To t-Test or not To t-test?: a *P*-Values-based Point of View in the ROC Curve Framework.

Albert Vexler,[a,*] and Jihnhee Yu[a],

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

**email:* avexler@buffalo.edu

ABSTRACT

A common statistical doctrine supported by many introductory courses and textbooks is that *t*-test type procedures based on normally distributed data points are anticipated to provide a standard in decision making. In order to motivate scholars to examine this convention, we introduce a simple approach based on graphical tools of receiver operating characteristic (ROC) curve analysis, a well-established biostatistical methodology. In this context, we propose employing a *p*-values-based method, taking into account the stochastic nature of *p*-values. We 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 extend the EPV concept to be considered in terms of the ROC curve technique. This provides expressive evaluations and visualizations of a wide spectrum of testing mechanisms' properties. We show that the conventional power characterization of tests is a partial aspect of the presented EPV/ROC technique. We desire that this explanation of EPV/ROC approach convinces researchers of the usefulness of the EPV/ROC approach for depicting different characteristics of decision-making procedures, in light of the growing interest regarding correct *p*-values-based applications.

Keywords: AUC; Expected p-value; P-value; Partial AUC; Partial Expected p-Value; Power; ROC curve; *t*-Test; Wilcoxon test.

1. INTRODUCTION

Albert Vexler is Professor, Department of Biostatistics, 715 Kimball Tower, 3435 Main St, The State University of New York, Buffalo, NY 14214, email: avexler@buffalo.edu
Jihnhee Yu is Associate Professor, Department of Biostatistics, 708 Kimball Tower, 3435 Main St, The State University of New York, Buffalo, NY 14214, email: jinheeyu@buffalo.edu

In this article, we introduce and extend a simple and objective statistical technique called the expected p-value (EPV) that allows us to compare characteristics associated with test statistics of interest. As an illustrative example, we reassess the performance of test statistics that are almost habitually used in every day statistical decision making (e.g., t-test and Wilcoxon test type procedures). In order to be concrete, we exemplify our points using a case-control study statement of problem. The central idea of the case-control study is the comparison of a group having the outcome of interest to a control group with regard to one or more characteristics. In health related experiments, the case group usually consists of individuals with a given disease, whereas the control group is disease free. Consider a biomarker example with myocardial infarction (MI). MI is commonly caused by blood clots blocking the blood flow of the heart leading heart muscle injury. Heart disease is a leading cause of death, affecting about 20% of population, regardless of ethnicity, according to the Centers for Disease Control and Prevention (e.g., Schisterman et al., 2001, 2002). The use of biomarkers to assist medical decision making, the diagnosis and prognosis of individuals with a given disease, is increasingly common in both clinical settings and epidemiological research. This has spurred an increase in exploration for and development of new biomarkers. The biomarker high density lipoprotein (HDL)-cholesterol is often used as a discriminant factor between individuals with and without MI disease. The HDL-cholesterol levels can be examined from a 12-hour fasting blood specimen for biochemical analysis at baseline, providing values of measurements regarding HDL biomarker to be collected on cases who survived an MI and on controls who had no previous MI. Note that, oftentimes measurements related to biological processes follow a log-normal distribution (see for details Limpert, 2001; Vexler et al., 2016: pp. 13-14). Thus, one may be interested in how often a log-transformed HDL cholesterol level of the case group, say X, outperforms a log-transformed

HDL cholesterol level of the case group, Y. Typically, this research statement is associated with the measure $\Pr(X > Y)$ that is assumed to be examined using n independent and normally distributed data points $X_1,...,X_n$ as well as m independent and normally distributed observations $Y_1,...,Y_m$ (e.g., Vexler et al., 2008). In this scenario, in order to test the hypothesis $H_0 : \Pr(X > Y) = 0.5$ versus $H_1 : \Pr(X > Y) > 0.5$, the traditional statistical literature commonly suggests using t-test type procedures (e.g., Browne, 2010). Researchers are encouraged to apply t-test type decision making mechanisms when the underlying data follow a normal distribution.

Student's *t*-test statistic and Welch's *t*-test statistic are mainstays in statistical practice and are introduced in most introductory statistical classes. Student's *t*-test statistic has the form

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

and Welch's *t*-test statistic is

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

where $\bar{X}$ is the sample mean based on $X_1,...,X_n$, $\bar{Y}$ is the sample mean based on $Y_1,...,Y_m$, $S_1^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ and $S_2^2 = \sum_{j=1}^{m}(Y_i - \bar{Y})^2/(m-1)$ are the unbiased estimators of the variances $\sigma_1^2 = Var(X_1)$ and $\sigma_2^2 = Var(Y_1)$, respectively and $S_p^2 = \{(n-1)S_1^2 + (m-1)S_2^2\}/(n+m-2)$ is the pooled sample variance.

In the two sample framework, if we observe normally distributed data points, the choice would be made between Student's *t*-test or Welch's *t*-test (e.g., Julious, 2005; Zimmerman and Zumbo, 2009), but it is anticipated that T_S-and T_W-based tests are somewhat better than the corresponding Wilcoxon rank-sum test (e.g., Ahmad, 1996). The Wilcoxon rank-sum test, which

is a nonparametric test, is generally recommended when the data are assumed to be from a non-normal distribution. In this paper, we target to propose a simple approach for examining this stereotypical view.

Conventional statistical power comparisons of various statistical procedures can hardly lead to a consistent decision that one method is preferred to other methods over a range of scenarios. The choice of a desired user-specified significance level α can largely affect the power properties of tests. Oftentimes, one method is more powerful than others with a certain α, but this conclusion would not be maintained with different values of α. Toward this end, an alternative idea can be used, i.e., the concept based on p-values in the comparison of test procedures. For this purpose, we need to consider that the p-value as a function of the data is a random variable with the probability distribution. This fact may hamper the usage of the p-value to examine the performance of statistical test procedures (e.g., Wasserstein and Lazar, 2016). The distribution of the p-value is conditional on either the null hypothesis being true or not, which needs to be taken into account to interpret the magnitude of the relevant p-value. That is, under the null hypothesis, typically, p-values exactly (or asymptotically) have the Uniform(0,1) distribution. Under the alternative hypothesis, a non-Uniform(0,1)-distribution will be assumed for the p-values where multiple factors including the difference between the null and true parameters and sample size affect the distribution. Dempster and Schatzoff (1965) proposed the concept of the expected significance level addressing the stochastic aspect of the p-value. The term EPV was coined by Sackrowitz and Samuel-Cahn (1999) who further investigated the approach relative to the expected significance level. In their paper, they touted the various potential usages of the EPV. Vexler et al. (2017) advanced the concept of the EPV, especially presenting the strong tie between the EPV concept and receiver operating characteristic (ROC)

curve methodology, a popular biomarker discriminant analysis tool (e.g., Vexler et al, 2016). Such a relationship between the EPV and ROC curve comes in handy for assessing and visualizing the properties of various decision-making procedures in the p-value-based context, since advanced methodologies relative to ROC curve and area under the ROC curve (AUC) are readily adaptable for EPV methodologies. This approach was successfully applied to construct optimal multiple testing procedures (Vexler et al., 2017). In this paper, we propose to use a partial expected p-value (pEPV) as a simple method for comparing statistical tests in an ROC curve framework especially with the partial AUC technique addressed extensively in the biostatistical literature. We will demonstrate that the presented EPV/ROC technique encompasses the conventional power characterization of tests.

This paper is organized as follows. Section 2 sets the notations related to the ROC curve methodology and the EPV definition. The key is that it is conceptually straightforward to associate the ROC curve tools with the EPV concept. The ROC-EPV connection implies new tools for examining statistical decision making procedures. In Section 3, we exemplify how an application of the EPV/ROC approach can show a critical concern regarding the t-tests' optimality. An applicability of the proposed method is illustrated through a real-life example of myocardial infarction disease in Section 4. In Section 5 we provide concluding remarks.

2. THE ROC CURVE, AUC AND EPV TERMINOLOGIES

2.1. ROC curve and AUC

The ROC curve analysis is a popular tool to assess the discriminability of different biomarkers. Briefly speaking, the ROC curve is a method to summarize and depict distance between two distribution functions. Suppose that random variables X and Y are from the continuous distribution functions F_X and F_Y. In a typical setting of a biomarker study (e.g., the MI disease

study introduced in Section 1), often we impose the meaning on the variables that X and Y are biomarker values from diseased and non-diseased subjects, respectively. Now, the ROC curve has the form

$$ROC(t) = 1 - F_X\{F_Y^{-1}(1-t)\}, 0 < t < 1, \quad (1)$$

where F_Y^{-1} represents the inverse or quantile function of F_Y, such that, $F_Y(F_Y^{-1}(\gamma)) = \gamma$, $0 < \gamma < 1$. Definition (1) clearly shows that the ROC curve is a special case of a probability-probability plot (P-P plot) (e.g., Vexler et al., 2016). In the plot of the points $\left(ROC(t), t\right), 0 < t < 1$, the farther apart the two distributions F_X and F_Y in terms of location gives rise to the more the ROC curve shift to the top left corner. With a biomarker that separates the diseased and non-diseased subjects well, the ROC curve will be coming close to the top left corner. If a biomarker has no discriminability, a diagonal line from the points (0,0) to (1,1) would be shown for the ROC curve. The ROC curve places tests (biomarkers values) on the same scale in the comparison of accurate discriminant ability. Such a feature allows us to assess different diagnostic biomarkers conveniently.

One summary index of the ROC curve is the AUC. The AUC expresses the overall performance of a biomarker, indicating that a larger value of the AUC implies a more accurate discriminating ability of a given marker. That is, values of the AUC vary from 0.5, in the case of no differentiation between the diseased and non-diseased patients, to 1, where the diseased and non-diseased patients are perfectly separated. The AUC can be expressed in the following succinct form (Bamber, 1975)

$$\text{AUC} = \int_0^1 ROC(t)dt = \Pr(X > Y).$$

Now, let us consider the partial area under the ROC curve (pAUC), the area under a part of the ROC curve. Specifically, the pAUC with two fixed a priori values t_0 and t_1 is expressed as

$$pAUC = \int_{t_0}^{t_1} ROC(t)dt .$$

The partial AUC summarizes the area of interest only under a part of the ROC curve, rather than summarizing the entire ROC curve.

2.2. EPV

Let $T(D)$ denote a test statistic, which is a random variable, depending on data *D*. We define its associated distribution F_i under the hypothesis $H_i, i = 0,1$, where the subscript *i* indicates the null (*i=0*) and alternative (*i=1*) hypotheses, respectively. With continuous F_i, we denote F_i^{-1} to be the quantile function of F_i, thus, $F_i(F_i^{-1}(\gamma)) = \gamma$, *i=0,1*. In this setting, without the loss of generality, we consider tests of the form: the event $T(D) > C$ rejects H_0, where *C* is a prefixed test threshold. The corresponding p-value is $1 - F_0(T(D))$. We define independent random variables T^0 and T^A that have distributions F_0 and F_1, respectively. Noting that

$$E(1 - F_0(T(D)) \mid H_1) = \int \{1 - F_0(u)\} dF_1(u) = \int \Pr\left(T^0 \geq u\right) dF_1(u) = \Pr\left(T^0 \geq T^A\right),$$

Sackrowitz and Samuel-Cahn (1999) proved that the expected p-value is

$$\text{EPV} = \Pr(T^0 \geq T^A). \quad (2)$$

The relationship (2) clearly indicates the strong connection between EPV and AUC. In the consideration of the fact that the ROC curve depicts a distance between the distribution functions F_0 and F_1, the relationship (2) leads us to rethink the EPV in the context of AUC, obtaining that EPV is 1-AUC.

Further, the value of 1-EPV has an interpretation of the uniform integration of the statistical power of a test over the range of significance level α from 0 to 1 as presented in the following notation,

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

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

A possible caveat of expression (3) is that the EPV summarize the power over the entire range of the significance level α where the most of the values of α are not of interest since they are not traditionally used in practice of statistical decision making (e.g. $\alpha > 0.1$). Instead, adapting the concept of the pAUC, the power can be integrated over significance levels of α in a specific interesting range. For some fixed upper level $\alpha_U \leq 1$, we have

$$pEPV = 1 - \int_0^{\alpha_U} \Pr\left\{p - value \leq \alpha \mid H_1\right\} d\alpha = 1 - \int_0^{\alpha_U} \Pr\left\{1 - F_0(T^A) \leq \alpha\right\} d\alpha$$

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

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

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

Generally, one can define the function $pEPV(\alpha_L, \alpha_U) = 1 - \int_{\alpha_L}^{\alpha_U} \Pr\left\{p - value \leq u \mid H_1\right\} du$.

Consider $\frac{d}{d\alpha}\left\{-pEPV(0,\alpha)\right\}$ that implies the power at a significance level of α. An essential property of efficient statistical tests is unbiasedness. A statistical test is unbiased when the probability of committing a Type I error is less than the significance level and a proper power is

greater than the significance level, i.e. Pr(reject $H_0 \mid H_0) \le \alpha$ and Pr(reject $H_0 \mid H_0$ is not true) $\ge \alpha$. In parallel with this definition, it is reasonable to consider the inequality

$$pEPV(0,\alpha) \le 1 - \int_0^{\alpha} \Pr\{p - value \le u \mid H_0\} du = 1 - \alpha^2 / 2,$$

since $p - value \sim Uniform[0,1]$ (i.e. $\Pr\{p - value \le u \mid H_0\} = u, u \in [0,1]$) under H_0 and we assume $H_1 \ne H_0$. In this case,

$$\frac{d}{d\alpha}\{pEPV(0,\alpha)\} = -\Pr(\text{reject } H_0 \mid H_0 \text{ is not true}) \text{ and } \frac{d}{d\alpha}\{1 - \alpha^2/2\} = -\alpha.$$

when Pr(reject $H_0 \mid H_0) = \alpha$. However, it is clear that the requirement $pEPV(0,\alpha) \le 1 - \alpha^2/2$ is weaker than that of Pr(p− value $< \alpha \mid H_0$ is not true) $\ge \alpha$. Thus, the EPV based concept extends the conventional power characterization of tests.

The EPV approach can provide an alternative approach to the Neyman-Pearson concept of testing statistical hypotheses (e.g., Vexler et al., 2016). The EPV corresponds to the integrated power of a test via all possible values of $\alpha \in (0,1)$ evaluating the performance of the test procedure globally. Smaller values of EPV indicate superior qualities of tests in a universal fashion. As opposed to this, the Neyman-Pearson lemma uses the concept that a viable statistical testing procedure maintains the Type I error rate under a user-specified significance level, α, together with maximizing the power in a uniform fashion. Consequently, superior test procedures may be different in general for different values of α. On the other hand, the EPV based approach enables comparison between decision-making rules to be more objective. Also, the EPV, a single number tool for assessing performance of testing procedures can rank-order the different procedures more easily.

3. COMPARISON OF THE TEST STATISTICS

In this section, we demonstrate the comparison of T_S, T_W and the Wilcoxon rank-sum test statistic using the EPV/ROC approach. The results shown in this section can be obtained analytically, since the distribution functions of the statistics T_S, T_W and the Wilcoxon rank-sum test statistic under H_0 and H_1 have specified forms. Alternatively, we provide the following R Code (R Development Core Team, 2002) that can be easily modified to be applied in order to evaluate various decision making procedures using the accurate Monte Carlo approximations to the EPV/ROC instruments. To exemplify the proposed approach for comparing the test statistics, the R Code below provides simulation-based evaluations of the ROC curves based on values of the test statistics related to the null and alternative hypotheses, using the R built-in procedure *pROC*. In this scenario, we assume, for example, that $X_1,...,X_n \sim N(0,1)$ and $Y_1,...,Y_m \sim N(0,1)$ under H_0, whereas $X_1,...,X_n \sim N(0,1)$ and $Y_1,...,Y_m \sim N(-0.5,10)$.

```
library(pROC)
N<-100000 #Number of the Monte Carlo data generations
W0<-array()
T0<-array()
W1<-array()
T1<-array()
n<-25 #The sample size n
m<-25 #The sample size m

for(i in 1:N){
x0<-rnorm(n,0,1) #Values of X from N(0,1) generated under the hypothesis H0
y0<-rnorm(m,0,1) #Values of Y from N(0,1) generated under the hypothesis H0
x1<-rnorm(n,0,1) #Values of X from N(0,1) generated under the hypothesis H1
y1<-rnorm(m,-0.5,10)  #Values  of  Y  from  N(-0.5,10)  generated  under  the
hypothesis H1

W0[i]<-wilcox.test(x0,y0,alternative = c(greater))$stat[[1]]/(n*m)
#Values of the Wilcoxon test statistic under H0
W1[i]<-wilcox.test(x1,y1,alternative = c(greater))$stat[[1]]/(n*m)
#Values of the Wilcoxon test statistic under H1

EV<-FALSE #This parameter indicates the use of Welch's t-test statistic
#EV<-TRUE #This parameter indicates the use of Student's t-test statistic
```

```
T0[i]<-t.test(x0,y0,alternative = c(greater),var.equal = EV)$stat[[1]]
#Values of the t-test statistic under H0
T1[i]<-t.test(x1,y1,alternative = c(greater),var.equal = EV)$stat[[1]]
#Values of the t-test statistic under H1
}

#Plotting the ROC cureves
Ind<-c(array(1,N),array(0,N))
W<-c(W0,W1)
T<-c(T0,T1)
plot.roc(Ind, W,type=l, legacy.axes=TRUE,xlab=t,ylab=ROC(t))
lines.roc(Ind,T,col=red,lty=2)
```

Figure 1 shows the obtained ROC curves $\left(ROC_T(t),t\right)$ and $\left(ROC_W(t),t\right)$, where

$$ROC_T(t)=1-F_{T_1}\{F_{T_0}^{-1}(1-t)\} \text{ and } ROC_W(t)=1-F_{W_1}\{F_{W_0}^{-1}(1-t)\}, t\in(0,1)$$

with the distribution functions F_{T_k} and F_{W_k} that correspond to the t-test statistic and the Wilcoxon rank-sum test statistic distributions under $H_k, k=0,1$, respectively.

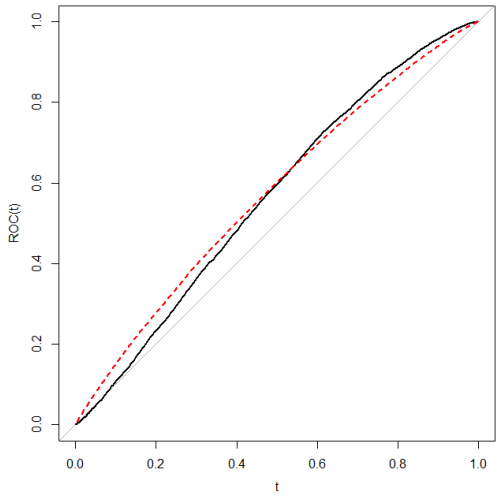

Figure 1. Values of the functions $ROC_T(t)$ (curve '- - - -') and $ROC_W(t)$ (curve '——') plotted against $t\in(0,1)$.

In the executed R Code, we focus on Welch's *t*-test statistic (the parameter `EV<-FALSE`), which is reasonable in the considered setting of data distributions' parameters. It is interesting to

remark that when examining Student's t-test statistic (the parameter `EV<-TRUE)` the graphs show that there are no significant differences between the relative curves. Similar observations are in effect, regarding the considerations shown below.

Analyzing EPV's for the one-sided, two-sample *t*- and Wilcoxon tests based on normally distributed data points ($n = m = 10, 20, 50$), Sackrowitz and Samuel-Cahn (1999) concluded that *The t test is best both for the Normal distribution (not surprising!) and the Uniform distribution*. In order to compute the EPV's corresponding to the considered example, we can execute the following code.

```
Troc<-roc(Ind, T)
Wroc<-roc(Ind, W)
EPV_t<-1-auc(Troc) # EPV of the t-test
EPV_W<-1-auc(Wroc) # EPV of the Wilcoxon test
```

Indeed, the computed EPV of the *t*-test is 0.431 that is smaller than 0.439 of that related to the Wilcoxon test. However, Figure 1 demonstrates that for $t \in (0, 0.5)$ the Wilcoxon test is somewhat better that the *t*-test. This motivates us to employ the pEPV for this analysis. Towards this end, we denote the function

$$G(\alpha) = \{ pEPV_W(0,\alpha) - pEPV_t(0,\alpha) \} / \, pEPV_t(0,\alpha) ,$$

where $pEPV_T(0,\alpha)$ and $pEPV_W(0,\alpha)$ are the function $pEPV(0,\alpha)$ defined in Section 2.2 and computed with respect to the t-test and the Wilcoxon rank-sum test, respectively. In order to depict the result we use the following code.

```
pEPV_t<-function(u) 1-auc(Troc, partial.auc=c(0,u))[[1]]
pEPV_W<-function(u) 1-auc(Wroc, partial.auc=c(0,u))[[1]]
G<-function(u) (pEPV_W(u)-pEPV_t(u))/(pEPV_t(u))
GV<-Vectorize(G)
plot(GV,0,1,xlab='alpha',ylab='G(alpha)')
```

Figure 2 presents the curve of $G(\alpha)$.

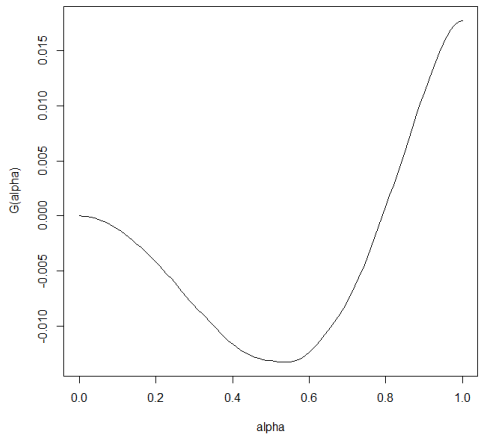

Figure 2. The relative comparison between the t-test and the Wilcoxon rank-sum test using their pEPV's via the function $G(\alpha)$ plotted against $\alpha \in (0,1)$.

In this case, it is clear that the Wilcoxon test outperforms the t-test, when the significance level $\alpha < 0.8$.

Let us fix $\alpha = 0.05$ and calculate the corresponding powers of the tests using the following code.

```
Wc<-quantile(W0,0.95) # the 95% critical value of the Wilcoxon test
Tc<-quantile(T0,0.95) # the 95% critical value of the t- test
PowW<-mean(1*(W1>=Wc)) #the power of the Wilcoxon test
PowT<-mean(1*(T1>=Tc)) #the power of the t-test
print(c(PowT,PowW))
```

We obtain that the power of the Wilcoxon test is 0.12, whereas the power of the t-test is 0.08.

Remark 1. In this study we applied the simulation-based computations via 100,000 Monte Carlo repetitions of the underlying data points. Assume a probability type parameter p is evaluated using 100,000 Monte Carlo repetitions of data points. In this case, we can anticipate the Monte Carlo error in order of $\pm 1.96\left(p(1-p)/100000 \right)^{1/2}$, taking into account the central limit theorem.

We also used 1,000,000 Monte Carlo repetitions in the context of the analysis shown in this section. The obtained results were very close to those presented in this note.

Remark 2. Assume we are interested in the measure $\Pr(X > Y)$. A fast way to evaluate $\Pr(X > Y)$ can be based on the R command

```
wilcox.test(X,Y,alternative=c(greater))$stat[[1]].
```

In the setting of the example mentioned in this section, one can use

```
nn<-5000000
x1<-rnorm(nn,0,1)
y1<-rnorm(nn,-0.5,10)
wilcox.test(x1,y1,alternative = c(greater))$stat[[1]]/(nn*nn)
```

In our simulation, the code above gives an approximated value of $\Pr(X > Y)$ under H_1 as 0.5192665 that corresponds to $\Pr(X > Y) = (2\pi)^{-1/2} \int_{-\infty}^{0.5} e^{-z^2/22} dz \approx 0.5181275$.

Remark 3. In various scenarios with $Var(X_1) = Var(Y_1)$ under H_1, it was observed using the function $G(\alpha), \alpha < 0.1,$ that the t-test procedures and the Wilcoxon rank-sum test provide approximately same properties.

4. DATA EXAMPLE

In this section, a real-life data example introduced in Introduction Section is presented in order to illustrate an applicability of the proposed method. The data set is 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 HDL-cholesterol. A

total of 61 measurements of HDL-cholesterol biomarker were evaluated by the study, when $n = 28$ data points were collected on cases who survived on MI and the other $m = 33$ on controls who had no previous MI.

Figure 3 depicts the histograms based on values of the log-transformed HDL cholesterol levels, $X_1,...,X_{28}$ and $Y_1,...,Y_{33}$, respectively. The Shapiro-Wilk test for normality provides the p-values 0.5799 and 0.1581 corresponding to $X_1,...,X_{28}$ and $Y_1,...,Y_{33}$, respectively.

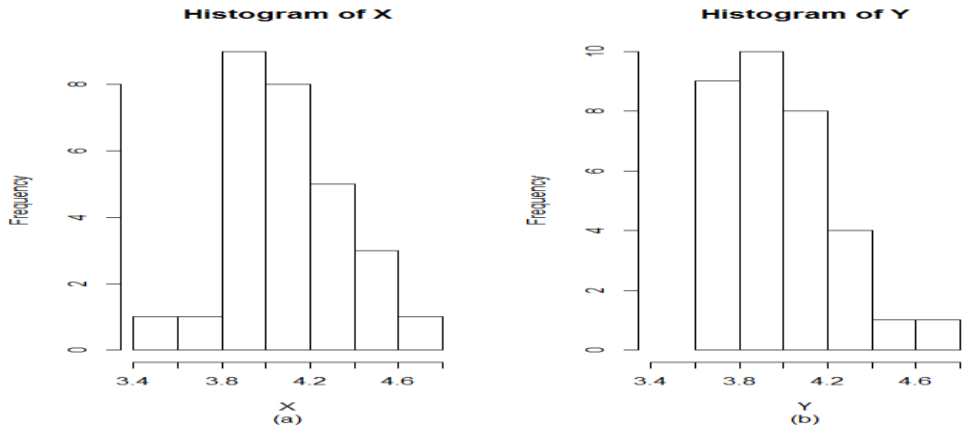

Figure 3. Histograms of the log-transformed biomarkers of interest corresponding to HDL-cholesterol cases (panel (a): the estimated mean and the estimated standard deviation of log(HDL-cholesterol) are 4.090 and 0.259, respectively;) and HDL-cholesterol controls (panel (b): the estimated mean and the estimated standard deviation of log(HDL-cholesterol) are 3.938 and 0.236, respectively).

The R code presented in Section 3 can be easily modified to provide the EPV/ROC analysis based on the real data. To this end the variables 'x0', 'y0' can be simulated corresponding to H_0 (e.g., as $X_1,...,X_n \sim N(0,1)$ and $Y_1,...,Y_m \sim N(0,1)$) and the variables 'x1', 'y1' can be sampled from the observed $X_1,...,X_n$ and $Y_1,...,Y_m$ in a bootstrap manner at each loop iteration in 'for(i in 1:N){…}'. Executing this procedure (with N<-50,000 simulations), we obtain the graphs shown in Figure 4 in parallel with Figures 2 and 3.

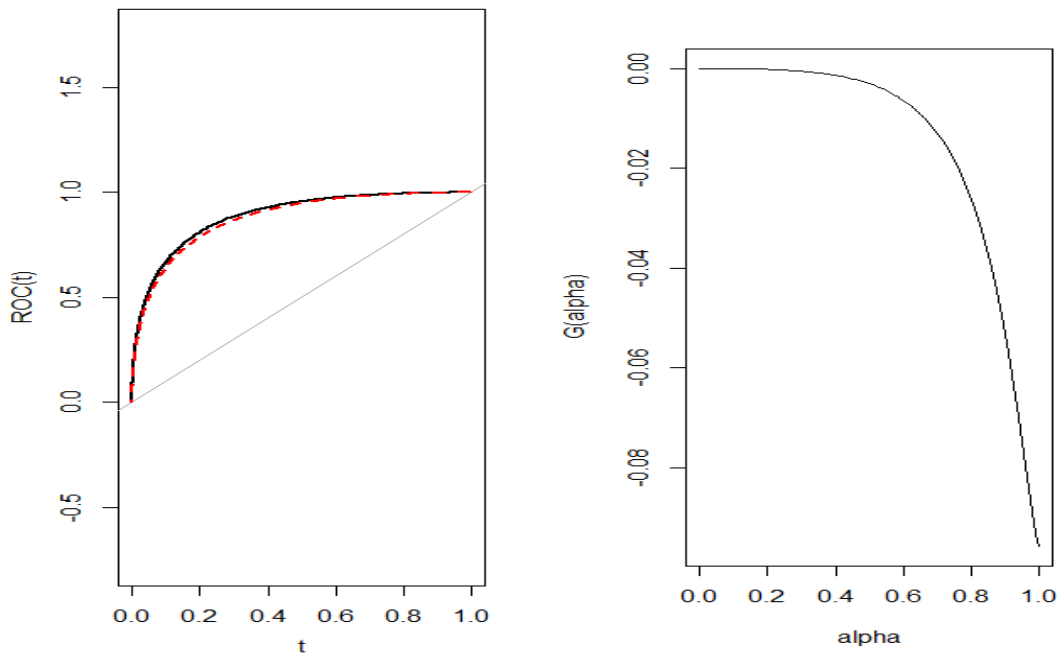

Figure 4. The HDL-cholesterol data based graphs related to 1) the functions $ROC_T(t)$ (curve '- - - -') and $ROC_W(t)$ (curve '——') plotted against $t \in (0,1)$; 2) the relative comparison between the t-test and the Wilcoxon test via their pEPV's, the function $G(\alpha)$ plotted against $\alpha \in (0,1)$.

In conjunction with the explorations presented in Section 3, Figure 4 induces applying the Wilcoxon rank-sum test rather than the t-test in the context of the HDL-cholesterol case control study. Note that, in this case, the Wilcoxon rank-sum test and Welch's *t*-test demonstrate the p-values of 0.045 and 0.0555, respectively.

5. CONCLUDING REMARKS

Our objective has been to show a simple technique that can provide intrinsic evaluations of statistical decision making strategies. The proposed approach involves correct p-value-based mechanisms. We have exemplified scenarios when the EPV/ROC concept yields a critical concern regarding uses, without doubt, of t-test type procedures, when underlying data are normally distributed. It has been demonstrated that the nonparametric Wilcoxon rank-sum test may clearly outperform the *t*-tests. Our desire to incorporate pre-analyses of test procedures

before their use in practical applications provided the EPV/ROC method, presuming its place in the toolkit of the well-equipped statistician. We have seen that the EPV/ROC technique is a very useful and succinct measurement tool of the performance of decision-making mechanisms.

There are many possible avenues of research that can be done in the context of the EPV/ROC methodology. One in particular is that asymptotic theoretical results can be developed with respect to the EPV/ROC concept in parametric and nonparametric frameworks. We also anticipate that efficient data-driven Bayesian type methods can be derived in order to assess test properties in terms of the EPV/ROC frame. These topics can warrant further strong empirical and methodological investigations.

ACKNOWLEDGEMENTS

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

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

REFERENCES

Ahmad, I. A. (1996). A Class of Mann-Whitney-Willcoxon Type Statistics. *The American Statistician*, **50**, 324-327.

Bamber, D. (1975). The Area above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph. *Journal of Mathematical Psychology* **12**(4): 387-415.

Browne, R. H. (2010). The t-test p Value and its Relationship to the Effect Size and P(X>Y). *The American Statistician,* **64:1**, 30-33.

Dempster, A. P. and Schatzoff, M. (1965). Expected significance level as a sensitivity index for test statistics. *Journal of the American Statistical Association*. **60**, 420-436.

Julious S. A. (2005). Why do we use pooled variance analysis of variance? *Pharmaceutical Statistics,* 4: 3-5.

Limpert, E., Stahel, W. A. and Abbt, M. (2001). 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,* **51**(5): 341-352.

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

Sackrowitz, H. and Samuel-Cahn, E. (1999). P values as random variables-expected p values. *The American Statistician.* **53**, 326-331.

Schisterman, E. F., Faraggi, D., Browne, R., Freudenheim, J., Dorn, J., Muti, P., Armstrong, D., Reiser, B. and Trevisan, M. (2001). Tbars and Cardiovascular Disease in a Population-Based Sample. *Journal of Cardiovascular Risk,* **8**(4): 219-225.

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

Vexler, A., Liu, A., Eliseeva, E. and Schisterman, E. F. (2008). Maximum Likelihood Ratio Tests for Comparing the Discriminatory Ability of Biomarkers Subject to Limit of Detection. *Biometrics,* **64**(3): 895-903.

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

Vexler, A., Yu, J, Zhao, Y. Hutson, A. D. and Gurevich, G. (2017). Expected P-values in Light of an ROC Curve Analysis Applied to Optimal Multiple Testing Procedures. *Statistical Methods in Medical Research.* DOI: 10.1177/0962280217704451. In Press.

Zimmerman, D. W and Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variance t tests. *Psiologica,* **30**, 371-390.

Wasserstein RL and Lazar N. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, **70,** 129-133.