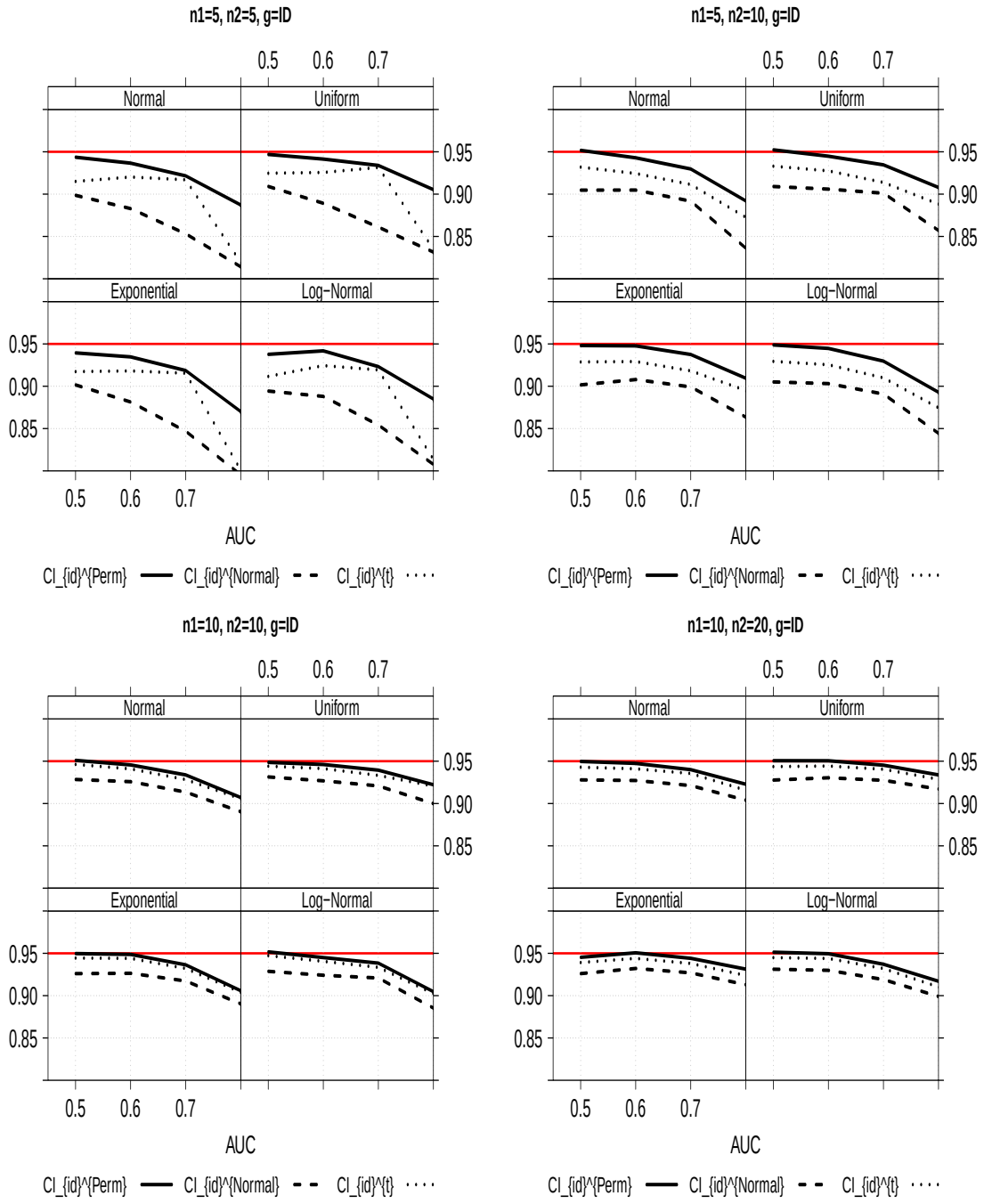

FIGURE 2. 95%-coverage probabilities of the three different confidence intervals CI_{id}^{Perm} as given in (4.4), CI_{id} as given in (3.2) and CI_t given in (3.4) for different sample size constellations: Top left $n_1 = 5$, $n_2 = 5$; top right $n_1 = 5$, $n_2 = 10$; down left $n_1 = 10$, $n_2 = 10$; down right $n_1 = 10$, $n_2 = 20$.

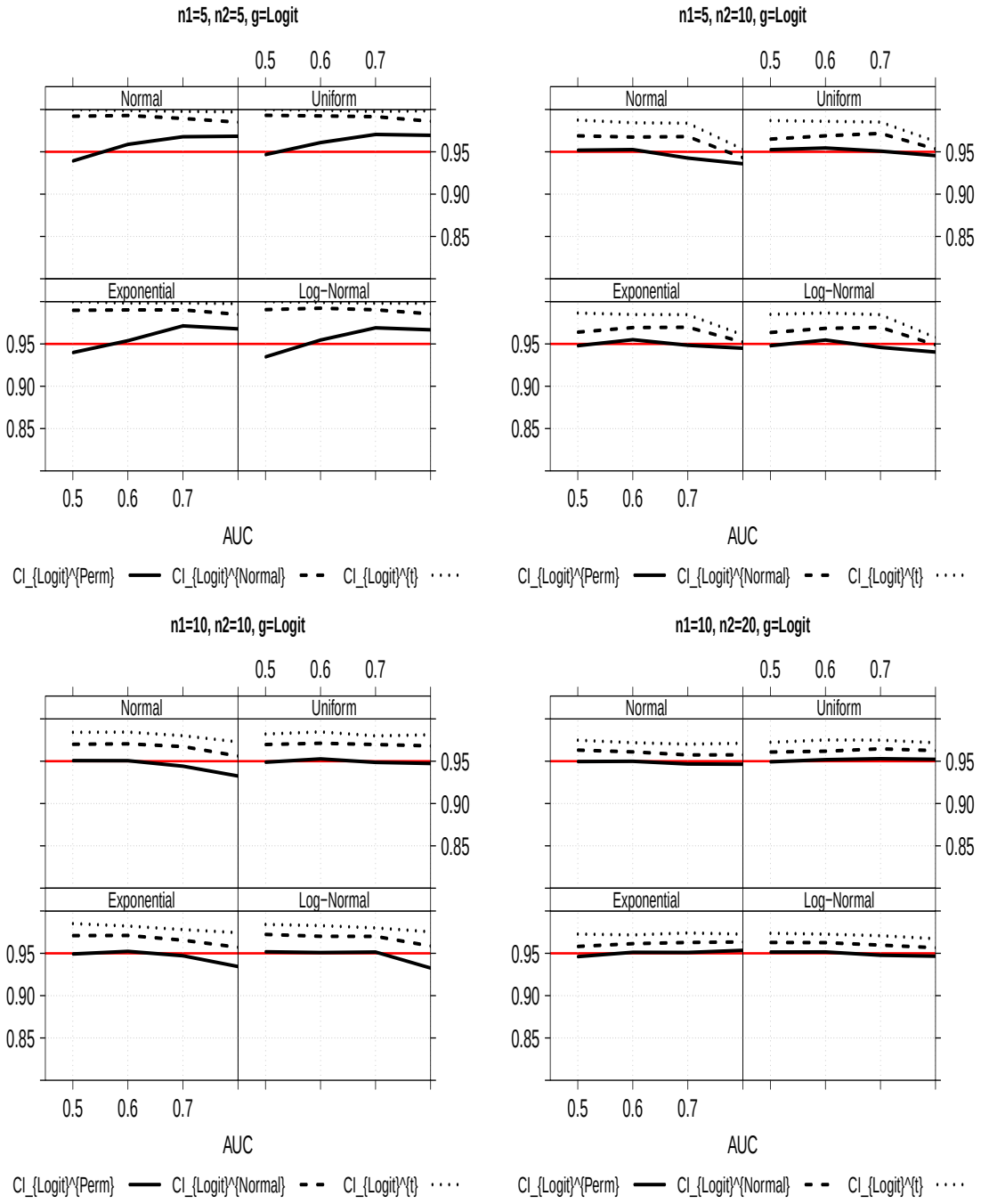

FIGURE 3. 95%-coverage probabilities of the three different confidence intervals CI^{Perm}_{Logit} as given in (4.4), CI^{Normal}_{Logit} as given in (3.6) using standard normal quantiles and CI^{t}_{Logit} given in (3.6) using $t_{1-\alpha/2,\widehat{f}}$ quantiles for different sample size constellations: Top left $n_1 = 5$, $n_2 = 5$; top right $n_1 = 5$, $n_2 = 10$; down left $n_1 = 10$, $n_2 = 10$; down right $n_1 = 10$, $n_2 = 20$.

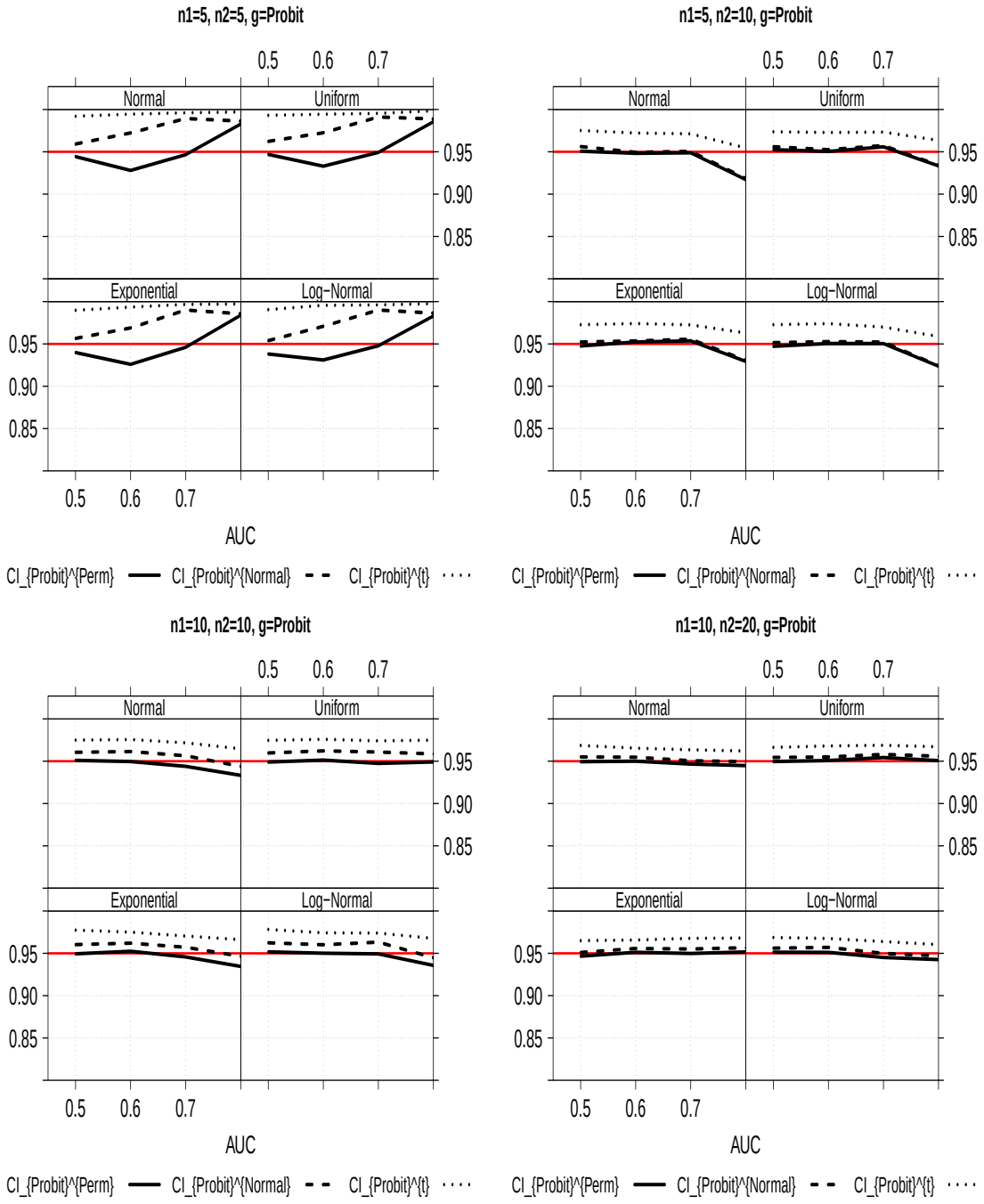

FIGURE 4. 95%-coverage probabilities of the three different confidence intervals CI^{Perm}_{Probit} as given in (4.4), CI^{Normal}_{Probit} as given in (3.7) using standard normal quantiles and CI^{t}_{Probit} given in (3.7) using $t_{1-\alpha/2,\widehat{f}}$ quantiles for different sample size constellations: Top left $n_1 = 5$, $n_2 = 5$; top right $n_1 = 5$, $n_2 = 10$; down left $n_1 = 10$, $n_2 = 10$; down right $n_1 = 10$, $n_2 = 20$.

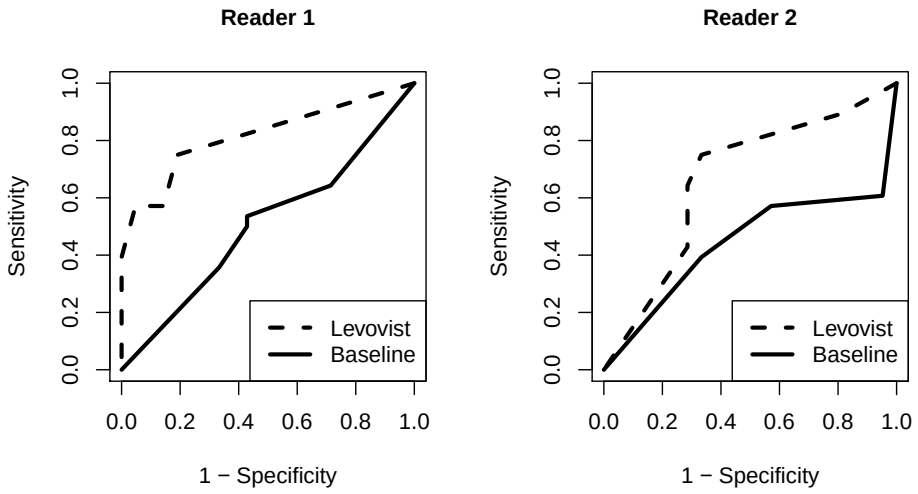

FIGURE 5. ROC curves of reader 1 and reader 2 for Levovist and the baseline.