Zapf *et al. BMC Medical Research Methodology* (2015) 15:43
DOI 10.1186/s12874-015-0025-y

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

A Wild Bootstrap approach for the selection of biomarkers in early diagnostic trials

Antonia Zapf[1*], Edgar Brunner[1] and Frank Konietschke[2]

Abstract

Background: In early diagnostic trials, particularly in biomarker studies, the aim is often to select diagnostic tests among several methods. In case of metric, discrete, or even ordered categorical data, the area under the receiver operating characteristic (ROC) curve (denoted by AUC) is an appropriate overall accuracy measure for the selection, because the AUC is independent of cut-off points.

Methods: For selection of biomarkers the individual AUC's are compared with a pre-defined threshold. To keep the overall coverage probability or the multiple type-I error rate, simultaneous confidence intervals and multiple contrast tests are considered. We propose a purely nonparametric approach for the estimation of the AUC's with the corresponding confidence intervals and statistical tests. This approach uses the correlation among the statistics to account for multiplicity. For small sample sizes, a Wild-Bootstrap approach is presented. It is shown that the corresponding intervals and tests are asymptotically exact.

Results: Extensive simulation studies indicate that the derived Wild-Bootstrap approach keeps and exploits the nominal type-I error at best, even for high accuracies and in case of small samples sizes. The strength of the correlation, the type of covariance structure, a skewed distribution, and also a moderate imbalanced case-control ratio do not have any impact on the behavior of the approach. A real data set illustrates the application of the proposed methods.

Conclusion: We recommend the new Wild Bootstrap approach for the selection of biomarkers in early diagnostic trials, especially for high accuracies and small samples sizes.

Keywords: AUC, Diagnostic study, Resampling, Simultaneous intervals, Wild bootstrap

Background

The aim of early diagnostic trials, particularly of biomarker studies, is often to select the most promising markers from a candidate set. For convenience, all different kinds of diagnostic tests, e.g., imaging techniques or biomarkers, will be denoted by *diagnostic tests* throughout the paper. In these studies, response variables are often not binary, but measured on a continuous, discrete or even ordinal scale and a cut-off value c has not yet been chosen. Therefore, the sensitivity (i.e. true positive proportion) and the specificity (true negative proportion) both being computed based on c cannot be used as selection criteria. In contrast, the Receiver Operating Characteristic (ROC) curve illustrates the overall diagnostic performance because it is independent of the chosen cut-off values (see, e.g., DeLong, DeLong and Clark-Pearson [1]). Because the ROC curve of a diagnostic test is invariant with respect to any monotone transformation of the test measurement scale, it is an adequate measure for comparing diagnostic tests being measured even on different scales. The Area Under the ROC-curve (AUC) represents an accuracy measure which is independent from the selected cut-off value c and which is invariant under any monotone transformation of the data. Therefore, it is an appropriate selection criterion for promising diagnostic tests, and in particular Xia et al. [2] (p. 286) state in their tutorial about translational biomarker discovery in clinical metabolomics that the *"AUC is widely used for performance comparison across different biomarker models"*.

As an example for the evaluation of different biomarkers we consider the ICM trial by Derichs et al. [3], which aims

*Correspondence: Antonia.Zapf@med.uni-goettingen.de
[1]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany
Full list of author information is available at the end of the article

to evaluate the diagnostic accuracy of intestinal current measurement (ICM) with regard to questionable cystic fibrosis (CF). This study was conducted with the approval of the local ethics committee, MH Hannover, Germany and all patients and/or parents and healthy controls gave their written informed consent. In this trial, a total of $N = 67$ children and adults were enrolled. The true disease state of the patients was defined by a composite gold standard, which consists of typical CF symptoms plus either a positive sweat test and/or gene mutations. By this definition 26 patients were classified into CF (referred to as cases) and 41 into 'CF unlikely' (referred to as controls). Furthermore, four biomarkers were considered: $\Delta I_{sc,carbachol}$, $\Delta I_{sc,cAMP/forskolin}$, and $\Delta I_{sc,histamine}$ (abbreviated by ΔI_{carb}, ΔI_{cAMP}, and ΔI_{hista}) as well as the sum of the three measured values, ΔI_{sum}. Boxplots of the data are displayed in Figure 1.

In the ROC-curves in Figure 2 the corresponding estimated AUC's are added. It can be readily seen that the diagnostic accuracy of ΔI_{carb}, ΔI_{cAMP}, and ΔI_{hista} is quite good, and that ΔI_{sum} perfectly differentiates the cases and the controls.

Thus, the remaining question is which biomarkers have sufficient diagnostic accuracy. There is no consensus about the threshold for sufficent diagnostic accuracy. Xia et al. [2] characterize a biomarker with an $AUC < 0.7$ as a quite "weak" biomarker. In their study about a blood-based biomarker panel for stratifying current risk for colorectal cancer Marshall et al. [4] accept a candidate model with an $AUC > 0.75$ as a predictive model. In contrast, Broadhurst and Kell [5] refer to an $AUC > 0.9$ as excellent and to an $AUC > 0.8$ as good. Depending on previous knowledge or expectations a threshold for the AUC as indicator for sufficient diagnostic accuracy should be chosen during the planning of the trial.

Note that the aim of such trials is not to test multiple hypotheses formulated in terms of AUC differences across the biomarkers, but to verify sufficient diagnostic accuracy for all biomarkers individually. Then comparing the lower limit of the confidence interval for the estimated AUC with this threshold indicates whether or not the diagnostic test has sufficient diagnostic accuracy. The "Guideline on the choice of the non-inferiority margin" of the European Medicines Agency [6] recommends to demonstrate non-inferiority by use of two-sided 95% or one-sided 97.5% confidence intervals.

If several diagnostic tests are evaluated in the same trial, it is important to adjust the confidence intervals for multiplicity. Otherwise there is a high risk that the accuracy of some diagnostic tests is overestimated. Xia et al. [2] (p.288) point out that *"The probability of finding a random association between a given metabolite and the outcome increases with the total number of comparisons"*. Furthermore they note that the Bonferroni correction is a simple but very conservative method. If the diagnostic tests are repeatedly measured on the same subjects, hence, these

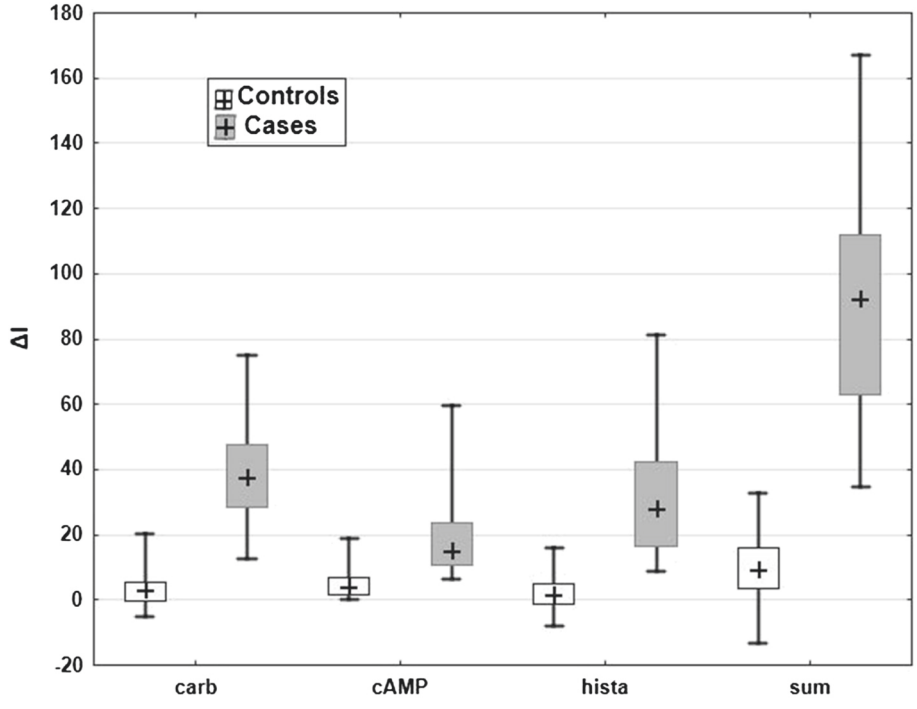

Figure 1 Boxplots of the biomarkers. Boxplot of the four biomarkers in the example, separately for cases and controls.

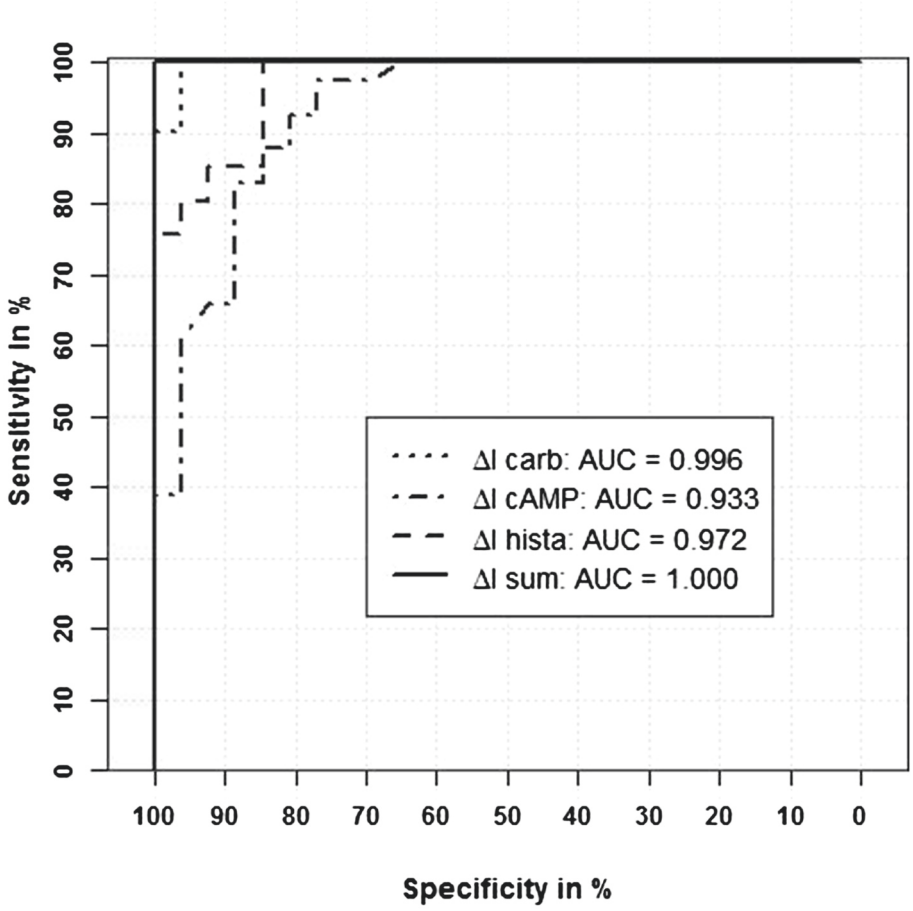

Figure 2 ROC-curves of the biomarkers. ROC-curves of the four biomarkers in the example and corresponding AUC's in the legend.

measurements are correlated in general. Therefore it is of highly practical importance to take into account these correlations in the estimation of the diagnostic accuracy.

The multiplicity expert group of the 'Statisticians in the Pharmaceutical Industry' [7] (p.258) states that *"The participants did, however, agree that for non-inferiority and equivalence trials, compatible simultaneous CIs for the primary endpoint(s) should be presented in all cases"*. Furthermore Strassburger and Bretz [8] recommend the use of single-step procedures if the aim is not to reject as many hypotheses as possible. Therefore we will confine ourselves to simultaneous confidence intervals from single-step procedures which are compatible with the results obtained by hypotheses tests. Among others, Hothorn et al. [9] proposed parametric simultaneous confidence intervals, which correspond to multiple contrast tests. However, since these parametric approaches are limited to normally distributed data, Konietschke et al. [10] proposed nonparametric multiple contrast tests and compatible asymptotic simultaneous confidence intervals for relative treatment effects for independent samples (based on some theoretical results developed by Brunner et al. [11]). In the particular case of two samples (cases and controls) the relative treatment effect is equivalent to the AUC (see Bamber [12]). In this article we will use this approach in the framework of diagnostic studies, but for paired samples in a multivariate layout.

The challenge in early diagnostic trials is often that smaller sample sizes and higher AUC's occur. For example in the systematic review of 10 studies about the diagnostic accuracy of pleural fluid NT-pro-BNP for pleural effusions of cardiac origin, performed by Janda and Swiston [13], the median total sample size was 104 (mean 112), and the pooled AUC was 98%. Wang et al. [14] reported in another systematic review about cardiac testing for coronary artery disease in potential kidney transplant recipients AUC's between 0.78 and 0.92. Kottas et al. [15] found that the Logit tranformation based confidence interval for a single AUC leads to slightly conservative results for small sample sizes. Here we suggest Wild Bootstrap based simultaneous confidence intervals to obtain robust methods for small sample sizes and potentially quite large

AUC's. Hereby, we generalize the method proposed by Arlot et al. [16] for multivarite high-dimensional normal data.

In this article nonparametric simultaneous confidence intervals for multiple AUC's in diagnostic studies are presented. Asymptotic intervals will be derived as well as intervals using the Wild Bootstrap approach. The properties of these simultaneous intervals are investigated in a simulation study regarding the type-I error rate and the statistical power. Furthermore, the results of all intervals are given for the example data set presented before in this section. In the next section we present the methods, including the statistical model with the corresponding hypotheses, and the point estimators with their asymptotic distribution. Furthermore multiple contrast tests and corresponding simultaneous confidence intervals (with or without Logit transformation) are derived, and the Wild Bootstrap approach is presented (in particular for small sample sizes). The results of a simulation study including robustness evaluations, and the application of the methods to the example presented above are given in the Results section. Finally, all results are summarized and discussed, and a recommendation is given.

Methods

Statistical model and hypotheses

We consider a within-subject multi-modality diagnostic trial given by independent and identically distributed random vectors

$$\mathbf{X}_{is} = \left(X_{is}^{(1)}, \ldots, X_{is}^{(d)}\right)' \sim \mathbf{F}_i,\ i = 0, 1(\text{control, case});\quad \text{subject } s = 1, \ldots, n_i, \tag{1}$$

with marginal distributions

$$X_{is}^{(\ell)} \sim F_{is}^{(\ell)}, \ell = 1, \ldots, d, \tag{2}$$

where d denotes the number of diagnostic tests. The partition of the data in cases ($i = 1$) or controls ($i = 0$) is based on the gold or reference standard, which is assumed to represent the true disease status of the subjects. In order to allow for continuous, discrete or even ordered categorical data in a unified way, we use the normalized version of the marginal distribution functions, i.e., $F_i^{(\ell)}(x) = \frac{1}{2}\left(F_i^{(+,\ell)}(x) + F_i^{(-,\ell)}(x)\right)$, where $F_i^{(+,\ell)}(x) = P\left(X_{i1}^{(\ell)} \leq x\right)$ denotes the right-continuous and $F_i^{(-,\ell)}(x) = P\left(X_{i1}^{(\ell)} < x\right)$ denotes the left-continuous version of the distribution function respectively. In the context of nonparametric models, the normalized version of the distribution function was first mentioned by Kruskal [17] and generally dates back to Lévy [18]. Later on, it was used by Ruymgaart [19], Munzel [20], Brunner and Puri [21], Kaufmann et al. [22], among others, to derive asymptotic results for rank statistics including the case of ties. We note that $F_i^{(\ell)}(x)$ may be arbitrary distribution functions, with the exception of the trivial case that both distributions are one-point distributions (see Lange and Brunner [23]).

The within-subject design given in (1), which means that all diagnostic tests are performed in each individual, is recommended in the EMA guideline about diagnostic agents [24] and refers to Design 1 in Brunner and Zapf [25].

For each of the d diagnostic tests the true AUC is given by

$$\begin{aligned} AUC^{(\ell)} &= P(X_{01}^{(\ell)} < X_{11}^{(\ell)}) + 0.5 \cdot P(X_{01}^{(\ell)} = X_{11}^{(\ell)}) \\ &= \int F_0^{(\ell)} dF_1^{(\ell)}\ , \ell = 1, \ldots, d. \end{aligned} \tag{3}$$

For a convenient derivation of asymptotic results, the AUC's are collected in the vector $\mathbf{AUC} = \left(AUC^{(1)}, \ldots, AUC^{(d)}\right)'$.

In order to select the most promising diagnostic tests from the candidate set of the d different methods, it is our aim to test the non-inferiority null hypotheses

$$\begin{aligned} H_0 &: \bigcap_{\ell=1}^{d} \left\{H_0^{(\ell)} : AUC^{(\ell)} \leq AUC_0\right\} \quad \text{versus} \\ H_1 &: \bigcup_{\ell=1}^{d} \left\{H_1^{(\ell)} : AUC^{(\ell)} > AUC_0\right\} \end{aligned} \tag{4}$$

with strong control of the familywise error rate (FWER) α simultaneously. The non-inferiority margin AUC_0 is assumed to have been fixed during the planning phase of the trial. Thus, the set of promising diagnostic tests consists of all markers, whose corresponding $AUC^{(\ell)}$ have been declared to be larger than AUC_0 by an adequate multiple testing procedure.

Point estimators and asymptotic distribution

Unbiased and L_2-consistent point estimators for the AUC's defined in (3) are derived by replacing the unknown distribution functions $F_0^{(\ell)}$ and $F_1^{(\ell)}$ by their empirical counterparts

$$\widehat{F}_i^{(\ell)}(x) = \frac{1}{n_i} \sum_{s=1}^{n_i} c(x - X_{is}^{(\ell)}),\ i = 0, 1;\ \ell = 1, \ldots, d,$$

where $c(x)$ denotes the normalized version of the count function, i.e. $c(x) \in \{0, \frac{1}{2}, 1\}$ corresponding to $\{x < 0, x = 0, x > 0\}$, respectively. The point estimator

$$\widehat{AUC}^{(\ell)} = \int \widehat{F}_0^{(\ell)} d\widehat{F}_1^{(\ell)} = \frac{1}{N}\left(\overline{R}_{1\cdot}^{(\ell)} - \overline{R}_{0\cdot}^{(\ell)}\right) + \frac{1}{2} \tag{5}$$

can easily be computed using the means $\overline{R}_{i\cdot}^{(\ell)} = n_i^{-1} \sum_{s=1}^{n_i} R_{is}^{(\ell)}$ of the (mid-) ranks $R_{is}^{(\ell)}$, $i = 0, 1$. Here, $R_{is}^{(\ell)}$ denotes the rank of $X_{is}^{(\ell)}$ among all $N = n_0 +$

n_1 observations $X_{01}^{(\ell)},\ldots,X_{0n_0}^{(\ell)}$, $X_{11}^{(\ell)},\ldots,X_{1n_1}^{(\ell)}$ per marker $\ell = 1,\ldots,d$. Further let $\mathbf{R}_{is} = \left(R_{is}^{(1)},\ldots,R_{is}^{(d)}\right)'$ denote the vectors of the midranks and let $\widehat{\mathbf{AUC}} = \left(\widehat{AUC}^{(1)},\ldots,\widehat{AUC}^{(d)}\right)'$ denote the vector of the point estimators.

Brunner et al. [11] have shown that the vector $\sqrt{N}(\widehat{\mathbf{AUC}} - \mathbf{AUC})$ follows, asymptotically, as $N \to \infty$, a multivariate normal distribution with expectation $\mathbf{0}$ and covariance matrix

$$\mathbf{V}_N = Cov(\sqrt{N}\mathbf{B}), \tag{6}$$

where $\mathbf{B} = \left(B^{(1)},\ldots,B^{(d)}\right)'$ denotes a random vector the components of which are sums of independent random variables

$$B^{(\ell)} = \frac{1}{n_1}\sum_{s=1}^{n_1} F_0^{(\ell)}(X_{1s}^{(\ell)}) - \frac{1}{n_0}\sum_{s=1}^{n_0} F_1^{(\ell)}(X_{0s}^{(\ell)}) + 1 - 2\cdot AUC^{(\ell)}. \tag{7}$$

The covariance matrix $\mathbf{V}_N$ with elements $v^{(\ell,m)}$, however, is unknown and has to be estimated. Let $R_{is}^{(i|\ell)}$ denote the so-called internal rank of $X_{is}^{(\ell)}$ among all n_i observations $X_{i1}^{(\ell)},\ldots,X_{in_i}^{(\ell)}$ for the diagnostic test ℓ in disease status group i, and let $\mathbf{R}_{is}^{(i)} = \left(R_{is}^{(i|1)},\ldots,R_{is}^{(i|d)}\right)'$ denote the vectors of these internal ranks. Furthermore, let

$$\mathbf{Z}_{is} = \frac{1}{N-n_i}\left(\mathbf{R}_{is} - \mathbf{R}_{is}^{(i)}\right) \tag{8}$$

denote the vectors of the normed placements

$$\widehat{F}_0^{(\ell)}(X_{1s}^{(\ell)}) = \frac{1}{n_0}\left(R_{1s}^{(\ell)} - R_{1s}^{(1|\ell)}\right) \text{ and}$$
$$\widehat{F}_1^{(\ell)}(X_{0s}^{(\ell)}) = \frac{1}{n_1}\left(R_{0s}^{(\ell)} - R_{0s}^{(0|\ell)}\right),$$

respectively. Then a consistent estimator of the covariance matrix is given by $\widehat{\mathbf{V}}_N = N\left(\widehat{\mathbf{V}}_{N,0}/n_0 + \widehat{\mathbf{V}}_{N,1}/n_1\right)$, where

$$\widehat{\mathbf{V}}_{N,i} = \frac{1}{n_i-1}\sum_{s=1}^{n_i}\left(\mathbf{Z}_{is} - \overline{\mathbf{Z}}_{i\cdot}\right)\left(\mathbf{Z}_{is} - \overline{\mathbf{Z}}_{i\cdot}\right)', \; i = 0,1. \tag{9}$$

Here, $\overline{\mathbf{Z}}_{i\cdot} = \frac{1}{n_i}\sum_{s=1}^{n_i}\mathbf{Z}_{is}$ denotes the vector of means of the normed placements. For more details we refer to Brunner et al. [11] and Kaufmann et al. [22].

Test statistics and confidence intervals

In order to test the null hypotheses formulated in (4), we first need to derive an univariate test statistic for testing the individual null hypothesis $H_0^{(\ell)} : AUC^{(\ell)} \le AUC_0$. It follows from the asymptotic multivariate normality of the vector $\sqrt{N}(\widehat{\mathbf{AUC}} - \mathbf{AUC})$ that $\sqrt{N}(\widehat{AUC}^{(\ell)} - AUC^{(\ell)})$ has, asymptotically as $N \to \infty$, a univariate normal distribution with mean 0 and variance $v^{(\ell,\ell)}$, i.e. $N(0, v^{(\ell,\ell)})$. Here, $v^{(\ell,\ell)}$ denotes the ℓ-th diagonal element of $\mathbf{V}_N$ in (6). Hence, by Slutzky's theorem, it follows that

$$T^{(\ell)} = \left(\widehat{AUC}^{(\ell)} - AUC^{(\ell)}\right)\sqrt{\frac{N}{\widehat{v}^{(\ell,\ell)}}} \xrightarrow{\mathcal{D}} N(0,1), \text{ as } N \to \infty, \tag{10}$$

where $\widehat{v}^{(\ell,\ell)}$ denotes the diagonal elements of $\widehat{V}_N$, defined in (9). In particular, each statistic is studentized with an individual consistent variance estimator and thus, the set of hypotheses and test statistics $\Omega = \left\{\left(H_0^{(\ell)}, T^{(\ell)}\right), \ell = 1,\ldots,d\right\}$ constitutes a joint-testing family in the sense of Gabriel [26]. Attention should be paid to the fact that the estimated variance $\widehat{v}^{(\ell,\ell)}$ is equal to zero if $\widehat{AUC}^{(\ell)} = 0$ or 1. Thus, the test statistic $T^{(\ell)}$ can not be computed. One possibility to solve this problem is to modify the data slightly (see the analysis of the example in the Results section).

A quite conservative selection approach can be derived by applying the Bonferroni method (denoted as 'Bonf'), i.e., the individual null hypothesis $H_0^{(\ell)} : AUC^{(\ell)} \le AUC_0$ will be rejected at multiple level α, if $T^{(\ell)} \le z_{1-\alpha/d,1}$, where $z_{1-\alpha/d,1}$ denotes the one-sided $(1-\alpha/d)$-quantile of the standard normal distribution. Asymptotic one-sided simultaneous confidence intervals for the treatment effects $AUC^{(\ell)}$ are then given by

$$CI_{Bonf}^{(\ell)} = \left[\widehat{AUC}^{(\ell)} - z_{1-\alpha/d,1}\sqrt{\frac{\widehat{v}^{(\ell,\ell)}}{N}}; 1\right]. \tag{11}$$

The global null hypothesis $H_0 : \mathbf{AUC} \le AUC_0 \cdot \mathbf{1}$ as defined in (4) will be rejected, if $\max\{T^{(1)},\ldots,T^{(d)}\} > z_{1-\alpha/d,1}$ or, equivalently, if the maximum of the lower limits of the confidence intervals $\max\{CI_{Bonf,l}^{(1)},\ldots,CI_{Bonf,l}^{(d)}\} > AUC_0$. Here $\mathbf{1} = (1,\ldots,1)'$ denotes a d-dimensional vector of 1s. The Bonferroni method is, however, a quite conservative selection approach (see Results section for more details). The reason for this is that the apparent correlations among the different pivotal quantitites $T^{(1)},\ldots,T^{(d)}$ are not taken into account by this method.

Multiple contrast tests and simultaneous confidence intervals

In order to use the correlation in the selection approach, it is our idea to apply the multiple contrast test principle (denoted by MCP), which uses the correlation among different test statistics. The key point of these procedures is to use the joint distribution of a set of statistics to adjust for multiplicity. Thus, the asymptotic multivariate distribution of the vector $\mathbf{T} = \left(T^{(1)},\ldots,T^{(d)}\right)'$ is required. The details are stated in the next theorem.

Theorem 1. Under the assumption that $N \to \infty$ such that $N/n_i \le N_0 < \infty$, $i = 0,1$, the vector $\mathbf{T}$ follows, asymptotically, a multivariate normal distribution

with expectation $\mathbf{0}$ and correlation matrix $\mathbf{R}$, where $\mathbf{R} = [r^{(\ell,m)}]_{\ell,m=1\ldots,d}$, and $r^{(\ell,m)} = \frac{v^{(\ell,m)}}{\sqrt{v^{(\ell,\ell)}v^{(m,m)}}}$.

The joint distribution of $\mathbf{T}$ can be used for the derivation of a simultaneous test procedure. Let $z_{1-\alpha,1}(\mathbf{R})$ denote the one-sided $(1-\alpha)$ equicoordinate quantile of the multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix $\mathbf{R}$, i.e., $N(\mathbf{0},\mathbf{R})$, that is

$$P\left(\bigcap_{\ell=1}^{d}\left\{T^{(\ell)} \le z_{1-\alpha,1}(\mathbf{R})\right\}\right) = 1-\alpha.$$

For details see Bretz et al. [27]. Then, the individual null hypothesis $H_0^{(\ell)} AUC^{(\ell)} \le AUC_0$ will be rejected at multiple level α, if

$$T^{(\ell)} \ge z_{1-\alpha,1}(\mathbf{R}). \tag{12}$$

Asymptotic one-sided simultaneous confidence intervals for $AUC^{(\ell)}$ are given by

$$CI_{MCP}^{(\ell)} = \left[\widehat{AUC}^{(\ell)} - z_{1-\alpha,1}(\mathbf{R})\sqrt{\frac{\widehat{v}^{(\ell,\ell)}}{N}}; 1\right]. \tag{13}$$

The global null hypothesis will be rejected if $\max\{T^{(1)},\ldots,T^{(d)}\} > z_{1-\alpha,1}(\mathbf{R})$ or if $\max\{CI_{MCP,l}^{(1)},\ldots, CI_{MCP,l}^{(d)}\} > AUC_0$. The correlation matrix $\mathbf{R}$, however, is unknown and must be replaced by a consistent estimator $\widehat{\mathbf{R}}$. We propose to replace $\mathbf{R}$ by $\widehat{\mathbf{R}}$ in the considerations above, where $\widehat{\mathbf{R}} = [\widehat{r}^{(\ell,m)}]_{\ell,m=1,\ldots,d}$ and $\widehat{r}^{(\ell,m)} = \frac{\widehat{v}^{(\ell,m)}}{\sqrt{\widehat{v}^{(\ell,\ell)}\widehat{v}^{(m,m)}}}$, respectively.

Simulation studies indicate, however, that the speed of convergence of $\mathbf{T}$ to a multivariate normal distribution is quite slow, particularly when smaller sample sizes and larger numbers of diagnostic tests are considered. In a variety of applications, see e.g. Zou and Yue [28] or Konietschke et al. [10], it turns out that the use of adequate transformations (e.g., the Logit-transformation) tend to increase the speed of convergence. Therefore, simultaneous confidence intervals with Logit transformation will be derived in the next section.

Multiple contrast tests and simultaneous confidence intervals with Logit transformation

To derive simultaneous Logit-transformed confidence intervals let

$$\mathbf{g}(\mathbf{AUC}) = \left(g(AUC^{(1)}),\ldots,g(AUC^{(d)})\right) : (0,1)^d \to \mathbb{R}^d,$$

denote the vector of Logit-transformed AUC's, where

$$g(AUC^{(\ell)}) = \log\left(\frac{AUC^{(\ell)}}{1-AUC^{(\ell)}}\right).$$

Furthermore, let

$$\mathbf{\Psi} = diag\left(\frac{1}{AUC^{(1)}(1-AUC^{(1)})},\ldots,\frac{1}{AUC^{(d)}(1-AUC^{(d)})}\right)$$

denote the diagonal Jacobian matrix of $\mathbf{g}(\mathbf{AUC})$. Under the additional assumption that $N \to \infty$ such that $N/n_i \to f_i$, it follows from Cramer's multivariate δ-theorem (see, e.g., Ferguson [29], Theorem 7.4) that

$$\sqrt{N}\left(\mathbf{g}\left(\widehat{\mathbf{AUC}}\right) - \mathbf{g}\left(\mathbf{AUC}\right)\right) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \mathbf{S}_N\right) \tag{14}$$

where $\mathbf{S}_N = \mathbf{\Psi}\mathbf{V}_N\mathbf{\Psi}'$ and $\mathbf{V}_N$ is given in (6). To estimate the asymptotic covariance matrix $\mathbf{S}_N$, let

$$\widehat{\mathbf{\Psi}} = diag\left(\frac{1}{\widehat{AUC}^{(1)}(1-\widehat{AUC}^{(1)})},\ldots,\frac{1}{\widehat{AUC}^{(d)}(1-\widehat{AUC}^{(d)})}\right)$$

denote the estimated Jacobian matrix of $\mathbf{g}(\mathbf{AUC})$ and note that the estimator $\widehat{\mathbf{S}}_N = \widehat{\mathbf{\Psi}}\widehat{\mathbf{V}}_N\widehat{\mathbf{\Psi}}$ is a consistent estimator of $\mathbf{S}_N$. Again there is a problem if $\widehat{AUC}^{(\ell)} = 0$ or 1. Here, $\widehat{\mathbf{\Psi}}$ and in turn $\widehat{\mathbf{S}}_N$ cannot be calculated. This problem is addressed in the analysis of the example in the Results section. To test the individual hypothesis $H_0^{(\ell)} : AUC^{(\ell)} \le AUC_0$ define the pivotal quantities

$$\widetilde{T}^{(\ell)} = \left(g(\widehat{AUC}^{(\ell)}) - g(AUC^{(\ell)})\right)\sqrt{\frac{N}{\widehat{s}^{(\ell,\ell)}}} \xrightarrow{\mathcal{D}} N(0,1),$$
$$N \to \infty,\ \ell = 1,\ldots,d, \tag{15}$$

where $\widehat{s}^{(\ell,\ell)}$ denotes the ℓ-th diagonal element of $\widehat{\mathbf{S}}_N$. The joint distribution of the vector $\widetilde{\mathbf{T}} = (\widetilde{T}^{(1)},\ldots,\widetilde{T}^{(d)})'$ is given in the next theorem.

Theorem 2. If $N \to \infty$ such that $N/n_i \to f_i < \infty$, then the vector $\widetilde{\mathbf{T}} = (\widetilde{T}^{(1)},\ldots,\widetilde{T}^{(d)})'$ follows, asymptotically, a multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix $\mathbf{R}$, where $\mathbf{R}$ is given in Theorem 1.

It follows from Theorem 2 that both the vectors $\mathbf{T}$ and $\widetilde{\mathbf{T}}$ have, asymptotically, as $N \to \infty$, the same joint distribution. Both the correlation matrices of $\mathbf{T}$ and $\widetilde{\mathbf{T}}$ asymptotically coincide due to the diagonal structure of $\mathbf{\Psi}$. Now, a simultaneous test procedure, which takes the correlation into account can be derived.

The individual null hypothesis $H_0^{(\ell)} : AUC^{(\ell)} \le AUC_0$ will be rejected at multiple level α, if

$$\widetilde{T}^{(\ell)} \ge z_{1-\alpha,1}(\widehat{\mathbf{R}}), \tag{16}$$

where $z_{1-\alpha,1}(\widehat{\mathbf{R}})$ denotes the one-sided equicoordinate quantile of the corresponding multivariate normal distribution where the correlation matrix $\mathbf{R}$ is replaced with the consistent estimator $\widehat{\mathbf{R}}$. One-sided simultaneous confidence intervals for $AUC^{(\ell)}$ are then given by

$$CI_{Logit}^{(\ell)} = \left[expit\left(g\left(\widehat{AUC}^{(\ell)}\right) - z_{1-\alpha,1}(\widehat{\mathbf{R}})\sqrt{\frac{\widehat{s}^{(\ell,\ell)}}{N}}\right), 1\right], \tag{17}$$

where $expit(y) = \frac{exp(y)}{1+\exp(y)}$ denotes the inverse Logit-transformation. The global null hypothesis $H_0 : \mathbf{AUC} \leq AUC_0 \cdot \mathbf{1}$ will be rejected, if $\max\left\{\widetilde{T}^{(1)}, \ldots, \widetilde{T}^{(d)}\right\} \geq z_{1-\alpha,1}(\widehat{\mathbf{R}})$, or if $\max\{CI^{(1)}_{Logit,l'} \ldots, CI^{(d)}_{Logit,l}\} > AUC_0$. Since the Logit-function is monotone, the procedure asymptotically controls the familywise error rate in the strong sense [26].

Small sample approximations with Wild Bootstrap

In the previous section approaches for the selection of diagnostic tests based on the AUC's have been derived. The procedures are based on the asymptotic joint distribution of the vectors $\mathbf{T}$ or $\widetilde{\mathbf{T}}$, respectively. The proposed approaches for selection of diagnostic tests are valid for large sample sizes. In order to investigate the accuracies of the procedures in terms of (i) controlling the pre-assigned type-I error level under the null hypothesis, (ii) maintaining the nominal coverage probability of the corresponding simultaneous confidence intervals, and (iii) their powers to detect certain alternatives, extensive simulation studies were conducted.

These simulation studies indicate, however, that both the statistics $\mathbf{T}$ in (12) and $\widetilde{\mathbf{T}}$ in (15) tend to result in liberal or conservative decisions in case of smaller sample sizes ($N \leq 100$) and larger AUC ($AUC \geq 0.8$). The results are in concordance with the simulation results proposed for univariate statistics by Kottas et al. [15] or Qin and Hotilovac [30]. Therefore, we propose a Wild Bootstrap approach to approximate their sampling distributions for small sample sizes.

Resampling procedures are widely known to be quite robust methods, even for small sample sizes. However, permutation methods cannot be used in this setup, since the distributions of the test statistics and the resampling statistics do not coincide, not even asymptotically (Pauly M, Asendorf T, Konietschke F: Permutation tests and confidence intervals for the area under the ROC curve, submitted). Simulation studies indicate that the use of the conventional Bootstrap from Efron [31] results in liberal conclusions, particularly when confronted with an $AUC \geq 0.7$ (see Table 1). Therefore, we did not further investigate the conventional Bootstrap. In contrast, the Wild Bootstrap approach ensures that the resampling distribution of the statistics mimics the distribution of $\mathbf{T}$ and $\widetilde{\mathbf{T}}$, asymptotically. The Wild Bootstrap technique is motivated by the residual bootstrap commonly applied in regression analysis [32-35], and in time-series testing problems [36-38]. It is also proposed in the context of survival analysis [39-42], and will be explained in the following.

Table 1 Empirical type-I error (theoretical 2.5%) of the normal Bootstrap for $d = 5$ and $N = 50$ with varying case-control-ratio and varying AUC

ccr	*AUC*				
	0.5	**0.6**	**0.7**	**0.8**	**0.9**
1 : 1	1.68%	2.28%	3.10%	5.00%	7.80%
1 : 4	1.90%	2.96%	4.70%	6.40%	12.10%

Let

$$\left(W_{01}, \ldots, W_{0n_0}, W_{11}, \ldots, W_{1n_1}\right) \tag{18}$$

denote independent and identically distributed random weights with $E(W_{is}) = 0$ and $Var(W_{is}) = 1$, which are independent of the data. We will investigate three different kinds of random weights W_{is} in our extensive simulation study:

- Rademacher weights: $P(W_{is} = 1) = P(W_{is} = -1) = \frac{1}{2}$.
- Standard normal weights: $W_{01}, \ldots, W_{1n_1} \sim N(0, 1)$.
- Uniform weights: $W_{01}, \ldots, W_{1n_1} \sim U\left[-\frac{\sqrt{12}}{2}, \frac{\sqrt{12}}{2}\right]$.

Let

$$\begin{aligned}\mathbf{Z}^*_{is} &= W_{is} \cdot \left(\mathbf{Z}_{is} - \overline{\mathbf{Z}}_{i\cdot}\right) \\ &= \left(W_{is} \cdot \left(Z^{(1)}_{is} - \overline{Z}^{(1)}_{i\cdot}\right), \ldots, W_{is} \cdot \left(Z^{(d)}_{is} - \overline{Z}^{(d)}_{i\cdot}\right)\right), \\ &\quad i = 0, 1,\ s = 1, \ldots, n_i,\end{aligned} \tag{19}$$

denote N resampling vectors, where $\mathbf{Z}_{is}$ is given in (8). Furthermore, let $\overline{\mathbf{Z}}^*_{i\cdot} = n_i^{-1} \sum_{k=1}^{n_i} \mathbf{Z}^*_{is} = \left(\overline{Z}^{*(1)}_{i\cdot}, \ldots, \overline{Z}^{*(d)}_{i\cdot}\right)'$ denote their means and let

$$\widehat{v}^{*(\ell,\ell)}_i = \frac{1}{n_i - 1} \sum_{s=1}^{n_i} \left(Z^{*(\ell)}_{is} - \overline{Z}^{*(\ell)}_{i\cdot}\right)^2$$

denote the empirical variance of $Z^{*(\ell)}_{i1}, \ldots, Z^{*(\ell)}_{in_i}$, $\ell = 1, \ldots, d$. In the next theorem it will be shown that the conditional resampling distribution of the vector

$$\begin{aligned}\mathbf{T}^* &= \left(T^{*(1)}, \ldots, T^{*(d)}\right)', \text{ where} \\ T^{*(\ell)} &= \sqrt{N} \frac{\overline{Z}^{*(\ell)}_{1\cdot} - \overline{Z}^{*(\ell)}_{0\cdot}}{\sqrt{\widehat{v}^{*(\ell,\ell)}_1 / n_0 + \widehat{v}^{*(\ell,\ell)}_0 / n_1}},\end{aligned} \tag{20}$$

mimics the distribution of both the vectors $\mathbf{T}$ and $\widetilde{\mathbf{T}}$, asymptotically.

Theorem 3. If $N \to \infty$ such that $\frac{N}{n_i}$ converges to some finite constant f_i, then the conditional distribution of $\mathbf{T}^*$ given the data $\mathbf{X}$ converges in probability to the multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix $\mathbf{R}$.

For proof see Additional file 1. Note that Theorem 3 is valid under the null as well as under the alternative, i.e., the resampling distribution mimics the distributions of $\mathbf{T}$ and $\widetilde{\mathbf{T}}$ for arbitrary values of $\mathbf{AUC} = \left(AUC^{(1)}, \ldots, AUC^{(d)}\right)'$. Next we will explain the computation of the simultaneous confidence intervals:

1. Given the data $\mathbf{X}$, compute the point estimators $\widehat{\mathbf{AUC}}$ and $\widehat{\mathbf{V}}_N$ as given in (5) and (9), respectively.
2. Generate $N = n_0 + n_1$ random weights $W_{01}, \ldots, W_{1n_1}$ as described in (18)
3. Compute $A_j^* := \max\{T^{*(1)}, \ldots, T^{*(d)}\}$ as given in (20).
4. Repeat the steps 2. - 3. *nboot* times (e.g. $nboot = 10,000$) and obtain the values $A_1^*, \ldots, A_{nboot}^*$.

5a. Compare each A_j^* with $\max\left\{\widetilde{\mathbf{T}}\right\}$. Then the individual p-value for $H_0^{(\ell)} : AUC^{(\ell)} \leq AUC_0$ is obtained from $\frac{1}{nboot}\sum_{j=1}^{nboot} \mathcal{I}\{\widetilde{T}^{(\ell)} \geq A_j^*\}$, where $\mathcal{I}\{\cdot\}$ denotes the indicator function.

5b. Estimate the quantile $z_{1-\alpha,1}(\mathbf{R})$ by the one-sided $(1-\alpha)$-quantile $z_{1-\alpha,1}^*$ of $A_1^*, \ldots, A_{nboot}^*$ to obtain the one-sided $(1-\alpha)$ simultaneous confidence intervals given by

$$CI_{WB}^{*(\ell)} = \left[expit\left(g\left(\widehat{AUC}^{(\ell)}\right) - z_{1-\alpha,1}^* \sqrt{\frac{\widehat{s}^{(\ell,\ell)}}{N}} \right), 1 \right]. \quad (21)$$

Results

Simulation results

We performed a simulation study to investigate the properties of the different approaches. All simulations were conducted with R environment, version 2.15.2. (R Development Core Team, 2010), each with 5, 000 simulation runs and 5, 000 bootstrap repetitions. The nominal type-I error was set to 2.5% one-sided and the global null hypothesis according to (4) was rejected, if at least one of the one-sided p-values was smaller than $\alpha = 2.5\%$. This means, the family wise error rate in the strong sense (FWER) is controlled, and the one-sided empirical type-I error should be closed to 2.5%. It is also possible to use the corresponding confidence intervals for decision. Then the global null hypothesis is rejected if the lower limit of at least one confidence interval was above AUC_0.

We generated multivariate normally distributed random vectors with compound symmetric correlation structure and defined the following scenario as standard scenario: a total sample size $N = 100$ with a case-control ratio (ccr) of $1:1$, $d = 5$ diagnostic tests and a correlation of $\rho = 0.9$ between the tests (motivated by [2,13,24]; and the example data set). The different parameters and conditions were varied afterwards as follows:

- The true AUC (0.5,..., 0.9)
- The number of diagnostic tests d (5, 10, 20)
- The total sample size N (50, 100, 200)
- The case-control ratio *ccr* (1:1, 1:2, 1:4, 1:9)
- The true correlation between the diagnostic tests ρ (0.3, 0.6, 0.9)
- The covariance structure in the data (compound symmetry, unstructured, and diagonal matrix with heterogeneous variances and positive or negative pairing)
- The distribution of the data (normal, skewed = log-normal, ordinal)

The different parameter constellations and all simulation results can be seen in the Additional file 2. Due to computational complexity, and its weak behavior in standard situations, we did not further investigate the conventional Bootstrap in our simulation study.

In a first step, this standard scenario was used for the comparison of the three random weights for the Wild Bootstrap: Rademacher (WB-Rade), standard normal (WB-Normal) and uniform (WB-Unif) weights. The results are displayed in the Additional file 3. For an AUC of 0.5 the three weights lead to nearly the same empirical type-I error and are quite conservative (empirical $\alpha \approx 0.015$). For larger AUC's the results are less conservative and for AUC's above 0.8 the empirical type-I error is around 2.5%. The Wild Bootstrap approach with uniform weights is, however, more conservative, while the standard normal and the Rademacher weights lead nearly to the same results. Therefore, and to present the simulation results more clearly, we only consider the standard normal weights in the following. The simulation results for the other weights are provided in the Additional file 2.

In practice often unadjusted (with the local type-I error α_0 equal to the global type-I error α) or Bonferroni adjusted confidence intervals for the single AUC's are used (see for example Shiotani et al. [43]). Therefore, in a second step, we compared these approaches (again for the standard scenario) using the multiple contrast test ('MCP'), the simultaneous Logit ('Logit') and the Wild Bootstrap ('WB-Normal') approach. In Figure 3 it becomes apparent that unadjusted intervals ('Unadj') lead to highly liberal conclusions (empirical type-I error 8 − 9%), while the Bonferroni correction ('Bonf') is too conservative (1.1 − 1.5%). Therefore we will not consider these approaches in the sequel. The MCP approach keeps the type-I error for an AUC of 0.5, but becomes more and more liberal for larger AUC's (up to 14% for $AUC = 0.9$). The empirical type I error of the Logit and the WB-Normal approach is comparable and between 1.5% and 2.9%. In the following we will investigate the influence of

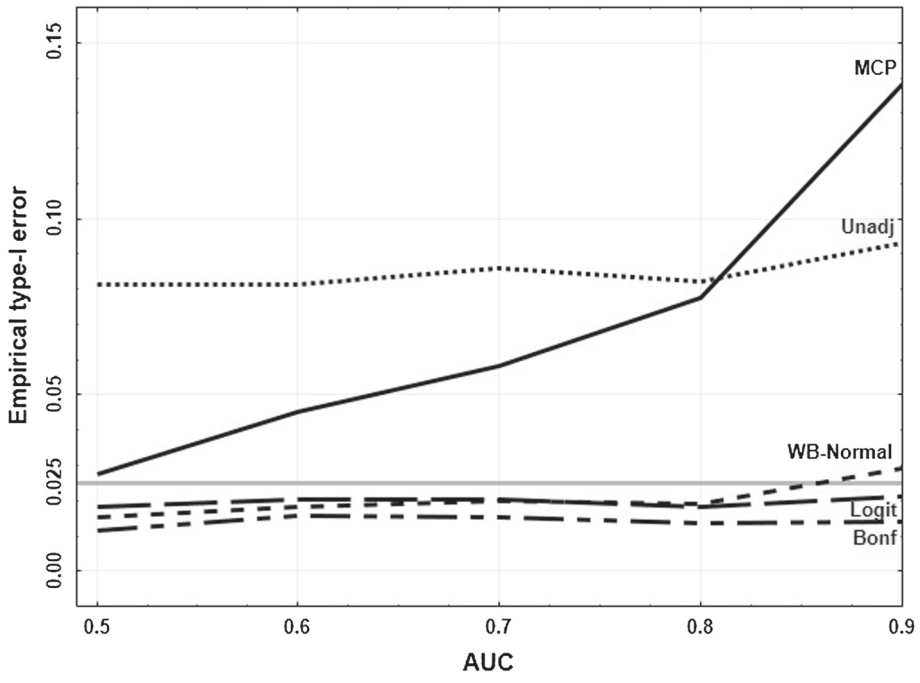

Figure 3 Empirical type-I error for varying AUC's. Empirical type-I error of the different approaches for the standard scenario (see text) with varying AUC's.

the different parameter settings on the type-I error of the Logit and the WB-Normal approach, and also of the MCP approach as the basis of both the approaches (despite of its liberal behavior).

The strength of the correlation, the type of the covariance structure and a skewed distribution do not have any impact on the behavior of the test (see figures and tables in the Additional files 2, 4, 5, and 6).

The impact of the sample size N and the number of diagnostic tests d is shown in Figure 4. As expected, for a larger sample size and a small number of diagnostic tests the type-I error is better exploited. As already seen in Figure 3 the Logit and the WB-Normal approach are comparable if $AUC \leq 0.8$ (independent of N and d). For larger AUC's, the WB-Normal approach leads to a larger empirical type-I error. On the one hand, this means that α is better exploited, on the other hand, this means that the results are liberal. The empirical type-I error of the Logit approach for $AUC = 0.9$ ranges from 1.3% to 2.1%, and of the WB-Normal approach from 2.2% to 2.9%.

If the case-control ratio (ccr) is not balanced, the empirical type-I error increases with increasing imbalance (see Figure 5). For an AUC of 0.8 or smaller both approaches are robust to an imbalance up to 1 : 4. For $AUC = 0.9$ the liberality of the WB-Normal approach is a disadvantage here, the empirical type-I error is above 2.5%. For a case-control ratio of 1 : 9, both approaches are far too liberal.

Ordinal data was generated using discretised normal distributions with a given AUC. For this data, representing a 5-point grading scale, the empirical type-I error decreases with increasing AUC ($AUC = 0.5$: Logit = 2.3%, WB-Normal = 2.2% to $AUC = 0.9$: Logit = 1.7%, WB-Normal = 1.6%). For details see Additional file 2.

The power was calculated for one example scenario ($N = 200$, $d = 5$, $ccr = 1 : 1$, $\rho = 0.9$, $AUC_0 = 0.7$), where the empirical type-I error of the Logit and of the WB-Normal approach was nearly the same. The true AUC is increasing from 0.7 (which is equal to AUC_0) to 0.85, according $\Delta AUC = 0, \ldots, 0.15$. The power of the two approaches is basically the same. For an ΔAUC of 0.1 (i.e. $AUC = 0.8$ vs. $AUC_0 = 0.7$) the power is greater than 80% (see Additional file 2).

Results for the analysis of the example

The point estimators for the AUC's are presented in the Background section in Figure 2. The number of 26 cases and 41 controls correspond to a case-control ratio of 1 : 1.6. The Spearman correlation coefficients between the biomarkers range from 0.64 to 0.95. For ΔI_{sum} the result was AUC=1. Because $logit(1) = \infty$, we modified the

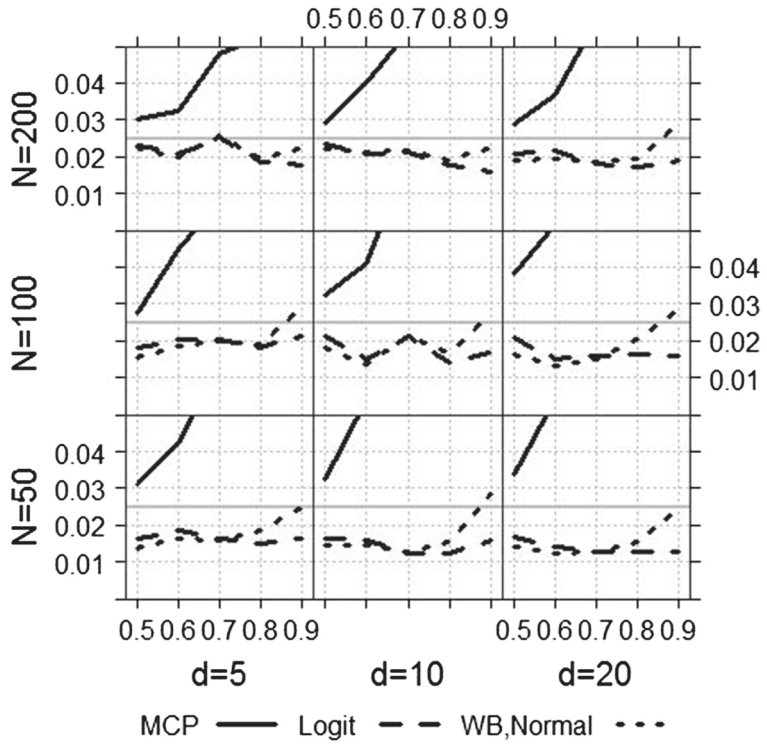

Figure 4 Empirical type-I error for varying *N* and *d*. Empirical type-I error of the MCP, the Logit and the WB-Normal approach for varying sample size and number of diagnostic tests.

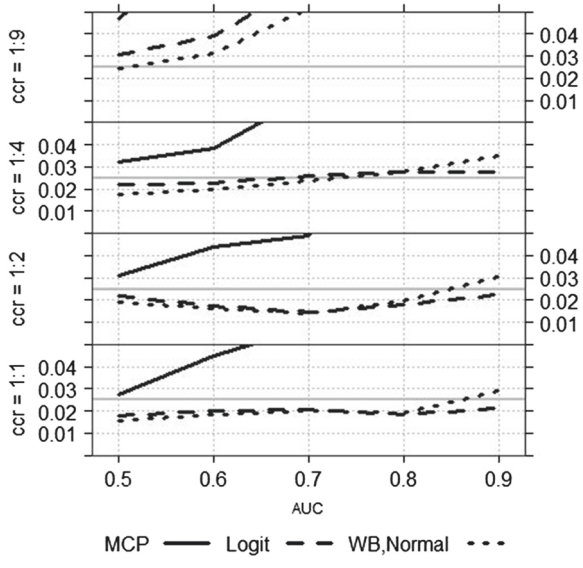

Figure 5 Empirical type-I error for varying ccr. Empirical type-I error of the MCP, the Logit and the WB-Normal approach for varying case-control ratios.

data for ΔI_{sum} such that we replaced the largest measurement of the controls with the smallest measurement of the cases. This minimal change leads to a point estimator for the AUC of 0.9999, and enables us to calculate the confidence intervals. This replacement strategy is conservative, since the effect is decreased, and the variance is increased. The one-sided 97.5% confidence intervals for all biomarkers using the MCP, the Logit, and the Wild Bootstrap approach are displayed in Figure 6. The results of the Wild Bootstrap with the three different weights differed just in the third decimal place. For consistency we displayed the WB-Normal approach here. The pattern of the results is the same for all four biomarkers. According to the simulation results, the MCP intervals are the shortest, the Logit intervals are the broadest, and the WB intervals are in between.

In the article of Derichs et al. [3] no threshold is defined. In Figure 6 four possible thresholds (0.8,0.85,0.9,0.95) are marked by solid horizontal lines. In Table 2 for each of these thresholds the numbers of selected biomarkers, depending on the individual approach, are listed. Apparently, the Logit approach is a more conservative selection criterion than the Wild Bootstrap approach. Although the MCP intervals are clearly shorter than the Wild Bootstrap intervals, the number of selected biomarkers is the same

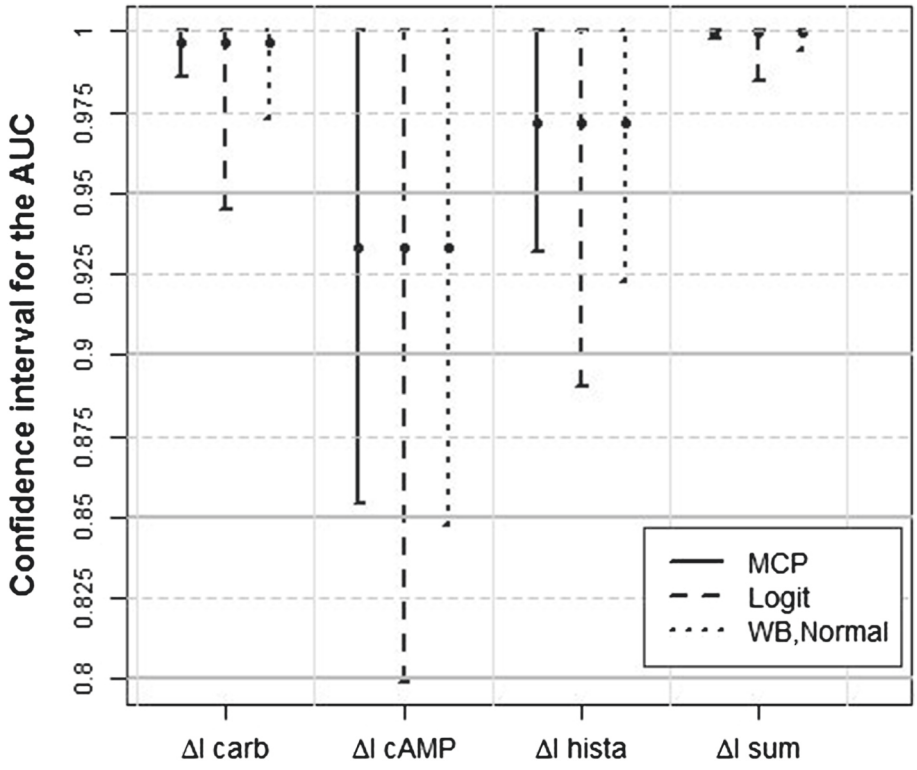

Figure 6 Confidence intervals for the biomarkers. One-sided 97.5% confidence intervals for the four biomarkers using the MCP, the Logit, and the WB-Normal approach.

for the MCP and the WB approach for three thresholds. Only for the threshold of 0.85 the MCP approach would select one biomarker more. Considering the simulation results of this section we would recommend to use the WB-Normal approach.

Discussion

It is widely discussed in the literature, whether the type-I error should be adjusted for multiplicity and whether the Bonferroni correction is an appropriate approach. Among many others, Wittes [44] states that lack of adjustment can lead to a misinterpretation of the study results as well as Bonferroni adjustment can do. Furthermore Perneger [45] states that "In summary, Bonferroni adjustments have, at best, limited applications in biomedical research, and should not be used when assessing evidence about specific hypotheses". Nevertheless, in practice often Bonferroni adjusted or even unadjusted confidence intervals for the single AUC's are used (see for example [43]). Konietschke et al. [10] proposed nonparametric multiple contrast tests and simultaneous confidence intervals for adequate correction of the type-I error, which take the dependencies within the data into account. Furthermore the authors recommended the transformation method (for example the Logit-transformation) to get less liberal results. However, Qin and Hotilovac [30] noticed that the Logit-transformed intervals are conservative for high accuracies. The reason is that the estimator $logit(\widehat{AUC})$ is quite unstable if $\widehat{AUC}$ is close to 0 or 1 because of a possibly larger variance. Obuchowski and Lieber [46] compared different confidence intervals for the AUC and concluded that for small sample sizes none of them provides adequate coverage for high accuracies.

Table 2 Number of selected biomarkers of the MCP, the Logit, and the WB-Normal approach for different thresholds (based on one-sided 97.5% confidence intervals)

Threshold	MCP	Logit	WB-Normal
0.8	4	3	4
0.85	4	3	3
0.9	3	2	3
0.95	2	1	2

Conclusion

In this article we derived a Wild Bootstrap approach, which exploits the type-I error much better than the Logit-approach, even for high accuracies and small

samples. Neither the strength of correlation, nor the structure of the covariance matrix, nor a skewed distribution, nor a moderate imbalanced case-control ratio has any impact on this desirable property of the Wild Bootstrap approach. Corresponding to these results we recommend to use the Wild Bootstrap approach with standard normally distributed weights for the selection of biomarkers in early diagnostic trials with the AUC as selection criterion.

Additional files

Additional file 1: Proof of Theorem 3.

Additional file 2: Tables of simulation results.

Additional file 3: Figure S1. Empirical type-I error of the Wild Bootstrap approach with the three different weights for the standard scenario (see article, Section "Simulation results") with varying AUC's.

Additional file 4: Figure S2. Empirical type-I error of the MCP, the Logit and the WB-Normal approach for varying strength of correlation.

Additional file 5: Figure S3. Empirical type-I error of the MCP, the Logit and the WB-Normal approach for different covariance structures (CS: compound symmetry, UN: unstructured, PP/NP: diagonal matrix with heterogeneous variances and positive/negative pairing).

Additional file 6: Figure S4. Empirical type-I error of the MCP, the Logit and the WB-Normal approach for normal and log-normal distributed data.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AZ, FK, and EB derived the Wild Bootstrap approach. AZ and FK performed the simulation study and wrote the article. EB revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the Federal Ministry of Education and Research [05M10MGB]. The authors thank Prof. Ballmann from the DRK-Kinderklinik Siegen for supporting this work by providing study data.

Author details
[1]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany. [2]Department of Mathematical Sciences, The University of Texas at Dallas, 800 W Campbell Road, 75080 Richardson, TX, USA.

Received: 15 October 2014 Accepted: 25 March 2015
Published online: 01 May 2015

References

1. DeLong E, DeLong D, Clark-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
2. Xia J, Broadhurst D, Wilson M, Wishart D. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics. 2013;9:280–99.
3. Derichs N, Sanz J, Von Kanel T, Stolpe C, Zapf A, Tümmler B, et al. Intestinal current measurement for diagnostic classification of patients with quastionable cystic fibrosis: validation and reference data. Thorax. 2010;65:594–9.
4. Marshall K, Mohr S, Khettabi F, Nossova N, Chao S, Bao W, et al. A blood-based biomarker panel for stratifying current risk for colorectal cancer. Int J Cancer. 2010;126:1177–86.
5. Broadhurst D, Kell D. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006;2:171–96.
6. EMA. Guideline on the choice of the non-inferiority margin. Doc. Ref. EMEA/CPMP/EWP/2158/99. 2005. www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500003636 (date of last access 13/04/15).
7. Phillips A, Fletcher C, Atkinson G, Channon E, Douiri A, Jaki T, et al. Multiplicity: discussion points from the statisticians in the pharmaceutical industry multiplicity expert group. Pharm Stat. 2013;12:255–9.
8. Strassburger K, Bretz F. Compatible simultaneous confidence bounds for the Holm procedure and other Bonferroni-based closed tests. Stat Med. 2008;27:4919–27.
9. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biometrical J. 2008;50:346–63.
10. Konietschke F, Hothorn L, Brunner E. Rank-based multiple test procedures and simultaneous confidence intervals. Electron J Stat. 2012;6:738–59.
11. Brunner E, Munzel U, Puri M. The multivariate nonparametric Behrens-Fisher problem. J Stat Planning Inference. 2002;108:37–53.
12. Bamber D. The area above the ordinal dominance graph and the area below receiver operating characteristic graph. J Math Psychol. 1975;12:387–415.
13. Janda S, Swiston J. Diagnostic accuracy of pleural fluid NT-pro-BNP for pleural effusions of cardiac origin: a systematic review and meta-analysis. BMC Pulmonary Med. 2010;10:58.
14. Wang L, Fahim M, Hayen A, Mitchell R, Baines L, Lord S. Cardiac testing for coronary artery disease in potential kidney transplant recipients. Cochrane Database Syst Rev. 2011;12:. DOI: 10.1002/14651858.CD008691.pub2.
15. Kottas M, Kuss O, Zapf A. A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. BMC Med Res Methodology. 2014;14:26.
16. Arlot S, Blanchard G, Roquain E. Some nonasymptotic results on resampling in high dimension, I: confidence regions. Ann Stat. 2010;38:51–82.
17. Kruskal W. A nonparametric test for the several sample problem. Ann Math Stat. 1952;23:525–40.
18. Lévy P. Calcul des Probabilitées. Paris: Gauthiers-Villars, Éditeurs; 1925.
19. Ruymgaart F. A unified approach to the asymptotic distribution theory of certain midrank statistics In: Raoult JP, editor. Statistique Non Parametrique Asymptotique vol. Lecture Notes on Mathematics, No. 821. Springer, Berlin Heidelberg; 1980. p. 1–18.
20. Munzel U. Linear rank score statistics when ties are present. Stat Probability Lett. 1999;41:389–95.
21. Brunner E, Puri M. Nonparametric methods in factorial designs. Stat Pap. 2001;42:1–52.
22. Kaufmann J, Werner C, Brunner E. Nonparametric methods for analysing the accuracy of diagnostic tests with multiple readers. Stat Methods Med Res. 2005;14:129–46.
23. Lange K, Brunner E. Sensitivity, specificity and ROC-curves in multiple reader diagnostic trials - a unified, nonparametric approach. Stat Methodology. 2012;9:490–500.
24. EMA. Guideline on clinical evaluation of diagnostic agents. Doc. Ref. CPMP/EWP/1119/98/Rev. 1. 2010. www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500003580 (date of last access 13/04/15).
25. Brunner E, Zapf A. Nonparametric ROC analysis for diagnostic trials In: Balkrishnan N, editor. Methods and Applications of Statistics in Clinical Trials vol. Volume 2: Planning, Analysis, and Inferential Methods. Hoboken, New Jersey: John Wiley & Sons; 2014. p. 471–83.
26. Gabriel K. Simultaneous test procedures - some theory of multiple comparisons. Ann Math Stat. 1969;40:224–50.
27. Bretz F, Landgrebe J, Brunner E. Multiplicity issues in microarray experiments. Methods Inf Med. 2005;44:431–7.
28. Zou G, Yue L. Using confidence intervals to compare several correlated areas under the receiver operating characteristic curves. Stat Med. 2012;32:5077–90.
29. Ferguson T. A Course in Large Sample Theory. London: Chapman & Hall; 1996.
30. Qin G, Hotilovac L. Comparison of non-parametric confidence interval for the area under the ROC curve of a continuous-scale diagnostic test. Stat Methods Med Res. 2008;17:207–21.
31. Efron B. Bootstrap methods: Another look at the Jackknife. Ann Stat. 1979;7:1–26.

32. Wu C. Jackknife, Bootstrap and other resampling methods in regression analysis. Ann Stat. 1986;14:1261–95.
33. Mammen E. When does Bootstrap work? Asymptotic results and simulations. New York: Springer; 1992.
34. Beran R. Diagnosing Bootstrap success. Ann Inst Stat Mathematics. 1997;49:1–24.
35. Janssen A. Nonparametric symmetry tests for statistical functionals. Math Methods Stat. 1999;8:320–43.
36. Kreiss J, Paparoditis E. Bootstrap for dependent data: a review, with discussion, and a rejoinder. J Korean Stat Soc. 2011;40:357–78.
37. Kreiss J, Paparoditis E. Bootstrapping locally stationary processes. J R Stat Soc - Ser B. 2014;77:267–90.
38. Konietschke F, Pauly M. Bootstrapping and permuting paired t-test type statistics. Stat Comput. 2014;24:283–96.
39. Lin D. Non-parametric inference for cumulative incidence functions in competing risks studies. Stat Med. 1997;16:901–10.
40. Beyersmann J, di Termini S, Pauly M. Weak convergence of the Wild Bootstrap for the Aalen-Johansen estimator of the cumulative incidence function of a competing risk. Scand J Stat. 2014;40:387–402.
41. Pauly M. Weighted resampling of martingale difference arrays with applications. Electron J Stat. 2011;5:41–2.
42. Dobler D, Pauly M. How to Bootstrap Aalen-Johansen processes for competing risks? Handicaps, solutions, limitations. Electron J Stat. 2014;8: 2779–803.
43. Shiotani A, Murao T, Kimura Y, Matsumoto H, Kamada T, Kusunoki H, et al. Identification of serum mirnas as novel non-invasive biomarkers for detection of high risk for early gastric cancer. Br J Cancer. 2013;109: 2323–30.
44. Wittes J. Clinical trials must cope better with multiplicity. Nat Med. 2012;18:1607.
45. Perneger T. What's wrong with Bonferroni adjustments. Br Med J. 1998;316:1236–8.
46. Obuchowski N, Lieber M. Confidence intervals for the receiver operating characteristic area in studies with small samples. Academic Radiology. 1998;5:561–71.