Expected *P*-values in Light of an ROC Curve Analysis Applied to Optimal Multiple Testing Procedures

Albert Vexler,[1,*] Jihnhee Yu[1], Yang Zhao[1], Alan D. Hutson[2] and Gregory Gurevich[3]

[1]Department of Biostatistics, The State University of New York, Buffalo, NY 14214, U.S.A.

[2]Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY 14263, U.S.A.

[3]Department of Industrial Engineering and Management, SCE-Shamoon College of Engineering, Beer Sheva 84100, Israel

**email:* avexler@buffalo.edu

Abstract

Many statistical studies report *p*-values for inferential purposes. In several scenarios, the stochastic aspect of *p*-values is neglected, which may contribute to drawing wrong conclusions in real data experiments. The stochastic nature of *p*-values makes their use to examine the performance of given testing procedures or associations between investigated factors to be difficult. We turn our focus on the modern statistical literature to address the expected *p*-value (EPV) as a measure of the performance of decision-making rules. During the course of our study we prove that the EPV can be considered in the context of receiver operating characteristic (ROC) curve analysis, a well-established biostatistical methodology. The ROC based framework provides a new and efficient methodology for investigating and constructing statistical decision-making procedures, including: (1) evaluation and visualization of properties of the testing mechanisms, considering, e.g., partial EPV's; (2) developing optimal tests via the minimization of EPV's; (3) creation of novel methods for optimally combining multiple test statistics. We demonstrate that the proposed EPV-based approach allows us to maximize the integrated power of testing algorithms with respect to various significance levels. In an application, we use the proposed method to construct the optimal test and analyze a myocardial infarction disease dataset. We outline the usefulness of the 'EPV/ROC' technique for evaluating different decision-making procedures, their constructions and properties with an eye towards practical applications.

Keywords

AUC; Benjamini-Hochberg procedure; Best combination; Bonferroni procedure; Bootstrap tilting method; Confidence region; Expected *p*-value; Multiple testing; *P*-value; Partial AUC; ROC curve.

1. Introduction

The *p*-value has long played a role in scientific research as a key decision-making tool with respect to hypothesis testing and dates back to Laplace in the 1770's (Stigler[1]). The concept of the *p*-value was popularized by Fisher[2] as an inferential tool and is where the first occurrence of the term "statistical significance" is found. In the frequentist hypothesis testing framework a test is deemed statistically significant if the *p*-value, which is a statistic having support over the real line in (0,1) space, is below some threshold known as the critical value. In a vast majority of studies the critical value is set at 0.05.

The obvious correct use of the *p*-value is to simply draw a conclusion of reject or do not reject the null hypothesis. This principle simplifies and standardizes statistical decision-making policies. However, the *p*-value is oftentimes misused and misinterpreted in the applied scientific literature where statistical decision-making procedures are involved. Many scientists misinterpret smaller *p*-values as providing stronger evidence against a null hypothesis relative to larger *p*-values. For example, some researchers draw conclusions regarding comparisons of associations between a disease and different factors using values of the corresponding *p*-values. In a hypothetical study, consider evaluating associations between a disease, say D, and two biomarkers, say A and B. It is not uncommon for scientists to conclude that the association between D and A is stronger than that between D and B if the *p*-value regarding the association between A and D is smaller than that of the association between B and D. This example demonstrates the non-careful use of the *p*-value's concept, since perhaps data obtained in a different but relevant experiment might provide the contradicting conclusion simply due to the stochastic nature of the *p*-value. These types of issues have led several scientific journals to discourage the use of *p*-values, with some scientists and statisticians encouraging their abandonment[3]. For example, the editors of the journal entitled *Basic and Applied Social Psychology* announced that the journal would no longer publish papers containing *p*-values-based studies since the statistics were too often used to support lower-quality research[4].

The *p*-value is a function of the data and hence it is a random variable, which too has a probability distribution. The subtlety in terms of those that try to interpret the magnitude of the relative *p*-value is that the distribution of the *p*-value is conditional on either the null hypothesis being true or not. Under the null hypothesis, and assuming no nuisance parameters, *p*-values have a Uniform[0,1] distribution. However, if the null hypothesis is false *p*-values have a non-Uniform[0,1] distribution for which the shape of the distribution varies across several factors including sample size and the distance of the parameter of interest from the hypothesized value (null). Hence, for the same exact null and alternative values the distribution of the *p*-value may be small or large simply as a function of the sample size (statistical power). In the era of "big data" it

would not be unusual to constantly find extremely small p-values simply as function a massively large sample size, but having nothing really to do with scientific question. Likewise a large observed *p*-value may simply be due to a very small sample size.

Statisticians have long recognized the deficiencies in terms of interpreting *p*-values relative to their stochastic nature and have tried to develop remedies to aid scientists in the interpretation of their data. For example, Lazzeroni et al.[5] developed prediction intervals for *p*-values in replication studies. This approach has certain critical points regarding the following problems. 1) In the frequentist context it is uncommon to create confidence intervals of random variables; 2) Under the null hypothesis *p*-values are distributed according to a Uniform[0,1] distribution, whereas in many scenarios, if we are sure the alternative hypothesis is in effect, the prediction interval for the *p*-value is not needed; 3) Prediction intervals for *p*-values can be directly associated with those for corresponding test statistics values, linking to just rejection sets of the test procedures.

The stochastic aspect of the *p*-value has been well studied by Dempster and Schatzoff[6] and Schatzoff[7] who introduced the concept of the *expected significance level*. Sackrowitz and Samuel-Cahn[8] developed the approach further and renamed it as the *expected p-value* (EPV). The authors presented the great potential of using EPVs in various aspects of hypothesis testing.

Comparisons of different test procedures, e.g. a Wilcoxon rank-sum test versus Student's t-test, based on their statistical power is oftentimes problematic in terms of deeming one method being the preferred test over a range of scenarios. One reason for this issue to occur is that the comparison between two or more testing procedures is dependent upon the choice of a pre-specified significance level α. One test procedure may be more or less powerful than the other one depending on the choice of α. Alternatively, one can consider the EPV concept in order to compare test procedures. In this paper, we show that the EPV corresponds to the integrated power of a test via all possible values of $\alpha \in (0,1)$. Thus, the performance of the test procedure can be evaluated globally using the EPV concept. Smaller values of EPV show better test qualities in a more universal fashion. This method is an alternative approach to the Neyman-Pearson concept of testing statistical hypotheses. In this paper, we present a framework for optimal decision making criteria based on the EPV. The famous Neyman-Pearson lemma[9] (also see Vexler et al.[10] for details) introduced us to the concept that a reasonable statistical testing procedure controls the Type I error rate at a pre-specified significance level,α, in conjunction with maximizing the power in a uniform fashion. Thus, for different values of α we may obtain different superior test procedures. On the other hand, the EPV based approach allows us to compare between decision-making rules in a more objective

manner. The global test performance of testing procedures can be measured by one number, the EPV, and hence tests can be more easily rank-ordered.

In this paper we further advance the concept of the EPV. We prove that there is a strong association between the EPV concept and the well-known receiver operating characteristic (ROC) curve methodology. The ROC curve technique is a very common biostatistical tool for describing the accuracy of different biomarkers in terms of predicting or diagnosing diseases. The area under the ROC curve (AUC) is a global summary index for measuring the diagnostic ability of a biomarker or combination of biomarkers to predict or diagnose disease[10-15]. It turns out that we can use well-established ROC curve and AUC methods to evaluate and visualize the properties of various decision-making procedures in the *p*-value based context. Further, we develop a ***partial expected p-value*** (pEPV) and introduce a novel method for visualizing the properties of statistical tests in an ROC curve framework. The 'ROC/EPV' framework is proposed to solve multiple testing problems (e.g., Dmitrienko et al.[16]).

Various experiments require rigorous statistical analyses involving the evaluation of sets consisting of more than one hypothesis. For example, in a case-control study, investigators may have expression levels of several thousand biomarkers measured to discriminate cases (disease) from controls (healthy). One may be interested in considering the discriminability of individual or different subsets of the biomarkers. An interesting issue is how one might combine the test-statistics for testing the discriminability of certain sets of biomarkers. In this case, we should take into account that biomarkers can be dependent as well as the fact that their values may be measured on different scales. One approach towards addressing the multiple testing problem is to use the classical Bonferroni method or Benjamini-Hochberg procedure (e.g., Benjamini and Hochberg[17]). In this paper, we propose a novel EPV method to combine different test statistics in the multiple testing framework. This approach is based on a principle of maximization of AUCs or partial AUCs. We show that in many scenarios the proposed methods outperform both the classical Bonferroni and Benjamini–Hochberg approaches. The novel EPV-based technique can be used to estimate confidence regions of a set of vector parameters based on the confidence intervals for each of the respective vector components.

This paper has the following structure: In Section 2, we define the EPV in the context of the ROC curve analysis. In Section 3, we demonstrate the common multiple testing issues and the current state-of-the-art solutions. We then propose novel methods for multiple testing problems that are shown to be superior in many instances. In Section 4, we present several examples to illustrate the proposed method as applied to the multiple testing problem. In Section 5 we carry out a Monte Carlo simulation study to evaluate the EPVs and the powers of the proposed decision-making

mechanisms as compared to the classical Bonferroni family-wise error rate and Benjamini–Hochberg false discovery rate approaches. Section 6 focuses on a real data example from a biomarker study associated with myocardial infarction (MI) disease, which shows the practicality and adaptability of the new method. In section 7, we present concluding remarks.

2. The EPV in the Context of an ROC Curve Analysis

In this section we present the following material: The formal definition of the EPV; an overview of ROC curves; and the association between the EPV and the AUC. We also provide a new quantity called the partial EPV (pEPV), which characterizes a property of decision-making procedures using the concept of partial AUCs.

2.1 *The Expected P-Value*

Let the random variable $T(D)$ represent a test statistic depending on data *D*. Assume F_i defines the distribution function of $T(D)$ under the hypothesis $H_i, i = 0,1$, where the subscript *i* indicates the null (*i=0*) and alternative (*i=1*) hypotheses, respectively. Given F_i is continuous we can denote F_i^{-1} to represent the inverse or quantile function of F_i, such that, $F_i(F_i^{-1}(\gamma)) = \gamma$, where $0 < \gamma < 1$ and *i=0,1*. In this setting, in order to concentrate upon the main issues, we will only focus on tests of the form: the event $T(D) > C$ rejects H_0, where *C* is a prefixed test threshold. Thus the *p*-value has the form $1 - F_0(T(D))$. Sackrowitz and Samuel-Cahn[8] proved that the expected *p*-value $E(1 - F_0(T(D)) \mid H_1)$ is

$$\text{EPV} = \Pr(T^0 \geq T^A), \tag{1}$$

where independent random variables T^0 and T^A are distributed according to F_0 and F_1, respectively. The simple example of the EPV is when $T^0 \sim N(\mu_1, \sigma_1^2)$ and $T^A \sim N(\mu_2, \sigma_2^2)$. Then the EPV can be expressed as

$$\text{EPV} = \Phi\left((\mu_1 - \mu_2)\left(\sigma_1^2 + \sigma_2^2\right)^{-1/2}\right)$$

where Φ is the cumulative standard normal distribution.

Note that the formal notation (1) is similar to that of the area under ROC curve. In this context, one can reconsider the EPV in terms of the area under ROC curve. In the next section, we outline the basic concepts of the ROC curve analysis.

2.2 *The ROC Curve and AUC based approach*

In biomedical research, a biomarker is frequently defined as a distinctive biological or biologically derived indicator of disease. For example, the prostate-specific antigen (PSA) biomarker is applied to diagnose prostate cancer; cardiac imaging biomarkers may be used to diagnose heart disease; the hemoglobin A1c (HbA1c) biomarker is known to be useful for diagnosing diabetes. The ROC curve analysis is an efficient approach for evaluating the discriminability of biomarkers. An ROC curve of a biomarker is a plot of its sensitivity (true positive rate) versus 1 minus its specificity (false positive rate). For excellent reviews of statistical methods involving ROC curves and their applications, we refer the reader to Zou et al.[11], Liu and Schisterman[12], Pepe[18], Zhou et al.[15] and Vexler et al.[10]. The AUC is a popular measure of the performance of a biomarker, with larger value of the area indicating a more accurate discriminating ability of a given marker (e.g., Liu et al.[19]; Vexler et al.[14]). Bamber[20] proved that the AUC can be expressed in the form

$$\mathrm{AUC} = \int_0^1 ROC(t)dt = \Pr(Y_D > Y_{\bar{D}}), \tag{2}$$

where $ROC(t) = 1 - F_{Y_D}\{F_{Y_{\bar{D}}}^{-1}(t)\}, 0 < t < 1,$ denotes the ROC curve, random variables Y_D and $Y_{\bar{D}}$ are from the distribution functions F_{Y_D} and $F_{Y_{\bar{D}}}$ that correspond to biomarker's measurements from diseased (D) and non-diseased ($\bar{D}$) subjects, respectively. Thus the AUC mechanism provides a convenient way to compare diagnostic biomarkers because the ROC curve places measurements for each biomarker on the same scale where they can be individually evaluated for accuracy.

The partial area under the ROC curve (pAUC) is the area under a portion of the ROC curve, oftentimes defined as the area between two false positive rates (FPRs). For example, the pAUC with two fixed *a priori* values for FPRs t_0 and t_1 is

$$pAUC = \int_{t_0}^{t_1} ROC(t)dt = \Pr\left\{Y_D > Y_{\bar{D}}, Y_{\bar{D}} \in \left(S_{Y_{\bar{D}}}^{-1}(t_1), S_{Y_{\bar{D}}}^{-1}(t_0)\right)\right\},$$

where S_{Y_D} and $S_{Y_{\bar{D}}}$ are the survival functions of the diseased and healthy group, respectively. To simplify this notation, we denote $q_0 = S_{\bar{D}}^{-1}(t_0)$ and $q_1 = S_{\bar{D}}^{-1}(t_1)$. Then

$$pAUC = \Pr\left\{Y_D > Y_{\bar{D}}, Y_{\bar{D}} \in (q_0, q_1)\right\}.$$

2.3 *The association between EPV-based characteristics and ROC curve methodology*

The area under the ROC curve is 1-EPV, which can be shown by (1) and (2). This connection between the EPV and the AUC induces new techniques for evaluating statistical test qualities via the well-established ROC curve methodology. Consider, for illustrative purposes, the following example related to applications of Student's and Welch's *t*-tests. The recent biostatistical literature

has extensively discussed which test, Student's *t*- or Welch's *t*-test, to use in practical applications. The questions in this setting are: What is the risk of using Student *t*-test when variances of the two populations are different as well as what is a loss in power when using Welch's *t*-test when the variances of the two populations are equal[21,22]? In order to apply the ROC curve analysis based on the EPV-concept, we denote Student's *t*-test statistic as

$$T_S = \left(\bar{X} - \bar{Y}\right)\left[S_p^2 \left(n^{-1} + m^{-1}\right)^{1/2} \right]^{-1},$$

and Welch's *t*-test statistic as

$$T_W = \left(\bar{X} - \bar{Y}\right)\left(S_1^2 n^{-1} + S_2^2 m^{-1}\right)^{-1/2},$$

where $\bar{X}$ is the sample mean based on the independent normally distributed data points $X_1,...,X_n$, $\bar{Y}$ is the sample mean based on the independent normally distributed observations $Y_1,...,Y_m$, $S_1^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ and $S_2^2 = \sum_{j=1}^{m}(Y_i - \bar{Y})^2/(m-1)$ are the unbiased estimators of the variances $\sigma_1^2 = Var(Y_1)$ and $\sigma_2^2 = Var(Y_2)$ respectively and $S_p^2 = \{(n-1)S_1^2 + (m-1)S_2^2\}/(n+m-2)$ is the pooled sample variance. Figure 1 depicts the ROC curves, $ROC(t) = 1 - F_1\{F_0^{-1}(t)\}, t \in (0,1)$, for each test when the distribution functions F_0, F_1 are correctly specified corresponding to underlying distributions of observations with $\delta = EX_1 - EY_1 =$ 0.7 and 1 under H_1. These graphs show that there are no significant differences between the relative curves. In the scenario $n = 10$, $m = 50$, $\sigma_1^2 = 4$, $\sigma_2^2 = 1$ Student's *t*-test demonstrates better performance than that of Welch's *t*-test. By using different values of *n*, *m*, σ_1^2, σ_2^2, δ, we attempt to detect cases when significant differences between the relevant curves are in effect. Applying different combinations of $n = 10, 20, ..., 150$; $m = 10, 20, ..., 150$; $\sigma_1^2 = 1, 2^2, ..., 10^2$ and $\sigma_2^2 = 1$, we could not derive a scenario when one of the considered tests clearly outperforms the other one with respect to the EPV. The corresponding AUC (1-EPV) values are given in Table 1. Thus, if the Type I error rates of the tests are correctly controlled, there are no critical differences between Student's *t*-test and Welch's *t*-test.

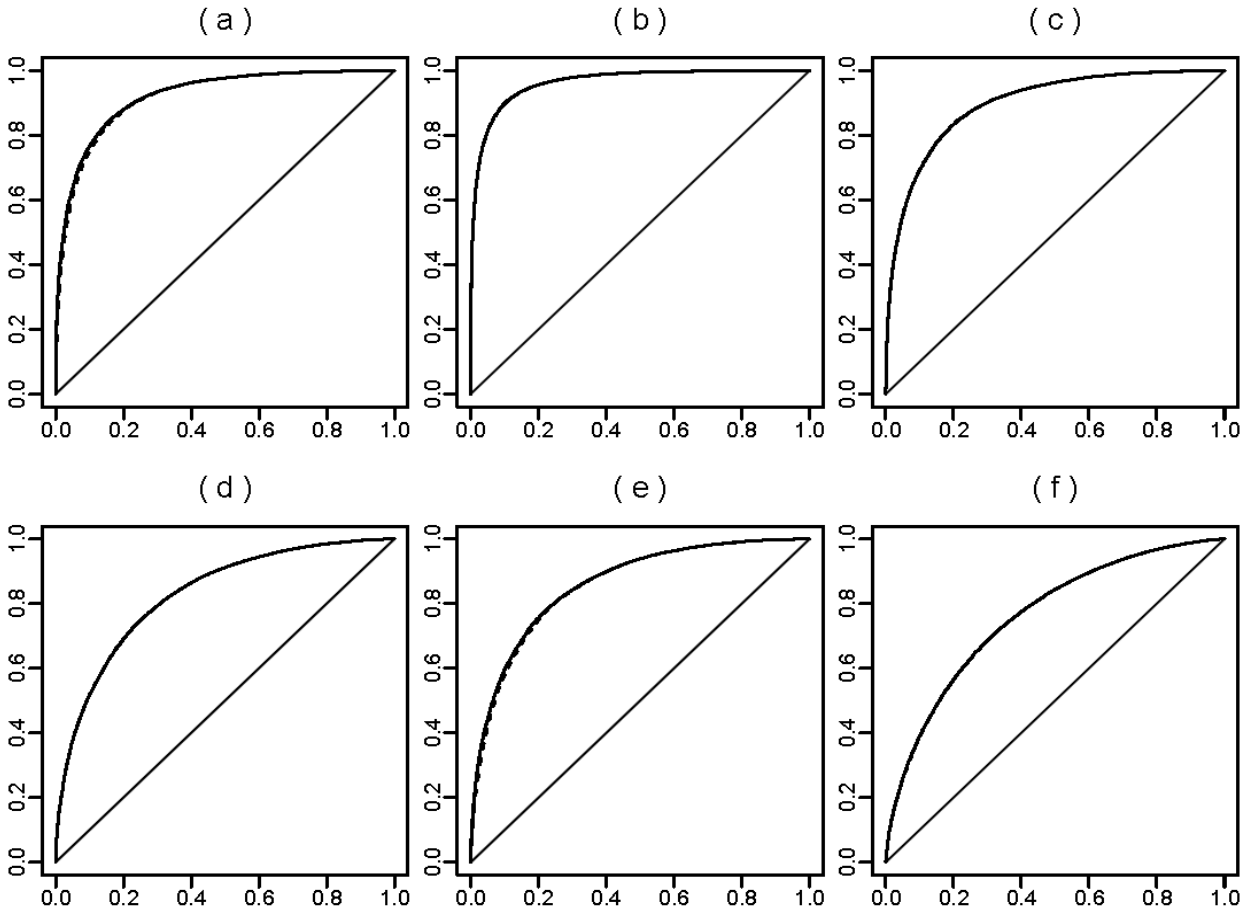

Figure 1. The ROC curves related to the Student's *t*-test ("——") and the Welch's *t*-test ("----"), where panel (a) represents the case of $n=10$, $m=50$, $\sigma_1^2=1$, $\sigma_2^2=1$, $\delta=0.7$; graph (b) represents the case of $n=20$, $m=40$, $\sigma_1^2=1$, $\sigma_2^2=1$, $\delta=0.7$; graph (c) represents the case of $n=40$, $m=20$, $\sigma_1^2=4$, $\sigma_2^2=1$, $\delta=0.7$; graph (d) represents the case of $n=40$, $m=20$, $\sigma_1^2=9$, $\sigma_2^2=1$, $\delta=0.7$; graph (e) represents the case of $n=10$, $m=50$, $\sigma_1^2=4$, $\sigma_2^2=1$, $\delta=1$ and graph (f) represents the case of $n=20$, $m=40$, $\sigma_1^2=9$, $\sigma_2^2=1$, $\delta=0.7$.

Table 1. The areas under the ROC curves of the Student's *t*-test (AUC_S) and Welch's *t*-test (AUC_W)

n	m	σ_1^2	σ_2^2	δ	AUC_S	AUC_W
10	50	1	1	0.7	0.9217	0.9172
20	40	1	1	0.7	0.9619	0.9611
40	20	4	1	0.7	0.8959	0.8962
40	20	9	1	0.7	0.8269	0.8271
10	50	4	1	1.0	0.8614	0.8569
20	40	9	1	0.7	0.7631	0.7626

In the next section, we define the pEPV concept using test-power aspects and the pAUC approach.

2.4 *The Connection between EPV and Power*

The value of the 1-EPV can be expressed in the form of the statistical power of a test through integration uniformly over the significance level α from 0 to 1; that is,

$$EPV = \Pr(T^0 \geq T^A) = \int_{-\infty}^{\infty} \Pr(T^A \leq t) dF_0(t) = \int_{-\infty}^{\infty} \Pr\left\{F_0(T^A) \leq F_0(t)\right\} dF_0(t) \tag{3}$$

$$= \int_1^0 \Pr\left\{1 - F_0(T^A) \geq \alpha\right\} d(1-\alpha) = \int_0^1 \left[1 - \Pr\left\{1 - F_0(T^A) \leq \alpha\right\}\right] d\alpha = 1 - \int_0^1 \Pr(p - \text{value} \leq \alpha \mid H_1) d\alpha.$$

The above expression of the EPV considers the weight of the significance level α from 0 to 1. It may appear to suffer from the defect of assigning most of its weight to relatively uninteresting values of α not typically used in practice, e.g. $\alpha \geq 0.1$. Alternatively, we can focus on significance levels of α in a specific interesting range by considering the partial expected *p*-value (pEPV); that is

$$pEPV = 1 - \int_0^{\alpha_U} \Pr\left\{p - value \leq \alpha\right\} d\alpha = 1 - \int_0^{\alpha_U} \Pr\left\{1 - F_0(T^A) \leq \alpha\right\} d\alpha \tag{4}$$

$$= 1 + \int_0^{\alpha_U} \Pr\left\{F_0(T^A) \geq 1 - \alpha\right\} d(1-\alpha) = 1 + \int_1^{1-\alpha_U} \Pr\left\{F_0(T^A) \geq z\right\} dz$$

$$= 1 - \int_{1-\alpha_U}^{1} \Pr\left\{F_0(T^A) \geq z\right\} dz = 1 - \int_{F_0^{-1}(1-\alpha_U)}^{\infty} \Pr\left\{F_0(T^A) \geq F_0(t)\right\} dF_0(t)$$

$$= 1 - \int_{F_0^{-1}(1-\alpha_U)}^{\infty} \Pr\left\{T^A \geq t\right\} dF_0(t) = 1 - \Pr\left\{T^A \geq T^0, T^0 \geq F_0^{-1}\left(1-\alpha_U\right)\right\}$$

at a fixed upper level $\alpha_U \leq 1$.

Remark: The Neyman-Pearson lemma framework for comparing, for example, two test-statistics, say M_1 and M_2, provides the following scheme: the Type I error rates of the tests should be fixed at a pre-specified significance level α, $\Pr(M_1 \text{ rejects } H_0 \mid H_0) \leq \alpha$ and $\Pr(M_2 \text{ rejects } H_0 \mid H_0) \leq \alpha$; then M_1 is superior with respect to M_2, if $\Pr(M_1 \text{ rejects } H_0 \mid H_1) > \Pr(M_2 \text{ rejects } H_0 \mid H_1)$. In general the power functions, $\Pr(M_1 \text{ rejects } H_0 \mid H_1)$ and $\Pr(M_2 \text{ rejects } H_0 \mid H_1)$, depend on α. Thus, for different values of α we may theoretically obtain diverse conclusions regarding the preferable test procedure. The EPV and pEPV concepts make the comparison more objective in a global sense. The test performance can be measured employing just the EPV or pEPV value by itself. Smaller values of the EPV or pEPV indicate more preferable test-procedure when comparing two or more tests. Equations (3) and (4) show that for a uniformly most powerful test (e.g., the likelihood ratio test), the EPV and pEPV will be the minimum as compared to any other tests with the same H_0 vs. H_1.

3. Multiple Testing Problems

In practice, biostatistical experiments can focus on several hypothesis tests. A current example is to test for differences in gene expression profiles between healthy and diseased populations across potentially thousands of tests. In the multiple testing task the concept of the family-wise experimental error rate (controlling the false positive rate across all tests) or the false discovery rate (a method to boost power by allowing a higher "known" proportion of false-discoveries to be declared statistically significant) needs to be considered in conjunction with the per comparison false positive rate. Multiple testing schemes adjust statistical inferences in an experiment for multiplicity by considering control of the family-wise error rate or false discovery rate at level α and therefore control the overall number of false-positive results depending upon the methodology. For an excellent review of statistical methods involving multiple testing problems, see Dmitrienko et al.[16]

In this section, in the context of the union-intersection test problems, we outline the classical Bonferroni procedure and the Benjamini-Hochberg (BH) approach. The new proposed method for multiple hypothesis testing problems is presented via the ROC based methodology for optimally combining biomarkers.

3.1 *The union-intersection test*

Oftentimes, multiple hypothesis testing problems are stated as union-intersection problems[23]. For example, assume *m* biomarkers are involved to test for their individual ability to diagnose a disease. Towards this end, biomarkers measurements are obtained from case and control populations and used to conduct and record values of *m* test statistics $B_1,...,B_m$ for the hypotheses $H_{01},...,H_{0m}$ against the alternative hypotheses $H_{11},...,H_{1m}$, respectively. The next study aim can focus on the global hypothesis H_0 defined as the intersection of the hypotheses tested versus the union of the alternative hypotheses (H_1): $H_0 : \bigcap_{i=1}^{m} H_{0i}$ vs. $H_1 : \bigcup_{i=1}^{m} H_{1i}$, where $i = 1,...,m$. Moreover, objectives of the study can be related to a subset, say *S*, of the biomarkers that require analyzing the hypothesis $H_0^S : \bigcap_{i \in S} H_{0i}$ vs. $H_0^S : \bigcup_{i \in S} H_{1i}$ .To address this problem correct decision-making algorithms based on $B_1,...,B_m$ need to be applied.

In modern scientific experiments, many large-scale hypotheses testing problems involve thousands of hypotheses as a joint family of interest. For example, in a DNA microarray experiment one may be interested in comparing the expression levels of a large number of genes in diseased subjects versus those in healthy subjects. The main goal of the experiment is to find a small group

of "interesting" genes among the numerous genes whose expression levels differ between the diseased group and healthy group.

It is clear that the setting considered above can be associated with a problem to estimate confidence regions of vector-parameters based on confidence interval estimates of their respective elements.

3.2 *The Bonferroni and the BH procedures*

Perhaps, the Bonferroni procedure is one of the most widely used methods for addressing multiple testing problems. To illustrate the Bonferroni procedure let p_i be the unadjusted *p*-value for testing the individual null hypotheses $H_{0i}, i = 1,...,m$. Then each p_i can be considered as a test-statistic for the respective $H_{0i}, i = 1,...,m$. Thus, we can develop a test statistic for $H_0 : \bigcap_{i=1}^{m} H_{0i}$ vs. $H_1 : \bigcup_{i=1}^{m} H_{1i}$ by constructing a decision-making rule based on $p_1,..., p_m$ with a Type I error rate α. The random variables $p_1,..., p_m$ are dependent in general. In this framework, the Bonferroni scheme rejects H_0 if the events $\{p_i \le \alpha / m\}$ are detected for all $i \in \{1,...,m\}$. This procedure is a very general method, which does not require any distributional assumptions and has a computational ease. However, the procedure tends to be conservative if the number of hypotheses is large or the test statistics are strongly correlated.

Benjamini and Hochberg[17] introduced a novel Bonferroni-type multiple adjusting procedure, which controls the false discovery rate (FDR) for a fixed value $q \in (0,1)$. We refer the reader to the multiple testing problems literature for details regarding the FDR definition and its interpretations. The BH procedure is based on $p_{(1)} \le p_{(2)} \le ... \le p_{(m)}$, the ordered *p*-values, where $p_{(i)}$ corresponds to the hypothesis $H_{0(i)}$. The procedure rejects all $H_{0(i)}, i = 1,...,k$, where $k = \max\{k \in \{1,...,m\} : p_{(i)} \le iq / n\}$, controlling the FDR at q.

3.3 *Combinations of biomarkers*

When multiple biomarkers are available, it is of common interest to combine the biomarkers to improve the diagnostic accuracy of a clinical test. One approach for doing this is to maximize the AUC based on functional combinations of biomarker measurements. This idea is employed as a basis for developing a novel method for combining the test statistics with respect to minimization of the EPV.

Let us assume that a number of K biomarkers are available and the random vector $\boldsymbol{M} = (M_1,..., M_K)^T$ represents the levels of the given biomarkers. The expression level of the biomarkers is denoted by $\boldsymbol{X} = (X_1,..., X_K)^T$ for the disease group and the expression level of

biomarkers is denoted by $\boldsymbol{Y} = (Y_1,...,Y_K)^T$ for the healthy group, with corresponding density functions $f = f(x_1,...,x_K)$ and $g = g(y_1,...,y_K)$, respectively. According to the Neyman-Pearson lemma, when the density functions f and g can be correctly specified, the likelihood ratio function,

$$LR(\mathbf{M}) = f(\mathbf{M}) / g(\mathbf{M})$$

yields the $AUC = \Pr\{LR(\mathbf{X}) \geq LR(\mathbf{Y})\}$ that is larger than the AUCs of any other combination of the biomarkers[18,24].

In practice the density functions f and g are generally unknown. In this case, combining multiple biomarkers using linear functions[25] is very popular due to its simplicity and the acceptability to clinicians[26]. Linear combination of biomarkers may be written as

$$l(\boldsymbol{\lambda}; \boldsymbol{M}) = \boldsymbol{\lambda}^T \boldsymbol{M} = M_1 + \lambda_2 M_2 + ... + \lambda_K M_K,$$

where $\boldsymbol{\lambda} = (1, \lambda_2,..., \lambda_K)^T$ is a K dimensional vector. The corresponding AUC has the form

$$AUC(\boldsymbol{\lambda}) = \Pr\{l(\boldsymbol{\lambda}; \boldsymbol{X}) > l(\boldsymbol{\lambda}; \boldsymbol{Y})\}.$$

The best linear combination (BLC) maximizes the AUC using the coefficient

$$\boldsymbol{\lambda}_0 = \arg\max_{\lambda} \{AUC(\boldsymbol{\lambda})\} = \arg\max_{\lambda} \Pr\left(\boldsymbol{\lambda}^T \mathbf{X} > \boldsymbol{\lambda}^T \mathbf{Y}\right).$$

Assuming the biomarkers of the healthy group and disease group follow normal distributions, Su and Liu[25] derived the best linear combination that yields the largest AUC. If the normality assumption is not met, Pepe and Thompson[27] and Chen et al.[26] have proposed nonparametric solutions to estimate the best linear combination. For example, one can maximize the Mann-Whitney U-statistic, an empirical estimate of AUC, by considering each linear combination $l(\boldsymbol{\lambda}, \boldsymbol{M})$, via the value of

$$AUC_e(\boldsymbol{\lambda}) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left\{l\left(\boldsymbol{\lambda}; X_i\right) \geq l\left(\boldsymbol{\lambda}; Y_j\right)\right\} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left\{\boldsymbol{\lambda}^T \left(X_i - Y_j\right) \geq 0\right\},$$

where I is the indicator function. Then the empirical best linear combination coefficient $\boldsymbol{\lambda}_e$ is

$$\boldsymbol{\lambda}_e = \arg\max_{\lambda} AUC_e(\boldsymbol{\lambda}) = \arg\max_{\lambda} \left[\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left\{\boldsymbol{\lambda}^T \left(X_i - Y_j\right) \geq 0\right\}\right].$$

The link between the notation of the EPV (1) and that of the AUC (2) leads us to use the methods regarding the combinations of biomarkers for developing the new method for combining the test statistics while minimizing the EPV.

3.4 ***Combinations of test statistics minimizing the EPV***

For simplicity, suppose that in a multiple test problem there are only two hypotheses H_{01} and H_{02} that need to be tested against the alternative hypotheses H_{11} and H_{12}. Assume we are interested in the union-intersection problem $H_0 : H_{01} \cap H_{02}$ vs. $H_1 : H_{11} \cup H_{12}$ and then the statistics $T_i, i = 1,2,$ are used to test H_{0i} vs. $H_{1i}, i = 1,2,$ respectively. Define two bivariate independent random variables $T_{H_0} = (T_1^0, T_2^0)$ and $T_{H_1} = (T_1^A, T_2^A)$. Let T_{H_k} be distributed according to the distribution function of (T_1, T_2) under H_k, $k = 0,1$. Then the likelihood ratio combination

$$LR(T_1, T_2) = f_{H_1}(T_1, T_2) / f_{H_0}(T_1, T_2)$$

results in the expression

$$EPV = \Pr\left\{ LR(T_1^0, T_2^0) \geq LR(T_1^A, T_2^A) \right\}$$

that is the minimum of the EPVs for any other combinations of the test statistics, where f_{H_0} and f_{H_1} are the joint density functions of T_{H_0} and T_{H_1}, respectively.

The linear combinations of the test statistics can be obtained in the form

$$l(\lambda; T_1, T_2) = T_1 + \lambda T_2,$$

and then the corresponding EPV is

$$EPV(\lambda) = \Pr(T_1^0 + \lambda T_2^0 \geq T_1^A + \lambda T_2^A).$$

Therefore, the BLC coefficient, λ_0, satisfies

$$\lambda_0 = \operatorname{argmin}_\lambda \left\{ EPV(\lambda) \right\} = \operatorname{argmin}_\lambda \Pr\left(T_1^0 + \lambda T_2^0 \geq T_1^A + \lambda T_2^A \right).$$

In this case, the test statistic $l(\lambda_0; T_1, T_2)$ is the resulting combination of the test statistics.

In the context mentioned above, regarding the pEPV, the linear combination of the test statistics is $l(\lambda; T_1, T_2) = T_1 + \lambda T_2$ and the BLC coefficient, λ_0, has the form

$$\lambda_0 = \operatorname{argmin}_\lambda \left\{ pEPV(\lambda) \right\} = \arg\max_\lambda \Pr\left\{ T_1^A + \lambda T_2^A \geq T_1^0 + \lambda T_2^0, T_1^0 + \lambda T_2^0 \geq F^{-1}_{T_1^0 + \lambda T_2^0}(1 - \alpha_U) \right\},$$

where $F_{T_1^0 + \lambda T_2^0}$ defines the distribution function of the random variable $T_1^0 + \lambda T_2^0$.

4. Examples

In this section, we illustrate the proposed methods developed above towards tackling several multiple testing problems. The examples with the BLC of test statistics based on minimizing the EPV will demonstrate that combining different multiple tests using the proposed method can be a routine task without relying on the specification of the theoretical joint distribution.

4.1 ***Example 1 (Parametric case)***

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ distribution. Consider the following hypotheses $H_0 : \mu = 0, \sigma^2 = 1$ vs. $H_1 : \mu > 0$ or $\sigma^2 \neq 1$.

The null hypothesis to be tested restricts the location and scale parameters, the two main parameters in the normal distribution. Assume values of the *t*-test statistic $T_1 = \bar{X}\sqrt{n}/s$ and the χ^2-test statistic $T_2 = (n-1)s^2$ are available only, where $\bar{X}$ and s^2 are the sample mean and the sample variance, respectively.

Since the data points are normally distributed and then their sample mean is independent from the sample standard deviation, one can prove that the joint distributions of T_1 and T_2 corresponding to $H_1 : (\mu = \mu_1 > 0$ or $\sigma^2 = \sigma_1^2 \neq 1)$ and H_0 have the forms shown in the Web Supplementary Materials. Then the LR test statistic, $LR(T_1, T_2)$, is described in the Web Appendix that also presents the formal notation of $MLR(T_1, T_2)$, an approximated $LR(T_1, T_2)$, corresponding to the case where μ_1 and σ_1 are assume to be unknown.

In the scenario of this example, regarding the BLC method, we can obtain the combined test statistic $T = T_1 + \lambda_0 T_2$, where $\lambda_0 = \arg\min_\lambda EPV(\lambda) = \arg\min_\lambda \Pr\left(T_1^0 + \lambda T_2^0 \geq T_1^A + \lambda T_2^A\right)$, where an explicit form of $\Pr\left(T_1^0 + \lambda T_2^0 \geq T_1^A + \lambda T_2^A\right)$ can be derived using the convolution calculations, or it can be accurately Monte Carlo approximated.

4.2 ***Example 2 (Nonparametric case)***

Let $X_1, ..., X_n$ be a random sample of size n from a population with mean μ and median M, and, say, we are interested in $H_0 : \mu = 0, M = M_0$ vs. $H_1 : \mu > 0$ or $M \neq M_0$, where M_0 is a specified value. The null hypothesis is about the two popular central tendencies. In Section 5, we examine this example choosing $M_0 = 0$ and 0.5. Assume that to test for $\mu = 0$, the *t*-test ($T_1 = \bar{X}\sqrt{n}/s$) was conducted; and to test for $M = M_0$, the one-sample sign test based on the sample median (T_2 based on $Med(X)$) was performed.

Ferguson[28] derived the joint distribution of the sample mean and the sample median. See the Web Supplementary Materials for details and definitions. We apply this result to obtain the large sample approximation to the likelihood ratio combination of the test statistics based on $X_1, ..., X_n$ with the density function f in the form

$$LR(T_1,T_2)=\frac{f(M_1)A_1}{f(M_0)A_0}\exp\left\{-\frac{1}{2}\left(A_1C-A_0D\right)\right\},$$

where $A_i=\left[1-\left(E\left|X_1-M_i\right|\right)^2/s_i^{\,2}\right]^{-1/2}, i=0,1,\ C=\left[T_1-n^{1/2}\mu_1/s_1\right]^2$

$-4E\left|X_1-M_1\right|(s_1)^{-1}\left(T_1-\sqrt{n}\mu_1/s_1\right)\left(T_2-M_1\right)f(M_1)\sqrt{n}+4f^2(M_1)n\left(T_2-M_1\right)^2,$

$D=T_1^2-4E\left|X_1-M_0\right|(s_0)^{-1}T_1\left(T_2-M_0\right)f(M_0)\sqrt{n}+4f^2(M_0)n\left(T_2-M_0\right)^2,$

$s_k^{\,2}=\sum_{i=1}^{n}(X_i-k\bar{X})^2/n,\ k=0,1,$ and μ_1 and M_1 are the mean and median under H_1.

If the parameters μ_1 and M_1 are unknown, we can use their estimators and a kernel estimation of f to calculate the approximate $LR(T_1,T_2)$.

In a similar manner to the test construction mentioned above, one can provide an approximation to the BLC of the test statistics. In this scenario, data-driven methods, e.g. the bootstrap methodology, can be used to approximate the coefficient $\lambda_0=\arg\min_\lambda\left\{EPV(\lambda)\right\}$ that is $\arg\min_\lambda \Pr\left(T_1^0+\lambda T_2^0\geq T_1^A+\lambda T_2^A\right)$. In Sections 5 and 6, we present more details regarding these algorithms.

4.3 ***Example 3 (a Goodness-of-fit problem)***

Consider the following hypothesis

$H_0:\mu=0, X$ is normally distributed vs. $H_1:\mu=\mu_1\neq 0$ or X is not normally distributed.

We can call this hypothesis a reinforced Goodness-of-Fit type statement where the hypothesis emphasizes the location difference. To test $\mu=0$, the two sided |*t*-test| is used and to test X is normally distributed the well-known Shapiro-Wilk test is used.

In the context of the BLC of the test statistics, we can obtain the combined test statistic $T=T_1+\lambda_0T_2$, where $\lambda_0=\arg\min_\lambda \Pr\left(T_1^0+\lambda T_2^0\geq T_1^A+\lambda T_2^A\right)$. To compute approximated values of λ_0 we use the fact that the distribution of $T_1^0+\lambda T_2^0$ has a known form for a fixed λ and the fact that the distribution of $T_1^A+\lambda T_2^A$ can be approximated via a bootstrap approach (see for details Sections 5 and 6).

5. Monte Carlo Study

We compared the performances of the developed tests with the classical Bonferroni and BH procedures by evaluating their average powers and EPVs. In the numerical studies, various settings of parameters under H_1 and different sample sizes are used, where the alternative parameters are

assumed to be known or unknown. With unknown alternative parameters, the test statistics are obtained based on the estimated parameters under H_1. In this case maximum likelihood estimation or bootstrap methods are used to construct the test statistics. Alternatively, we also employed novel bootstrap tilting methods for constructing the test statistics. See details related to the bootstrap tilting method in the Web Supplementary Materials. The testing strategies that we compare in the simulations are the approximated likelihood ratio combination of test statistics (called MLR in legends used in this section); the BLC of test statistics obtained using known alternative parameters (BLC); the best linear combinations of test statistics with estimated alternative parameters (BLC_1); the estimated best linear combinations of test statistics via bootstrap methods (BOOT); the best linear combinations of test statistics minimizing pEPV with $\alpha_U = 0.1$, when alternative parameters are known (BLC_p); the best linear combinations of test statistics minimizing pEPV ($\alpha_U = 0.1$) with estimated alternative parameters (BLC_{1p}); the best linear combinations of test statistics minimizing pEPV ($\alpha_U = 0.1$) via bootstrap methods ($BOOT_p$) and the best linear combinations of test statistics minimizing EPV with bootstrap tilting method ($BOOT_t$).

The Monte Carlo power and average power of considered tests are evaluated. To obtain the average Monte Carlo (MC) power, we evaluate the powers at different significant levels of 0.005, 0.01, 0.025, 0.05 and 0.1 and average them.

5.1 ***Example 1 (The Parametric case)***

Regarding Example 4.1, in the Monte Carlo setting, we considered 180 different scenarios based on combinations of the sample sizes, n of 30, 50, 75, 100, 150 and 200; μ of 0, 0.05, 0.15, 0.25, 0.35 and 0.45; and σ_1^2 of 1, 1.1, 1.2, 1.3 and 1.4. Five-thousand sample generations were carried out per each scenario. Table 2 shows that the percentage of the 180 scenarios that the average power of the proposed methods is higher than that of both the Bonferroni and BH procedures.

Table 2. The average Monte Carlo power comparison between the proposed methods and the classical techniques, the Bonferroni and BH procedures.

MLR	BLC	BLC_1	BOOT	BLC_p	BLC_{1p}	$BOOT_p$
88%	98%	85%	83%	100%	88%	89%

For example, there are 88% of the 180 scenarios that the average MC power of the MLR is higher than both the Bonferroni and BH procedures; 98% of the 180 scenarios that the average MC power of the BLC obtained using known alternative parameters is higher than both the Bonferroni and BH; 85% of the 180 scenarios that the average MC power of the BLC based on estimated alternative

parameters is higher than both Bonferroni and BH, etc. Table 3 shows that the percentage of the 180 scenarios that the Monte Carlo EPV of the proposed methods is lower than both the Bonferroni and BH.

Table 3. The Monte Carlo EPV comparison between the proposed methods and the classical techniques, the Bonferroni and BH procedures.

MLR	BLC	BLC_1	BOOT	BLC_p	BLC_{1p}	$BOOT_p$
80%	99%	73%	73%	99%	80%	79%

For example, there are 80% of the 180 scenarios that the MC EPV of the MLR is lower than both the Bonferroni and BH procedures; 99% of the 180 scenarios that the EPV of the BLC with known alternative parameters is lower than both the classical procedures; 73% of the 180 scenarios that the EPV of the BLC with bootstrap method is lower than both the classical procedures.

Web Table S1 in the Web Supplementary Materials shows the MC EPVs of the proposed methods and the classical procedures under some of the simulation scenarios. Among all the considered scenarios, the BLC has more cases with smaller EPVs. To justify the conducted results, Web Table S2 presents the MC Type I error rates of the proposed methods, the Bonferroni and BH procedures at the significant level $\alpha = 0.05$ under control. The MC Type I error rates are controlled well when the tests are implemented.

5.2 ***Example 2 (The Nonparametric case)***

Consider the statement of Section 4.2. In this example, we deal with median values of $M_0 = 0$ and 0.5 under H_0 in settings of two different underlying distributions, the normal and the exponential distributions, respectively. Testing relative to $M_0 = 0$ is in Example 2.1 and testing relative to $M_0 = 0.5$ is in Example 2.2.

5.2.1 ***Example 2.1***

The simulations are executed using samples from normal distributions. We consider 30 scenarios based on the combinations of the sample size of 30, 50, 75, 100, 150 and 200, and the alternative parameters $\mu_1 = M_1$ (data points are normally distributed), where the parameters can have the values of 0.05, 0.15, 0.25, 0.35 and 0.45, including, e.g., the cases (n, μ_1, M) $= (30, 0.05, 0.05),\ (100, 0.25, 0.25)$.

Table 4 shows that the percentage of the 30 scenarios that the average MC power of the proposed methods is higher than that of both the Bonferroni and the BH procedures. For example,

there are 96.67% of the scenarios that the average power of the MLR of the test statistics is higher than that of both the Bonferroni and BH procedures, etc.

Table 4. The average MC power comparison between the proposed methods and the classical procedures, the Bonferroni and BH schemes.

MLR	BLC_1	BOOT	BLC_{1p}	$BOOT_p$	$BOOT_t$
96.67%	96.67%	96.67%	96.67%	96.67%	96.67%

While obtaining the outputs to calculate Table 4, we observed that the classical procedures can slightly outperform the proposed schemes in the scenarios based on relatively large samples ($n \geq 150$) with the alternative parameter $\mu_1 \geq 0.35$. In these cases, the considered procedures provided the average MC power values $\simeq 1$. Perhaps, in this study, Monte Carlo errors of the numerical experiments can have critical roles when $n \geq 150$ and $\mu_1 \geq 0.35$.

Table 5 shows that the percentage of the 30 scenarios that the EPV of the proposed methods is lower than that of both Bonferroni and BH procedures. The table displays that 100% of the scenarios that the EPV of the MLR is lower than that of both Bonferroni and BH schemes; 86.67% of the 30 scenarios that the EPV of the BLC with the estimated alternative parameters is lower than that of both the Bonferroni and BH schemes; etc.

Table 5. The EPV comparison between the proposed methods and the classical schemes, the Bonferroni and BH procedures.

MLR	BLC_1	BOOT	BLC_{1p}	$BOOT_p$	$BOOT_t$
100.00%	86.67%	83.33%	80.00%	83.33%	86.67%

Web Table S3 in the Web Supplementary Materials presents the EPVs of the proposed methods and the classical procedures in the different simulation scenarios. Among all the simulation scenarios, the BLC estimated using the bootstrap tilting method more frequently registers the smaller EPVs. Web Table S4 shows the MC Type I error rates of the proposed methods, the Bonferroni and BH procedures at the significant level $\alpha = 0.05$. The results show that the Type I Error rates for the proposed methods are well controlled.

5.2.2 ***Example 2.2***

Under the exponential distribution, consider the following hypothesis

$$H_0 : \mu = 0, M = 0.5 \text{ vs. } H_1 : \mu = \mu_1 > 0 \text{ or } M = M_1 \neq 0.5.$$

Under the null hypothesis, we considered values of the random variable $a\xi - a$ with $\xi \sim \text{Exp}(1)$ and $a = (\ln(2) - 1)^{-1} / 2$.

In this example we set up 32 scenarios based on the sample sizes of 30, 50, 75, 100, 150 and 200. For the alternative parameters, we have the mean and median (μ_1, M_1) of $(0.1, 0.5)$, $(0.1, 0.6)$ and $(0.2, 0.5)$.

Table 6 displays that the percentage of the 32 scenarios that the average MC power of the proposed methods is higher than that of both the Bonferroni and BH procedures.

Table 6. The average MC power comparison between the proposed methods and the classical schemes, the Bonferroni and BH procedures.

MLR	BLC_1	BOOT	BLC_{1p}	$BOOT_p$	$BOOT_t$
62.50%	65.61%	75.00%	71.88%	78.13%	81.25%

Table 7 shows that the in a large percentage of the 32 scenarios the EPV of the proposed methods is lower than that of both the Bonferroni and BH procedures.

Table 7. The EPV comparison between the proposed methods, the classical schemes and the Bonferroni and BH procedures

MLR	BLC_1	BOOT	BLC_{1p}	$BOOT_p$	$BOOT_t$
100.00%	62.50%	77.50%	70.00%	77.50%	62.50%

Tables 6 and 7 demonstrate that the proposed methods have the better power and lower EPV than those of the Bonferroni and BH methods.

Web Table S5 in the Web Supplementary Materials shows the detailed EPV of the proposed methods and the classical procedures in the different simulation scenarios. Among all the simulation scenarios, the BLC estimated using the bootstrap tilting method has more cases with smaller EPVs. Web Table S6 reports that the corresponding Type I error rates are controlled well when the tests are implemented at the significant level $\alpha = 0.05$.

5.3 ***Example 3***

To test the reinforced goodness-of-fit test (Section 4.3), we choose the sample sizes of 30, 50, 75 and 100, and under the alternative hypothesis, we simulate data from *t*-distributions with $df = \{2,5,25\}$ and the mean $\mu_1 = \{0,0.5\}$ and data from Laplace distribution with mean $\mu_1 = 0$. Thus, 28 designs of the generated samples are stated. Table 8 shows that the percentage of the 28 scenarios that the average MC power of the proposed methods is higher than that of both the Bonferroni and BH procedures. There are 100% of the 28 scenarios that the average MC power of the BLC is higher than that of both the Bonferroni and BH procedures; 83.33% scenarios that the

average power of the BLC approximated using the bootstrap method is higher than both Bonferroni and BH schemes, etc.

Table 8. The average MC power comparison between the proposed methods and the classical procedures, the Bonferroni and BH schemes.

BLC	BOOT	BLC_p	$BOOT_p$
100.00%	83.33%	100.00%	83.33%

In a similar manner to that of Table 8, Table 9 presents the percentage of the 28 scenarios that the EPV of the proposed methods is lower than that of both the Bonferroni and BH procedures.

Table 9. The EPV comparison between the proposed methods and the Bonferroni and BH procedures.

BLC	BOOT	BLC_p	$BOOT_p$
100.00%	96.43%	100.00%	96.43%

Tables 8 and 9 demonstrate that the proposed methods outperform the classical algorithms, the Bonferroni and BH techniques. Supporting this conclusion, Web Table S7 in the Web Supplementary Materials shows the detailed EPV of the proposed methods and the classical procedures under some of the simulation scenarios. Web Table S8 indicates that the Type I error rates of the considered tests are controlled well at the significant level $\alpha = 0.05$.

Remark: The Monte Carlo powers at $\alpha = 0.05$ of the considered tests were also compared in the context of Examples 5.1-5.3. In this aspect we observed that in most of situations the proposed methods are superior to the classical procedures. For example, we provide Table 10 that corresponds to Example 5.3.

Table 10. The MC power at $\alpha = 0.05$ of the proposed methods and the Bonferroni and BH procedures.

N	*Distribution*	μ	BON	BH	BOOT	$BOOT_p$
30	*t, df=2*	0	0.0651	0.0700	0.1606	0.1814
30	*t, df=25*	0	0.0364	0.0404	0.0546	0.0544
30	*t, df=2*	0.5	0.2899	0.3044	0.3534	0.3423
30	*t, df=25*	0.5	0.6671	0.6876	0.6121	0.5784
75	*t, df=2*	0	0.1133	0.1179	0.3010	0.3353
75	*t, df=25*	0	0.0389	0.0441	0.0536	0.0586
75	*t, df=2*	0.5	0.6436	0.6556	0.7109	0.6946
75	*t, df=25*	0.5	0.9787	0.9817	0.8937	0.8774
30	Laplace	0	0.0410	0.0451	0.0884	0.0884
50	Laplace	0	0.0469	0.0506	0.0983	0.0983
75	Laplace	0	0.0503	0.0555	0.1234	0.1234
100	Laplace	0	0.0467	0.0511	0.1513	0.1513

6. Real Data Example

Myocardial infarction is commonly caused by blood clots blocking the blood flow of the heart leading heart muscle injury. The heart disease is leading cause of death affecting about or higher than 20% of populations regardless of different ethnicities according to the Centers for Disease Control and Prevention (e.g., Schisterman et al.[29,30]).

We illustrate the application of the proposed approach based on a sample from a study that evaluates biomarkers related to the myocardial infarction (MI). The study was focused on the residents of Erie and Niagara counties, 35-79 years of age. The New York State department of Motor Vehicles drivers' license rolls was used as the sampling frame for adults between the age of 35 and 65 years, while the elderly sample (age 65-79) was randomly chosen from the Health Care Financing Administration database. We consider the biomarker "high density lipoprotein (HDL)-cholesterol" that is often used as a discriminant factor between individuals with and without MI disease (e.g., Schisterman et al[29,30]). The HDL-cholesterol levels were examined from a 12-hour fasting blood specimen for biochemical analysis at baseline. A total of 366 measurements of HDL biomarker were evaluated by the study. The sample of 105 biomarker values was collected on cases who survived on MI and the sample of 261 measurements on controls who had no previous MI.

Note that, oftentimes measurements related to biological processes follow a log-normal distribution (see for details Limpert[31] and Vexler et al.[10], pp. 13-14). Thus, we are interested in whether the HDL cholesterol of the control group, say X, has the same log-normal distribution as that of the case group, Y. Web Figure S1 in the Web Supplementary Materials depicts the histograms based on values of $\log(X)$ and $\log(Y)$, respectively. Our hypothesis to test can be expressed as

$$H_0: X \sim Y, X \sim \text{Lognormal} \text{ vs. } H_1: H_0 \text{ is not true.}$$

To test for $X \sim Y$, the Wilcoxon-test (T_1) was performed, and to test $\log(X) \sim \text{N}$, the Shapiro-Wilk test (T_2) was used. To examine the proposed 'EPV/ROC' technique, we apply the following algorithm: 1) since under the null hypothesis the distribution function of (X, Y) is specified, the mean and the variance of $\{\log(X), \log(Y)\}$ were estimated and then 100,000 random variables were generated from the corresponding log-normal distribution to Monte-Carlo-approximate the distribution of (T_1^0, T_2^0) (alternatively, the parametric bootstrap approach can be applied in this stage); 2) using the HDL cholesterol data, 10,000 bootstrap resamples were obtained from X and Y, respectively, to empirically approximate the distribution function of (T_1^A, T_2^A); 3) the outputs of the

stages 1) - 2) were applied to compute an approximation to $\lambda_0 = \arg\min_\lambda \Pr\left(T_1^0 + \lambda T_2^0 \geq T_1^A + \lambda T_2^A\right)$ employing the Mann-Whitney U-statistic method and the R command[32] *wilcox.test(* $T_1^0 + \lambda T_2^0$, $T_1^A + \lambda T_2^A$ *)[[1]]*; 4) the combined test statistic $T = T_1 + \lambda_0 T_2$ was conducted.

Thus we could compare the average power and the EPVs of the BLC of the test statistics and those of the Bonferroni and BH procedures. Cases with the power<0.05 we could associate with the Type I error rates. The algorithm mentioned above was repeatedly used employing randomly selected subsamples from the data of sizes 25, 50, 75, 100 from the case and control groups to evaluate robustness of the test procedures. Figures 2 and 3 depict the average result via 1,000 repetitions at each fixed sample size, *n*, regarding the average power (we evaluated the powers at different significant levels of 0.005, 0.01, 0.025, 0.05 and 0.1 and average them) and the EPVs of the estimated BLC (called 'BOOT' in legends used in Figures 2 and 3), the Bonferroni and BH procedures plotted again the sample sizes.

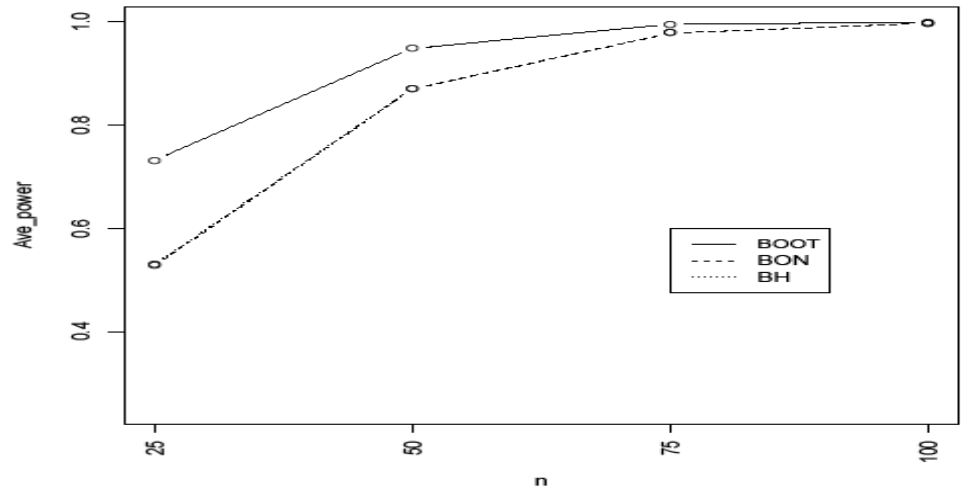

Figure 2. The average power of the approximated BLC of the test statistics, the Bonferroni and BH procedures plotted against the sample sizes.

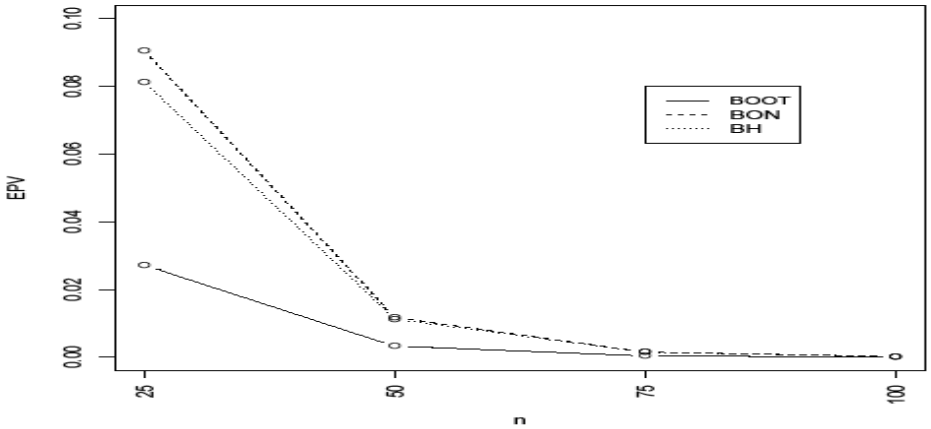

Figure 3. The EPVs of the approximated BLC the test statistics, Bonferroni and BH procedures plotted against the sample sizes.

Figures 2 and 3 demonstrate that the BLC has the higher power and lower EPVs than those of the Bonferroni and BH procedures. As the sample size increases, the changes of the power and EPV of the proposed method is smaller than those of the classical procedures. This implies that the proposed method has better characteristics comparing with the classical schemes based on relatively small sample sizes, while the differences between the tests vanish when the sample size increases.

Note that it is reasonable to assume that consistent test procedures based on relatively large samples have the much of the same operating characteristics, e.g. power one tests. In this example, we observed that the decision-making schemes under consideration simultaneously offer to reject the null hypothesis, when more than $n = 100$ data points were employed.

7. Discussion

We have seen that the EPV is a very useful and succinct tool as a measurement of performances in decision-making mechanisms. We have proposed a novel methodology to analyze and visualize characteristics of tests procedures. To this end the 'EPV/ROC' concept has been introduced. This approach provides us new and efficient perspectives to develop and examine statistical decision-making policies, including those related to: the partial EPV considerations, associations between the EPV and the power of tests, visualization of testing mechanisms' properties; developments of optimal tests minimizing the EPVs and creations of new methods for optimally combining multiple test statistics. Many possible researches can be done based on the concept we introduced in this paper. For example, a large sample theory can be developed to evaluate the EPVs in several parametric and nonparametric scenarios; Bayesian type methods can be developed in order to evaluate test properties in the 'EPV/ROC' frame. The proposed technique can be easily applied to

obtain confidence region estimation of vector-parameters based on their elements' confidence interval estimates. These topics need further strong empirical and methodological investigations. In the context of the multiple testing problem, the Bonferroni method and the Benjamini-Hochberg procedure were applied in this paper for an illustrative purpose only. The bootstrap tools including the Bootstrap tiling have been shown to be very reasonable in the test properties evaluations.

We hope that the proposed concept convinces the medical practitioners regarding the usefulness of 'EPV/ROC' methodology to evaluate different decision-making procedures, their constructions and properties.

8. Supplementary Materials

The supplemental material consists of the following aspects: Web Appendices referenced in Sections 4, 5 and the outlined bootstrap tilting method (Web_Supplementary.pdf).

Acknowledgements

Dr. Hutson's effort was supported by Roswell Park Cancer Institute and National Cancer Institute (NCI) grant P30CA016056.

Drs. Vexler and Yu's efforts were supported by the National Institutes of Health (NIH) grant 1G13LM012241-01.

We gratefully acknowledge the editor and the two anonymous referees for their very helpful and constructive comments and suggestions, which led to significant improvements to the paper.

References

1. Stigler SM. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass: Belknap Press of Harvard University Press. London, 1986.
2. Fisher R. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd. London, 1925.
3. Wasserstein RL and Lazar N. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 2016; **70**: 129-133.
4. Trafimow D and Marks M. Editorial in Basic and Applied Social Pschology. *Basic and Applied Social Pschology* 2015; **37**: 1-2.
5. Lazzeroni LC, Lu Y and Belitskaya-Levy I. *P*-values in genomics: Apparent precision masks high uncertainty. *Molecular Psychiatry* 2014; **19**: 1336-1340.
6. Dempster AP and Schatzoff M. Expected significance level as a sensitivity index for test statistics. *Journal of the American Statistical Association* 1965; **60**: 420-436.
7. Schatzoff M. Sensitivity comparisons among tests of the general linear hypotheses. *Journal of the American Statistical Association* 1966; **61**: 415-435.

8. Sackrowitz H and Samuel-Cahn E. *P* values as random variables-expected *p* values. *The American Statistician* 1999; **53**: 326-331.

9. Neyman J and Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 1993; **231**: 694–706.

10. Vexler A, Hutson AD and Chen X. *Statistical Testing Strategies in the Health Sciences*. Chapman & Hall/CRC: New York, 2016.

11. Zou KH, Liu A, Bandos AI, Ohno-Machado L, and Rockette HE. *Statistical evaluation of diagnostic performance: topics in ROC analysis*. CRC Press: New York, 2001.

12. Liu A and Schisterman EF. Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* 2003; **45**: 631-644.

13. Vexler A, Liu A, Eliseeva E, and Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics* 2008a; **64**: 895–903.

14. Vexler A, Schisterman EF and Liu A. Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine* 2008b; **27**: 280-296.

15. Zhou XH, Obuchowsk NA and Mcclish DK. *Statistical methods in diagnostic medicine*. John Wiley & Sons: New York, 2009.

16. Dmitrienko A, Tamhane AC and Bretz F (Eds.). *Multiple testing problems in pharmaceutical statistics*. CRC Press: New York, 2009.

17. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 1995; **57**: 289-300.

18. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press Inc.: New York, 2003.

19. Liu A, Schisterman EF and Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine* 2001; **24**: 37-47.

20. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**: 387-415.

21. Julious SA. Why do we use pooled variance analysis of variance? *Pharmaceutical Statistics* 2005; **4**: 3-5.

22. Zimmerman DW and Zumbo BD. Hazards in choosing between pooled and separate-variance t tests. *Psiologica* 2009; **30**: 371-390.

23. Roy SN. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics* 1953; **24**: 220-238.

24. Liu CL, Liu A and Halabi S. A min-max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine* 2011; **30**: 2005-2014.

25. Su JQ and Liu JS. Linear combintion of multiple diagnostic markers. *Journal of the American Statistical Association* 1993; **88**: 1350-1355.

26. Chen X, Vexler A and Markatou M. Empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers. *Computational Statistics & Data Analysis* 2015; **82**: 186–198.

27. Pepe MS and Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics* 2000; **1**: 123-140.

28. Ferguson TS. Asymptotic Joint distribution of Sample Mean and a Sample Quantile. Unpublished, 1998. Available at http://www.math.ucla.edu/~tom/papers/unpublished/meanmed.pdf

29. Schisterman EF, Faraggi D, Browne R, Freudenheim J, Dorn J, Muti P, Armstrong D, Reiser B and Trevisan M. Tbars and cardiovascular disease in a population-based sample. *Journal of Cardiovascular Risk* 2001; **8**: 219-225.

30. Schisterman EF, Faraggi D, Browne R, Freudenheim J, Dorn J, Muti P, Armstrong D, Reiser B and Trevisan M. Minimal and best linear combination of oxidative stress and antioxidant biomarkers to discriminate cardiovascular disease. *Nutrition, Metabolism, and Cardiovascular Disease* 2002; **12**: 259-266.

31. Limpert E, Stahel WA and Abbt M. Log-Normal Distributions across the Sciences: Keys and Clues on the Charms of Statistics, and How Mechanical Models Resembling Gambling Machines Offer a Link to a Handy Way to Characterize Log-Normal Distributions, Which Can Provide Deeper Insight into Variability and Probability—Normal or Log-Normal: That Is the Question. *BioScience* 2001; **51**: 341-352.

32. R Development Core Team. *R: A language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2002. http://www.R-project.org.

Web-Supplementary Material

Expected *P*-values in Light of an ROC Curve Analysis Applied to Optimal Multiple Testing Procedures

Albert Vexler,[1,*] Jihnhee Yu[1], Yang Zhao[1], Alan D. Hutson[2] and Gregory Gurevich[1]

[1]Department of Biostatistics, The State University of New York, Buffalo, NY 14214, U.S.A.

[2]Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY 14263, U.S.A.

**email:* avexler@buffalo.edu

The joint distributions of the test statistics T_1 and T_2 from Section 4.1 "*Example 1 (Parametric case)*"

If the data are normally distributed then their sample mean is independent from the corresponding sample standard deviation. Thus, one can prove that the joint distributions of T_1 and T_2 corresponding to $H_1:(\mu=\mu_1>0$ or $\sigma^2=\sigma_1^2\neq 1)$ and H_0 have the following forms

$$f_{H_1}(T_1,T_2)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\left(\frac{1}{\sqrt{n-1}}T_1\sqrt{T_2}-\sqrt{n}\mu_1\right)^2\left(2\sigma_1^2\right)^{-1}\right\}\frac{T_2^{\left(\frac{n-1}{2}-1\right)}e^{-\frac{T_2}{2\sigma_1^2}}}{\left(2\sigma_1^2\right)^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}\sqrt{\frac{T_2}{n-1}}, \quad \text{(A.1)}$$

$$f_{H_0}(T_1,T_2)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{1}{\sqrt{n-1}}T_1\sqrt{T_2}\right)^2\right\}\frac{T_2^{\left(\frac{n-1}{2}-1\right)}e^{-\frac{T_2}{2}}}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}\sqrt{\frac{T_2}{n-1}}, \quad \text{(A.2)}$$

where $\Gamma(u)$ is the gamma function.

Using (A.1) and (A.2), we obtain the LR test statistic,

$$LR(T_1,T_2)=\frac{f_{H_1}(T_1,T_2)}{f_{H_0}(T_1,T_2)}=\exp\left\{-\frac{1}{2\sigma_1^2}\left(\frac{T_1\sqrt{T_2}}{\sqrt{n-1}}-\sqrt{n}\mu_1\right)^2-\frac{T_2}{2\sigma_1^2}+\frac{1}{2}\left(\frac{T_1\sqrt{T_2}}{\sqrt{n-1}}\right)^2+\frac{T_2}{2}\right\}\sigma_1^{-n}.$$

If μ_1 and σ_1 are unknown, we can use their maximum likelihood (ML) estimators $\hat{\mu}_1$ and $\hat{\sigma}_1$ to derive the approximated $LR(T_1,T_2)$ given as

$$MLR(T_1,T_2)=\exp\left\{-\frac{1}{2\hat{\sigma}_1^2}\left(\frac{T_1\sqrt{T_2}}{\sqrt{n-1}}-\sqrt{n}\hat{\mu}_1\right)^2-\frac{T_2}{2\hat{\sigma}_1^2}+\frac{1}{2}\left(\frac{T_1\sqrt{T_2}}{\sqrt{n-1}}\right)^2+\frac{T_2}{2}\right\}\hat{\sigma}_1^{-n}.$$

Ferguson's (1998) result used in Section 4.2 *"Example 2 (Nonparametric case)"*

Proposition. Let $X_1,...,X_n$ be i.i.d. with distribution function $F(x)$, density $f(x)$, mean μ, finite variance σ^2 and median M. Then the asymptotic joint distribution of the sample mean, $\bar{X}_n$, and the sample median, Y_n, satisfies

$$\sqrt{n}\left[\begin{pmatrix}\bar{X}_n\\ Y_n\end{pmatrix}-\begin{pmatrix}\mu\\ M\end{pmatrix}\right]\xrightarrow{L}N\left[\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\sigma^2 & \{2f(M)\}^{-1}E|X_1-M|\\ \{2f(M)\}^{-1}E|X_1-M| & \{4f^2(M)\}^{-1}\end{pmatrix}\right].$$

The bootstrap tilting method

In the classical nonparametric bootstrap method, one assigns the resampling weight $1/n$ to each observed data point. The bootstrap tilting method changes the resampling weight from $1/n$ to different weights with respect to the constraint imposed by the null hypothesis, similar to empirical likelihood methods. For example, suppose we are interested in

$$H_0: EX=0 \text{ vs. } H_1: EX\neq 0.$$

We can find the probability weights corresponding to $H_0: EX=0$, using the approaches found in the empirical likelihood literature (e.g., Vexler et al., 2016) is used. For simplicity, let $X_1,...,X_n$ be independent and identically distributed data points. Then the empirical likelihood function has the form $L_p=\prod_{i=1}^n w_i$, where the component $w_i\in(0,1), i=1,...,n$ maximize the likelihood L_p. Under H_0, the constraints of interests are $\sum_{i=1}^n w_i=1$ and $\sum_{i=1}^n w_i X_i=0$. Empirical estimation of the weights w_i can be accomplished using the method of Lagrange multipliers. This provides bootstrap with resampling weights $w_i, i=1,...,n,$ from the data in a similar manner to the parametric bootstrap approach.

Web Tables related to Section 5.

Table S1. Section 5.1: The EPV of the proposed methods, the Bonferroni (BON) and the BH procedures.

n	μ_1	σ_1^2	BON	BH	MLR	BLC	BLC_1	BOOT	BLC_p	BLC_{1p}	$BOOT_p$
30	0.00	1.0	0.567	0.486	0.494	0.479	0.499	0.497	0.512	0.503	0.504
30	0.05	1.1	0.471	0.403	0.421	0.386	0.383	0.395	0.389	0.436	0.447
30	0.15	1.2	0.341	0.293	0.287	0.287	0.282	0.280	0.288	0.293	0.318
30	0.25	1.3	0.221	0.190	0.176	0.175	0.182	0.189	0.177	0.168	0.170
30	0.35	1.4	0.132	0.114	0.098	0.094	0.092	0.141	0.097	0.104	0.096
50	0.00	1.0	0.571	0.489	0.496	0.498	0.497	0.501	0.499	0.496	0.498
50	0.05	1.1	0.448	0.382	0.399	0.365	0.443	0.373	0.369	0.372	0.356
50	0.15	1.2	0.278	0.238	0.230	0.217	0.250	0.238	0.219	0.221	0.234
50	0.25	1.3	0.143	0.123	0.108	0.100	0.121	0.104	0.105	0.122	0.105
50	0.35	1.4	0.064	0.056	0.043	0.036	0.040	0.038	0.038	0.038	0.038
100	0.00	1.0	0.576	0.493	0.503	0.497	0.503	0.503	0.508	0.498	0.502
100	0.05	1.1	0.394	0.338	0.355	0.307	0.341	0.319	0.310	0.310	0.323
100	0.15	1.2	0.171	0.147	0.137	0.127	0.138	0.126	0.139	0.124	0.125
100	0.25	1.3	0.052	0.046	0.035	0.029	0.033	0.056	0.039	0.030	0.043
100	0.35	1.4	0.013	0.011	0.006	0.005	0.007	0.005	0.006	0.005	0.007
200	0.00	1.0	0.574	0.491	0.502	0.502	0.492	0.503	0.505	0.501	0.498
200	0.05	1.1	0.311	0.268	0.284	0.238	0.272	0.246	0.239	0.239	0.308
200	0.15	1.2	0.071	0.062	0.051	0.050	0.044	0.047	0.052	0.045	0.068
200	0.25	1.3	0.008	0.007	0.004	0.003	0.007	0.006	0.004	0.003	0.003
200	0.35	1.4	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table S2. Section 5.1: The Monte Carlo Type I error rates of the proposed methods, the Bonferroni and the BH procedures at the significant level $\alpha = 0.05$.

n	BON	BH	MLR	BLC	BLC_1	BOOT	BLC_p	BLC_{1p}	$BOOT_p$
30	0.0496	0.0496	0.0478	0.0523	0.0501	0.0480	0.0537	0.0521	0.0489
50	0.0499	0.0502	0.0490	0.0487	0.0470	0.0489	0.0478	0.0491	0.0494
75	0.0471	0.0475	0.0539	0.0500	0.0491	0.0501	0.0506	0.0474	0.0523
100	0.0537	0.0543	0.0579	0.0519	0.0483	0.0510	0.0520	0.0476	0.0499
150	0.0493	0.0496	0.0449	0.0467	0.0497	0.0483	0.0509	0.0487	0.0509
200	0.0499	0.0505	0.0453	0.0496	0.0513	0.0520	0.0536	0.0517	0.0501

Table S3. Section 5.2.1: The EPVs of the proposed methods and the Bonferroni (BON) and BH procedures.

$\mu_1 = M_1$	n	BON	BH	MLR	BLC_1	BOOT	BLC_{1p}	$BOOT_p$	$BOOT_t$
0.05	30	0.674	0.585	0.489	0.561	0.45	0.568	0.487	0.43
0.15	30	0.573	0.496	0.413	0.372	0.471	0.674	0.53	0.42
0.25	30	0.414	0.358	0.3	0.406	0.628	0.779	0.631	0.165
0.35	30	0.255	0.223	0.188	0.132	0.775	0.865	0.773	0.089
0.45	30	0.137	0.12	0.102	0.073	0.074	0.074	0.047	0.049
0.05	50	0.657	0.567	0.479	0.557	0.577	0.563	0.399	0.451
0.15	50	0.497	0.428	0.368	0.264	0.331	0.273	0.26	0.246

0.25	50	0.288	0.249	0.219	0.122	0.142	0.127	0.138	0.147
0.35	50	0.129	0.113	0.101	0.047	0.06	0.049	0.062	0.066
0.45	50	0.046	0.04	0.036	0.015	0.019	0.017	0.022	0.02
0.05	75	0.647	0.559	0.477	0.516	0.381	0.54	0.381	0.585
0.15	75	0.43	0.369	0.326	0.75	0.181	0.755	0.181	0.768
0.25	75	0.19	0.164	0.151	0.086	0.065	0.088	0.065	0.082
0.35	75	0.056	0.05	0.048	0.022	0.018	0.023	0.018	0.019
0.45	75	0.012	0.011	0.011	0.004	0.004	0.004	0.004	0.004
0.05	100	0.635	0.544	0.47	0.39	0.365	0.429	0.365	0.594
0.15	100	0.367	0.313	0.282	0.513	0.146	0.673	0.145	0.195
0.25	100	0.124	0.107	0.101	0.05	0.039	0.052	0.039	0.045
0.35	100	0.025	0.022	0.022	0.008	0.007	0.01	0.007	0.008
0.45	100	0.003	0.003	0.003	0.001	0.001	0.001	0.001	0.001
0.05	150	0.605	0.518	0.455	0.337	0.377	0.345	0.371	0.433
0.15	150	0.268	0.229	0.214	0.099	0.098	0.099	0.0937	0.808
0.25	150	0.053	0.047	0.046	0.016	0.019	0.016	0.018	0.018
0.35	150	0.005	0.005	0.005	0.001	0.001	0.001	0.001	0.001
0.45	150	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
0.05	200	0.587	0.5	0.444	0.324	0.315	0.325	0.315	0.313
0.15	200	0.199	0.17	0.163	0.074	0.071	0.074	0.071	0.072
0.25	200	0.023	0.022	0.021	0.007	0.007	0.008	0.007	0.007
0.35	200	0.001	0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001
0.45	200	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table S4. Section 5.2.1: The MC Type I error rates of the proposed methods, the Bonferroni and BH procedures at the significant level $\alpha = 0.05$.

n	BON	BH	MLR	BLC_1	BOOT	BLC_{1p}	$BOOT_p$	$BOOT_t$
30	0.032	0.037	0.051	0.049	0.050	0.049	0.053	0.048
50	0.037	0.038	0.054	0.047	0.046	0.055	0.056	0.054
75	0.037	0.040	0.052	0.047	0.046	0.049	0.047	0.046
100	0.037	0.039	0.048	0.048	0.049	0.049	0.050	0.049
150	0.036	0.039	0.050	0.050	0.049	0.050	0.051	0.051
200	0.036	0.038	0.049	0.047	0.047	0.053	0.055	0.046

Table S5. Section 5.2.2: The EPVs of the proposed methods e and the Bonferroni (BON) and BH procedures.

μ_1	M_1	n	MLR	BON	BH	BLC_1	BLC_p	BOOT	$BOOT_p$	$BOOT_t$
0.0	0.6	30	0.498	0.657	0.576	0.532	0.601	0.535	0.586	0.53
0.1	0.5	30	0.434	0.6	0.527	0.374	0.375	0.377	0.377	0.364
0.1	0.6	30	0.468	0.622	0.544	0.424	0.428	0.428	0.476	0.421
0.2	0.5	30	0.259	0.42	0.375	0.213	0.215	0.217	0.217	0.211
0.0	0.6	50	0.484	0.647	0.56	0.651	0.655	0.646	0.652	0.65

0.1	0.5	50	0.405	0.572	0.498	0.437	0.474	0.486	0.344	0.436
0.1	0.6	50	0.457	0.605	0.526	0.596	0.609	0.583	0.587	0.596
0.2	0.5	50	0.191	0.341	0.304	0.48	0.159	0.158	0.158	0.48
0.0	0.6	75	0.463	0.626	0.545	0.678	0.679	0.525	0.391	0.678
0.1	0.5	75	0.38	0.546	0.48	0.439	0.4485	0.312	0.312	0.439
0.1	0.6	75	0.45	0.588	0.509	0.603	0.6147	0.373	0.373	0.603
0.2	0.5	75	0.13	0.259	0.231	0.258	0.263	0.11	0.11	0.258
0.0	0.6	100	0.449	0.609	0.526	0.683	0.695	0.36	0.441	0.683
0.1	0.5	100	0.348	0.518	0.454	0.434	0.441	0.288	0.288	0.434
0.1	0.6	100	0.433	0.565	0.486	0.59	0.591	0.352	0.352	0.59
0.2	0.5	100	0.092	0.202	0.181	0.081	0.085	0.08	0.08	0.081
0.0	0.6	150	0.429	0.595	0.515	0.714	0.739	0.741	0.747	0.714
0.1	0.5	150	0.318	0.482	0.422	0.463	0.248	0.263	0.249	0.463
0.1	0.6	150	0.408	0.528	0.451	0.584	0.594	0.664	0.606	0.584
0.2	0.5	150	0.041	0.115	0.104	0.045	0.047	0.045	0.045	0.045
0.0	0.6	200	0.398	0.567	0.49	0.764	0.768	0.535	0.535	0.764
0.1	0.5	200	0.276	0.436	0.383	0.217	0.224	0.219	0.219	0.217
0.1	0.6	200	0.386	0.494	0.423	0.728	0.728	0.289	0.289	0.728
0.2	0.5	200	0.019	0.069	0.062	0.025	0.028	0.026	0.025	0.026
0.0	0.6	300	0.355	0.53	0.459	0.808	0.813	0.829	0.829	0.808
0.1	0.5	300	0.21	0.365	0.321	0.174	0.175	0.174	0.174	0.174
0.1	0.6	300	0.36	0.45	0.383	0.213	0.216	0.774	0.216	0.213
0.2	0.5	300	0.004	0.024	0.022	0.009	0.009	0.009	0.009	0.009
0.0	0.6	500	0.278	0.447	0.389	0.883	0.894	0.867	0.877	0.883
0.1	0.5	500	0.129	0.264	0.235	0.111	0.112	0.111	0.111	0.111
0.1	0.6	500	0.282	0.339	0.289	0.149	0.162	0.775	0.776	0.15
0.2	0.5	500	<0.001	0.003	<0.001	0.001	0.001	0.001	0.001	0.001

Table S6. **Section 5.2.2:** The MC Type I error rates of the proposed methods and the Bonferroni and BH procedures at the significant level $\alpha = 0.05$.

n	MLR	BON	BH	BLC_1	BLC_p	BOOT	$BOOT_p$	$BOOT_t$
30	0.051	0.057	0.056	0.049	0.047	0.049	0.047	0.053
50	0.049	0.048	0.048	0.049	0.052	0.048	0.052	0.048
75	0.048	0.047	0.047	0.050	0.049	0.05	0.051	0.046
100	0.049	0.046	0.046	0.049	0.051	0.048	0.052	0.048
150	0.044	0.041	0.041	0.056	0.052	0.055	0.051	0.052
200	0.053	0.043	0.043	0.051	0.046	0.051	0.046	0.051
300	0.050	0.047	0.047	0.050	0.047	0.051	0.046	0.051
500	0.049	0.042	0.042	0.055	0.048	0.054	0.049	0.050

Table S7. Section 5.3: The EPV of the proposed methods and the Bonferroni (BON) and BH procedures.

η	Distribution	df	μ_1	BLC	BON	BH	BLC_p	BOOT	$BOOT_p$
30	t	2	0	0.344	0.529	0.438	0.357	0.346	0.35
30	t	5	0	0.461	0.635	0.529	0.467	0.464	0.462
30	t	25	0	0.498	0.662	0.548	0.5	0.493	0.489
30	t	2	0.5	0.199	0.296	0.245	0.204	0.199	0.203
30	t	5	0.5	0.115	0.163	0.133	0.115	0.116	0.121
30	t	25	0.5	0.067	0.102	0.084	0.068	0.067	0.072
50	t	2	0	0.272	0.469	0.392	0.286	0.284	0.272
50	t	5	0	0.443	0.628	0.521	0.451	0.444	0.445
50	t	25	0	0.497	0.665	0.551	0.498	0.498	0.499
50	t	2	0.5	0.113	0.18	0.153	0.116	0.133	0.151
50	t	5	0.5	0.046	0.065	0.053	0.049	0.047	0.047
50	t	25	0.5	0.018	0.029	0.024	0.018	0.018	0.021
75	T	2	0	0.201	0.39	0.329	0.214	0.2	0.204
75	t	5	0	0.424	0.612	0.506	0.434	0.424	0.429
75	t	25	0	0.494	0.663	0.547	0.496	0.495	0.491
75	t	2	0.5	0.058	0.098	0.084	0.058	0.06	0.089
75	t	5	0.5	0.015	0.022	0.018	0.016	0.017	0.018
75	t	25	0.5	0.004	0.007	0.006	0.004	0.005	0.005
100	t	2	0	0.144	0.338	0.29	0.155	0.147	0.144
100	t	5	0	0.396	0.609	0.504	0.408	0.411	0.399
100	t	25	0	0.489	0.665	0.554	0.49	0.485	0.485
100	t	2	0.5	0.028	0.056	0.048	0.029	0.028	0.03
100	t	5	0.5	0.006	0.009	0.007	0.006	0.011	0.005
100	t	25	0.5	0.001	0.002	0.001	0.001	0.001	0.001
30	Laplace		0	0.423	0.623	0.51	0.442	0.422	0.429
50	Laplace		0	0.391	0.604	0.495	0.415	0.396	0.391
75	Laplace		0	0.346	0.58	0.477	0.367	0.346	0.353
100	Laplace		0	0.309	0.568	0.469	0.333	0.308	0.45

Table S8. Section 5.3: The MC Type I error rates of the proposed methods and the Bonferroni and BH procedures at the significant level $\alpha = 0.05$.

n	BON	B&H	BLC	BLCp	BOOT	$BOOT_p$
30	0.036	0.039	0.050	0.052	0.049	0.048
50	0.039	0.043	0.053	0.049	0.052	0.050
75	0.037	0.042	0.051	0.051	0.049	0.050
100	0.036	0.041	0.050	0.047	0.049	0.050

Web Histograms related to Section 6.

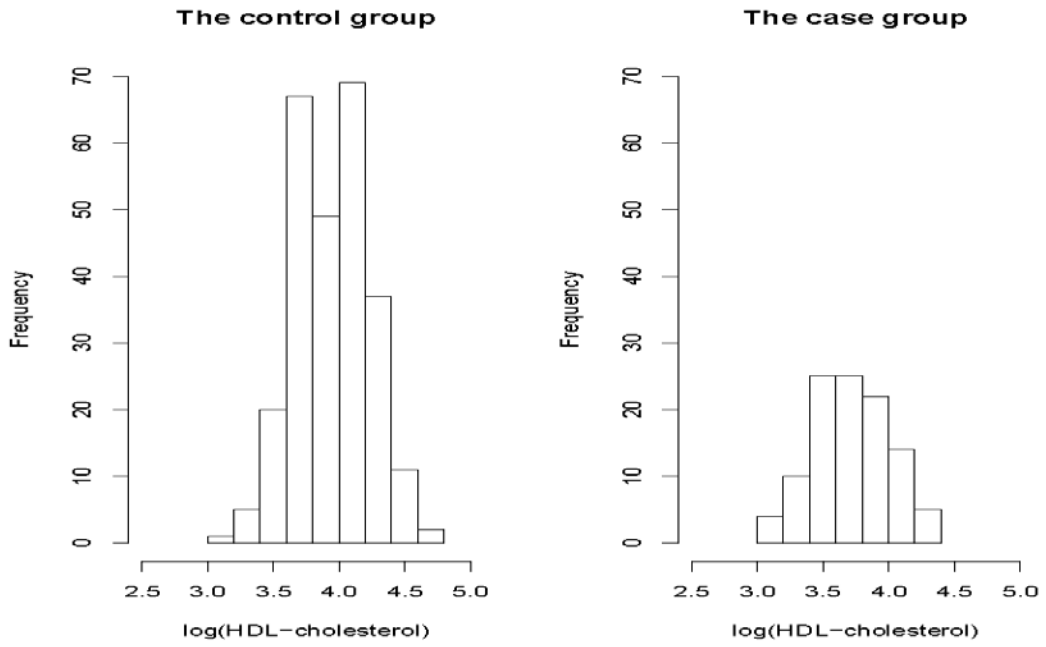

Figure S1. Histograms of the log-transformed biomarkers of interest corresponding to HDL-cholesterol controls (the estimated mean and the estimated standard deviation of log(HDL-cholesterol) are 3.94 and 0.29, respectively) and HDL-cholesterol cases (the estimated mean and the estimated standard deviation of log(HDL-cholesterol) are 3.72 and 0.292, respectively).

REFERENCE

Ferguson T.S. (1998). Asymptotic Joint distribution of Sample Mean and a Sample Quantile. Unpublished. Available at http://www.math.ucla.edu/~tom/papers/unpublished/meanmed.pdf

Vexler, A., Hutson, A. D. and Chen, X. (2016). *Statistical Testing Strategies in the Health Sciences.* Chapman & Hall/CRC, New York