Pak. J. Statist.
2020 Vol. 36(1), 13-28

AREA UNDER THE ROC CURVE AS EFFECT SIZE MEASURE

Pablo Martinez-Camblor[1§], Sonia Perez-Fernandez[2] and **Susana Diaz-Coto[2]**
[1] Biomedical Data Sciences Department, Geisel School of Medicine at Dartmouth, USA. Email: pablo.martinez.camblor@dartmouth.edu
[2] Department of Statistics, Oviedo University, Spain
Email: perezsonia@uiovi.es, susanadcoto@hotmail.com
[§] Corresponding author

ABSTRACT

The area under the receiver-operating characteristic (ROC) curve, AUC, is frequently used as diagnostic capacity measure and also as a goodness of fit index in logistic regression models. Given a considered (bio) marker, it measures the difference in the location parameters between two independent groups, each of them containing the positive and the negative subjects. Therefore, it can be interpreted as the *effect size* of the group on the studied variable (biomarker). In the present manuscript, we study the AUC interpretation within the two-sample problem. Both the non-parametric and the parametric tests join with the case where the marker is a categorical variable are considered. The use of the AUC as *effect size* provides a measure which is comparable and interpretable in different context: parametric, non-parametric and categorical cases are considered. Finally, in order to illustrate the problem, a real-world dataset is explored.

KEYWORDS

Area under the ROC curve; Effect size; Mann-Whitney test; ROC curve; T-test; Two-sample test.

AMS Classification.

1. INTRODUCTION

Since it appeared in the World War II in the radar signal detection context, the receiver-operating characteristic, ROC, curve has been deeply studied and employed in a vast variety of knowledge fields (Green and Sweet, 1966). Probably, the most active areas of investigation are machine learning (see, for instance, Fawcett; 2006) and biomedicine (Pepe, 2003) among many others). Both theoretical and practical aspects of the ROC curve have been deeply considered; the reader is referred to the essential monograph of Zhou, Obuchowski and McClish (2002) for a complete overview and to Martinez-Camblor (2017) for a most recent revision. In addition, several generalizations have been proposed: for instance, Mossman (1999) proposed a ROC surface useful in trichotomous decision tasks; Heagerty, Lumley and Pepe (2000) dealt with the case of time-dependent outcomes and Martinez-Camblor et al. (2017) proposed a ROC curve generalization for non-monotone relationship. Moreover, although other measures have been proposed (see, for instance Hilden (1991) or Ma et al. (2013)), the area under the ROC curve, AUC, is still the most

popular index for summarizing diagnostic capacity (Faraggi and Reiser (2002)); it can also be (and, in fact, it is) used as goodness of fit measure in logistic regression models.

The standard ROC curve construction involves two independent populations (one of positive and another one of negative subjects) and, given a studied (bio) marker, the AUC measures how different their location parameters are. As it is well-known, the AUC stands for the probability that the value of the marker in a randomly chosen positive subject will be higher than the value of the marker in a randomly chosen negative subject (this point will be deeper discussed in section 2). Hence, when we are studying the behavior of a variable on two different groups; i.e., when we are dealing with a conventional two-sample problem, the AUC can also be considered as a measure of how different the location parameters are; i.e., as an *effect size* measure of the group on the studied variable.

This work deals with the area under the ROC curve from the two-sample test approach. Main objective is not to develop novel results but revendicate the use of the AUC as effect size measure and pointing out the good behavior of it in a vast range of situations. Rest of the paper is organized as follows: in section 2 some theoretical properties of both the receiver-operating characteristic curve and the AUC are pointed out. Section 3 is devoted to the study of the non-parametric estimation for the AUC and its relationship with the non-parametric Mann-Whitney test. In section 4, the traditional Student-Welch test is related with the plug-in AUC estimator. Categorical variables are investigated in section 5 and, in section 6, the advantages of using the AUC as effect size measure is illustrated from real-world application. Finally, in Section 7, we presented our main conclusions. Technical aspects and simulation results are provided as appendix.

2. THE ROC CURVE AND THE AUC

Let χ and ξ be two random variables representing the values of the studied continuous measure, marker, in two different groups which contain, for the moment, the negative and the positive subjects, respectively. The ROC curve, $\mathcal{R}(\cdot)$, is a plot of the sensitivity (S_E); i.e. the ability of the marker to classify positive subjects as positives versus the complementary of the specificity ($1 - S_P$) of this marker; i.e., the inability of the test to recognize negative subjects as negative. Equivalently, the ROC curve is the geometric place of the points $\{1 - F_\chi(u), 1 - F_\xi(u)\}$ for each u $\in \mathbb{R}$, where $F_\chi(\cdot)$ and $F_\xi(\cdot)$ stand for the cumulative distribution function (CDF) for χ and ξ, respectively. And hence, for each $t \in [0.1]$,

$$\mathcal{R}(\mathrm{t}) = 1 - F_\xi\left(F_\chi^{-1}(1-t)\right). \tag{1}$$

Martinez-Camblor, Carleos and Corral (2011) related the ROC curve with the CDF by the equality,

$$\begin{aligned}\mathcal{R}(t) &= 1 - F_\xi\left(F_\chi^{-1}(1-t)\right) \\ &= \mathcal{P}\{\xi > F_\chi^{-1}(1-t)\} = \mathcal{P}\{1 - F_\chi(\xi) \leq t\} = F_{1-F_\chi(\xi)}(\mathrm{t}),\end{aligned}$$

hence, the ROC curve inherits the CDF properties. In addition, it is invariant under monotone increasing transformations of the measurement scale. Figure 1 depicts the densities for two normal distributed variables, left, and the resulting ROC curve, right.

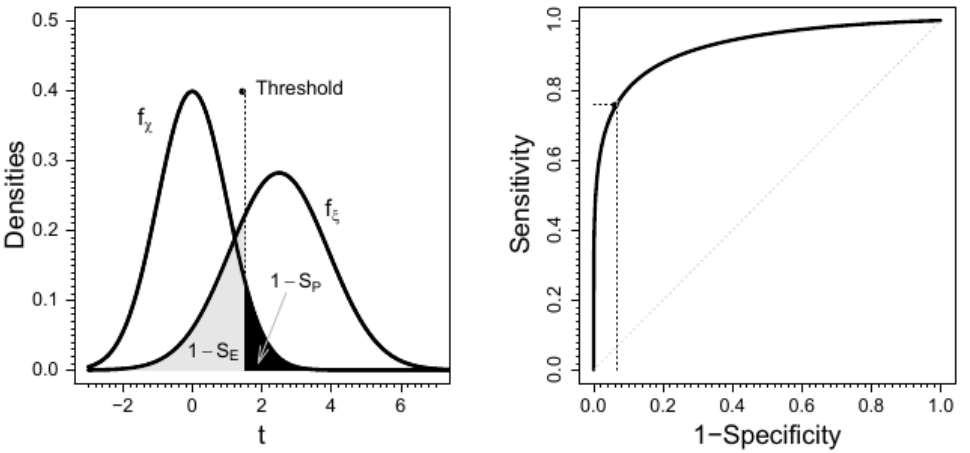

Figure 1: At Left, Densities for the Two Involved Populations, Both Follow Normal Laws. At Right, Resulting ROC Curve

The area under the ROC curve, $\mathcal{A}$, is directly defined as integral of the ROC curve between 0 and 1; i.e.,

$$\mathcal{A} = \int_0^1 \mathcal{R}(t)dt = 1 - \int F_\xi(t)dF_\chi(t) = 1 - \mathbb{E}[\mathcal{P}\{\xi \leq t|\chi = t\}] = \mathcal{P}\{\chi \leq \xi\} \tag{2}$$

The ROC curves comparison is frequently performed from the direct AUC comparisons. Besides, validity of the ROC curve is checked from the null $H_0: \mathcal{A} = 1/2$. Note that, in general, rejecting this hypothesis implies that $F_\xi(t) \neq F_\chi(t)$, however the inverse is not true; $F_\xi(t) \neq F_\chi(t)$, does not imply that $\mathcal{A} \neq 1/2$. By assuming that both χ and ξ are normally distributed (this assumption can be replaced by a weaker one, see Section 4) with means μ_χ and μ_ξ, respectively, then by equation (2),

$$\mathcal{A} = \mathcal{P}\{\chi \leq \xi\} = \mathcal{P}\left\{\mathcal{N}(0,1) \leq (\mu_\xi - \mu_\chi)/\sqrt{(\sigma_\xi^2 + \sigma_\chi^2)}\right\} = \Phi\left(\frac{\mu_\xi - \mu_\chi}{\sqrt{\sigma_\xi^2 + \sigma_\chi^2}}\right) \tag{3}$$

where σ_ξ^2 *and* σ_χ^2 are the variances of the random variables ξ and χ, respectively, and $\Phi(\cdot)$ stands for the CDF of a standard normal distributed random variable. Therefore, in this case $H_0: \mathcal{A} = 1/2$ and $H_0^*: \mu_\chi \neq \mu_\xi$ are equivalent hypotheses. In general, there is not a direct relationship between H_0 and the medians equality. However, difference between the medians often implies $\mathcal{A} = 1/2$ and vice versa.

3. NON-PARAMETRIC AUC ESTIMATION

Although different smooth estimators for the ROC curve have been proposed (see, for instance, Zou, Hall and Shapiro (1997)), the empirical one; resulting of replacing in the equation (1) the unknown CDFs for their respective empirical cumulative distribution functions (ECDF), is still the most frequently used estimator. Then, let $\mathrm{X_n} = \{\mathrm{x_1}, \dots, \mathrm{x_n}\}$ and $Y_m = \{y_1, \dots, y_m\}$ be two independent samples drawn from the random variables χ and ξ, respectively, for each $t \in [0.1]$, the empirical ROC curve estimator, $\hat{\mathcal{R}}(\cdot)$ $(= \hat{\mathcal{R}}_{m,n}(\cdot))$, is defined by

$$\hat{\mathcal{R}}(\mathrm{t}) = 1 - \hat{F}_{\mathrm{m}}\left(\mathrm{Y_m}, \hat{F}_n^{-1}(\mathrm{X_n}, 1-\mathrm{t})\right), \tag{4}$$

where $\hat{F}_m(\mathrm{Y_m}, \cdot) = (1/\mathrm{m})\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathbb{I}_{(-\infty, \mathrm{x_i}]}(\cdot)$ ($\mathbb{I}$ stands for the usual indicator function) and $\hat{F}_n^{-1}(X_n, \cdot) = inf\{s: \hat{F}_n(X_n, s) \geq \cdot\}$. The properties of $\hat{\mathcal{R}}(\cdot)$ are well-known. Hiseh and Turnbull (1996) proved its uniform consistence. In addition, they enunciated the following result.

Theorem 1:

Under the above notation, if $\mathrm{n/m} \to \lambda > 0$ and $F_\chi(\cdot)$, $F_\xi(\cdot)$ have continuous densities, $f_\chi(\cdot)$, $f_\xi(\cdot)$ respectively, such that $f_\xi\left(F_\chi^{-1}(1-\mathrm{t})\right)/f_\chi\left(F_\chi^{-1}(1-\mathrm{t})\right)$ is bounded on any subinterval (a, b) of $(0,1)$, then there exists a probability space on which one can define sequences of two independent versions of Brownian bridges, $\mathcal{B}_1^{(\mathrm{n})}, \mathcal{B}_2^{(\mathrm{n})}$, such that

$$\sqrt{\mathrm{n}} \cdot \{\hat{\mathcal{R}}(\mathrm{t}) - \mathcal{R}(t)\} = \sqrt{\lambda} \cdot \mathcal{B}_1^{(\mathrm{n})}\left\{F_\xi\left(F_\chi^{-1}(1-\mathrm{t})\right)\right\} + \frac{f_\xi\left(F_\chi^{-1}(1-t)\right)}{f_\chi\left(F_\chi^{-1}(1-t)\right)} \cdot \mathcal{B}_2^{(n)}\{(1-t)\} + o\left(n^{-1/2} \cdot (\log n)^2\right) \text{ a. s}$$

The proof of this theorem is followed from the empirical and quantile process theory (see, for instance Theorem 4.4.1 in Csörgő and Révész (1981) and Theorem 3.2.4 in Csörgő (1983)).

The empirical estimator for the AUC, $\hat{\mathcal{A}}$ $(= \hat{\mathcal{A}}_{m,n})$, is the resulting of replacing the unknown CDFs in the equation (2) for their empirical estimators, then

$$\hat{\mathcal{A}} = 1 - \int \hat{F}_m(Y_m, t)\mathrm{d}\hat{F}_n(X_n, t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \hat{F}_m(Y_m, x_i)$$
$$= 1 - \frac{1}{\mathrm{n} \cdot \mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathbb{I}_{(-\infty, \mathrm{y_j}]}(\mathrm{x_i}) = \frac{1}{\mathrm{n} \cdot \mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathbb{I}_{\{\mathrm{x_i} \leq \mathrm{y_j}\}}, \tag{5}$$

under the above assumptions, $\hat{\mathcal{A}}$ *is* an unbiased estimator for the AUC and, obviously, it is equal to the Mann-Whitney two-sample statistic. Its properties are well-known and, if $F_\xi(t) = F_\chi(t)$, (usual null hypothesis), its asymptotic and exact distributions have already been derived. In addition, Hsieh and Turnbull (1996) enunciated the following result,

Theorem 2:

Under the Theorem 1 assumptions, it holds

$$\sqrt{\mathrm{n}} \cdot \{\hat{\mathcal{A}} - \mathcal{A}\} \rightarrow_{\mathrm{n}} \mathcal{N}(0, \sigma), \tag{6}$$

with $\sigma^2 = ||F_\xi(F_\chi^{-1})|| + \lambda \cdot ||F_\chi(F_\xi^{-1})||$, and where, given an arbitrary measurable function $g(\cdot)$, $||g|| = \int_0^1 g^2 - \left(\int_0^1 g\right)^2$.

□

Computations related with the proof of this result are provided in the Appendix A.

The variance reported in Theorem 2 depends on the unknown CDFs. Hence, in order to make inference on $\hat{\mathcal{A}}$ it must be estimated. Although applying a plug-in method is a possible solution, there is some debate. In Cleves (2002) three different methods for approximating this variance were studied from Monte Carlo simulations; observed results suggested that, although asymptotically equivalents, methods can lead to different estimations for small sample sizes.

In the two-sample context, any group is more relevant than the other; and then, AUCs below 1/2 must not be unusual. The AUC interpretation will be how different the studied variable behavior is within the different considered groups and, therefore, values close to zero or close to one indicate more evidence in favor that the studied behavior is different.

4. AUC ESTIMATION UNDER NORMALITY

Since the ROC curve is invariant under monotone increasing transformations of the measurement scale, i.e., given a monotone increasing transformation, $H(\cdot)$, the ROC curve referred to the variables χ and ξ is the same than the one referred to the variables $H(\chi)$, and $H(\xi)$, due to for each $t \in [0.1]$ directly,

$$\mathcal{R}(t) = 1 - F_\xi\left(F_\chi^{-1}(1-t)\right) = 1 - F_\xi\left(H^{-1}\left(H\left(F_\chi(1-t)\right)\right)\right),$$

the normality is usually replaced by the *binormality* which assumes that there exists such transformation. The binormal assumption is slightly weaker than normality. However, due to the objective of this section is to discuss about the use of the AUC as effect size measure in parametric contexts, we consider that both χ and ξ are normally distributed with means μ_χ, μ_ξ and variance σ_χ^2 and σ_ξ^2, respectively. Applying the plug-in method on equation (3) we obtain an adequate estimator for the real AUC, $\mathcal{A}$. Then let $X_n = \{x_1, \ldots, x_n\}$ and $Y_m = \{y_1, \ldots, y_m\}$ be two independent samples drawn from the normally distributed variables χ and ξ, respectively. Then, for each $t \in [0.1]$, we define the parametric AUC estimator by,

$$\hat{\mathcal{A}}_{\mathcal{N}} = \Phi\left(\frac{\bar{Y}_m - \bar{X}_n}{\sqrt{\hat{S}_{\xi,m}^2 + \hat{S}_{\chi,n}^2}}\right) \tag{7}$$

where $\overline{X_n}$ and $\hat{S}^2_{\chi,n}$ are the mean and the pseudo-variance referred to the sample X_n and $\overline{Y_m}$ and $\hat{S}^2_{\xi,m}$ are the mean and the pseudo-variance sample referred to Y_m. The following result guarantees that $\hat{\mathcal{A}}_{\mathcal{N}}$ is asymptotically normal distributed (the proof is provided in the Appendix A).

Theorem 3:

Under the above conditions, it holds the convergence

$$\sqrt{n} \cdot \{\hat{\mathcal{A}}_{\mathcal{N}} - \mathcal{A}\} \rightarrow_n \mathcal{N}(0, v), \tag{8}$$

with $v^2 = \varphi(p) \cdot \sqrt{\left(\lambda \cdot \sigma_\xi^2 + \sigma_\chi^2\right)/\left(\sigma_\xi^2 + \sigma_\chi^2\right)}$, $\varphi(\cdot)$ denotes the density of a standard normal distribution and $p = \left(\mu_\xi - \mu_\chi\right)/\sqrt{\sigma_\chi^2 + \sigma_\xi^2}$.

□

Again, the variance depends on unknown parameters which must be estimated.

Simulation study (see Appendix B) shows the good performance of the proposed method in both the AUC and the variance estimation.

The standard Student-Welch test (Welch (1947)), commonly used for checking the equality of means equality without assuming equality of variance, is based on the statistics,

$$\mathrm{t_{SW}} = \sqrt{\mathrm{n} \cdot \mathrm{m}} \cdot \frac{\overline{\mathrm{Y_m}} - \overline{\mathrm{X_n}}}{\sqrt{n \cdot \hat{S}^2_{\xi,m} + m \cdot \hat{S}^2_{\chi,n}}}. \tag{9}$$

As it is well-known, under the null ($\mu_\xi = \mu_\chi$), the exact distribution of t_{SW} is approximately a Student T with

$$df = \frac{n \cdot m \cdot (n-1) \cdot (m-1) \cdot \left[m \cdot \hat{S}^2_{\chi,n} + n \cdot \hat{S}^2_{\xi,m}\right]}{m^2 \cdot (m-1) \cdot S^4_{\chi,n} + n^2 \cdot (n-1) \cdot S^4_{\xi,m}}$$

degrees of freedom. Relationship between t_{SW} and $\hat{\mathcal{A}}_{\mathcal{N}}$ is clear:

$$t_{SW} = \sqrt{n \cdot m} \cdot \frac{\overline{Y_m} - \overline{X_n}}{\sqrt{m} \cdot \sqrt{\lambda_n \cdot \hat{S}^2_{\xi,m} + \hat{S}^2_{\chi,n}}} = \sqrt{n} \cdot \frac{\sqrt{\hat{S}^2_{\xi,m} + \hat{S}^2_{\chi,n}}}{\sqrt{\lambda_n \cdot \hat{S}^2_{\xi,m} + \hat{S}^2_{\chi,n}}} \cdot \frac{\overline{Y_m} - \overline{X_n}}{\sqrt{\hat{S}^2_{\xi,m} + \hat{S}^2_{\chi,n}}}$$

$$= \sqrt{n} \cdot \frac{\sqrt{\hat{S}^2_{\xi,m} + \hat{S}^2_{\chi,n}}}{\sqrt{\lambda_n \cdot \hat{S}^2_{\xi,m} + \hat{S}^2_{\chi,n}}} \cdot \Phi^{-1}\left(\hat{\mathcal{A}}_{\mathcal{N}}\right), \tag{10}$$

where $\lambda_n = n/m$. The sole difference is how each estimator deals with unbalance sample sizes. Obviously, the relationship is direct for $\lambda_n = 1$.

5. CATEGORICAL VARIABLES

Categorical data appear frequently in practice. When one wants to investigate the behavior of a categorial variable in two different groups (we can label these groups by g_1 and g_2, respectively), usually, the data are described in a so-called *contingency table* (also referred by a cross-tabulation or crosstab); the result of some independence test is reported and, probably, some among a wide variety of indices is also provided (see, for instance, Agresti (2012)). Suppose that the random variables χ and ξ denote the values of the studied feature on the groups g_1 and g_2, respectively. Suppose that they can take values on the set by $\mathcal{I} = \{I_1, \ldots, I_k\}$ ($k \in \mathbb{N}$ and, usually, small) and let Y_m and X_n be two random samples drawn from χ and ξ, respectively. The samples are determined by the number of individuals in each category; i.e., $a_1, \cdots a_k$ for Y_m and $b_1, \cdots b_k$ for X_n, respectively. Table 1 stands for a usual contingence table; a_i/m and b_i/n estimate $p_{i,P} = \mathcal{P}\{\xi = I_i\}$ $(= \mathcal{P}\{I_i|g_1\})$ and $p_{i,N} = \mathcal{P}\{\chi = I_i\}$ $(= \mathcal{P}\{I_i|g_2\})$ with $1 \leq i \leq k$, respectively.

Table 1
Contingence table: a_i and b_i stand for the number of subjects in the sample Y_m and X_n respectively, which take the value I_i ($1 \leq i \leq k$). $\boldsymbol{a} = \sum_{i=1}^{k} a_i = m$ and $\mathbf{b} = \sum_{i=1}^{k} b_i = n$

	I_1	I_2		I_k	Total
g_1	a_1	a_1	...	a_1	a $(= m)$
g_2	b_1	b_2	...	b_k	b $(= n)$
Total	$a_1 + b_1$	$a_2 + b_2$	...	$a_k + b_k$	$a + b$ $(= m + n)$

The chi-square test is based on the difference between the observed quantities and the expected ones (assuming the null); when we are interested in checking the independence between the variables, chi-square test based on the Table 1 will be,

$$\chi^2_{n,m} = \frac{1}{n \cdot m} \sum_{i=1}^{k} \frac{(a_i \cdot n - b_i \cdot m)^2}{a_i + b_i}, \tag{11}$$

which, as it is well-known, follows a χ^2_{k-1} distribution.

For categorical variables, involved sensitivities and specificities can be directly expressed in a table (similar to Table 1) and, therefore, the ROC curve is not so frequently used. However, section 4.2 of Zhou, Obuchowski and McClish (2002) is devoted to the ROC curve use, estimation and testing in ordinal-scale data. Due to, in this context, $\mathcal{P}\{\chi = \xi\} \neq 0$, the AUC is defined as,

$$\hat{\mathcal{A}}_{\mathrm{k}} = \mathcal{P}\{\xi > \chi\} + 1/2 \cdot \mathcal{P}\{\xi = \chi\}, \tag{12}$$

k stands for the number of categories. Obviously, for continuous random variables, this definition and (2) are equivalent. Figure 2 represents the AUC for a binary variable (left) and for an ordinal variable with three categories (right).

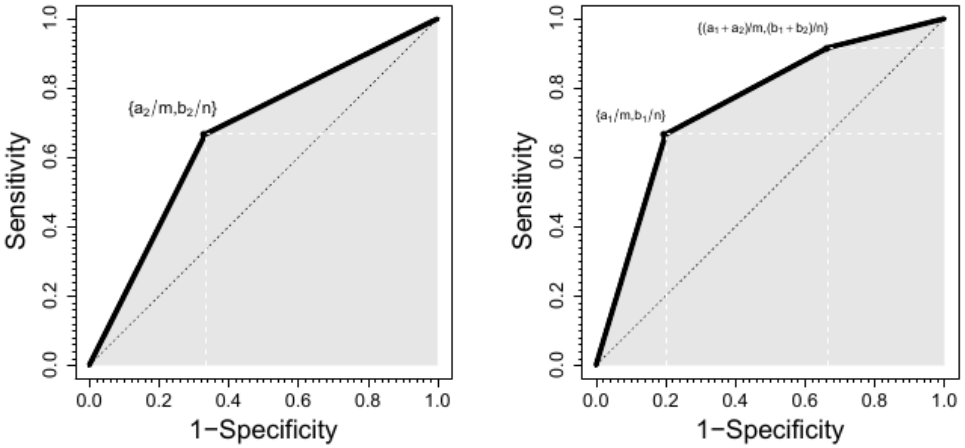

Figure 2: At Left, Densities for the Two İnvolved Populations, Both Follow Normal Laws. At Right, Resulting ROC Curve

From the trapezoidal rule, it is easy to check that Table 1 leads to an AUC value of

$$\hat{\mathcal{A}}_{\mathrm{k}} = \frac{1}{2mn}\sum_{i=1}^{k-1}\left(2\sum_{j=0}^{i} b_j + b_{i+1}\right)\cdot a_{i+1}, \tag{13}$$

with $\mathrm{a}_0 = 0 = \mathrm{b}_0$. The projections method can be directly used in order to obtain the asymptotic normality of the above expression (see, for instance, van der Vaart (1998)).

The case where *k=2* is probably the most relevant, especially in biomedicine, where indices such the relative risk ($\mathrm{RR} = p_{1,P}\cdot\left(m\cdot p_{2,P} + n\cdot p_{2,N}\right)/p_{2,P}\cdot\left(m\cdot p_{1,P} + n\cdot p_{1,N}\right)$ or the odd ratio ($\mathrm{OR} = p_{1,N}\cdot p_{2,P}/p_{1,P}\cdot p_{2,N}$) are commonly reported in the specialized literature (notice that they are not appropriated when any of the involved probabilities takes the value 0). In this case, the real AUC will be (see Figure 2),

$$\begin{aligned}\hat{\mathcal{A}}_2 &= 1/2\cdot p_{2,N}\cdot p_{2,P} + p_{1,N}\cdot p_{2,P} + 1/2\cdot p_{1,N}\cdot p_{1,P}\\ &= 1/2\cdot\left(1-p_{1,N}\right)\cdot p_{2,P} + p_{1,N}\cdot p_{2,P} + 1/2\cdot p_{1,N}\cdot\left(1-p_{2,P}\right)\\ &= 1/2\cdot\left(p_{1,N} + p_{2,P}\right).\end{aligned} \tag{14}$$

Therefore, $\hat{\mathcal{A}}_2$ is the average of the true-positive and the true-negative rates. Although the relationship is not direct and even same theoretical AUCs can drive to different ORs (RRs), all indices are strongly associated; in fact, for a fixed problem, correlations between the AUC and log(OR) estimations are always close to 1 (data not shown). This association (see Table 2 in the Appendix B) endorses the use of the AUC as effect size measure. Of course, depending of the case, it can be more, or less, appropriated than other indices. This situation is the usual one and it is the main cause of the number of existing indices.

6. REAL-WORLD APPLICATION

Santos-Juanes et al. (2015) published a study in which psoriasis (Ps) is related with multiple measurements of dyslipidemia (DL) which include levels of triglycerides (TGs), low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol and total cholesterol (TCh). In that work, in order to determine the association between DL and Ps, variables were categorized and then, the odd ratios (OR) were computed. Thresholds were selected to follow the most recent guidelines recommendations. Of course, these cut-off points can be discussed, and they are often moved. The use of AUCs as effect size measurement avoids the need of categorizing the variables. For instance, Figure 3 depicts the observed effects by population subgroups for TCh and TGs. Those effects were measured by using the AUC. Both parametric and non-parametric approaches were considered; however, in this case (661 cases vs. 661 controls) differences between both methods were almost negligible.

The conclusions derived by using the AUC as effect size are consistent with those using ORs from arbitrary categorized variables and reported in the original paper. The use of AUC avoids categorizing the data and the loss of information derived from. In addition, both the non-parametric and the parametric estimations were similar, i.e., the estimated effects do not depend on the chosen procedure.

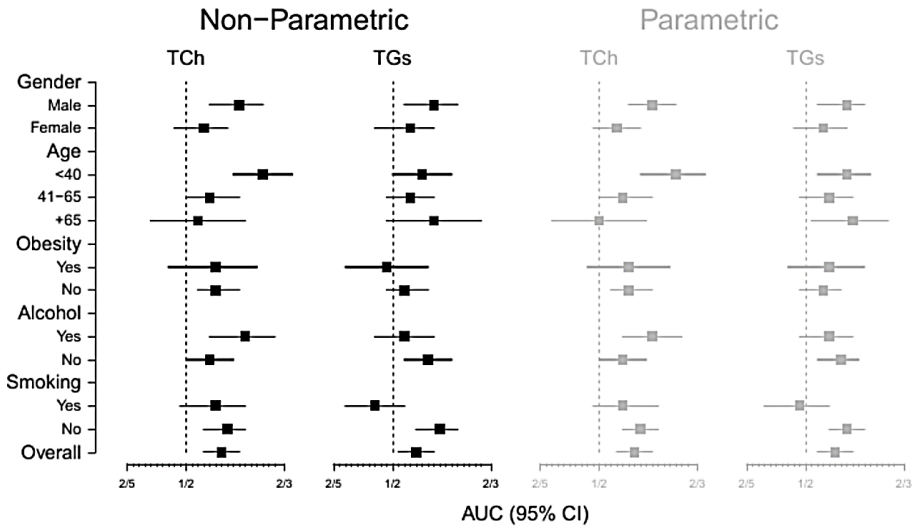

Figure 3: Forest Plot for the Non-Parametric and Parametric AUC for Total Cholesterol (TCh) and Triglycerides (TGs) Levels in Psoriasis vs. Controls by subgroups

7. MAIN CONCLUSIONS

Due to *p*-values depend on the sample size, reporting indices which measure the size of the observed effect is strongly recommended, in those studies where the sample sizes are mandatorily small, for instance, in the study of infrequent diseases (see, for instance, Kraemer and Kupfer (2005) or McCough and Faraone (2009), for two dissertations about the need of going beyond the *p*-value in biomedical research) but also in those studies involving the so called *big data* in which small differences usually achieved small *p*-values. Recently, Demidenko (2016), in a paper with the suggestive title of *the p-value you can't buy*, introduced the *D*-value as an alternative *p*-values. This *D*-value is strongly related with the AUC in the parametric case.

The use of effect size measures is a standard practice when the studied variable is categorical and, for instance, ORs or RRs are frequently reported in biomedical literature. However, categorization the variables, is the conventional procedure when these are continuous. In this manuscript, the authors propose to use the AUC as effect size measure.

The use of the area under the ROC curve as effect size measure in two-sample problems has a number of advantages: *i)* it is a well-known index which does not depend on the used units and with a clear probability interpretation (shared with the *p*-value); *ii)* it has direct relationship with traditional statistics/indices in which the standard inference is based on, in both the non-parametric and the parametric cases, even for categorical variables; *iii)* due to the AUC has been widely studied from both theoretical and practical approaches, there exists a number of generalizations in which it can be applied, for instance, in time-dependent problems (see, for instance, Heagerty, Lumley and Pepe (2000) and references therein) or in more general distribution comparisons (Martinez-Camblor, Corral, Rey et. al. (2014); *iv)* in 2×2 tables, the AUC-values below or above 1/2 have a linear interpretation, notice that there is not a linear scale in the RR or OR interpretation.

The use of AUC on a real-world example concludes with the same conclusion that the original work in which the variables were categorized. Besides, obtained results were similar for both parametric and non-parametric estimations. Authors consider that all these evidences endorse the use of AUC as effect measurement and, although some technical aspect must be carefully considered, the use on paired sample should be also studied.

APPENDIX A:
THEOREMS PROOF'S

In this section we provided demonstrations for Theorem 2 and Theorem 3. Theorem 2 was already enunciated in Hsieh and Turnbull (1996), here we include the computes in detail.

Theorem 2's Proof:

If $n/m = \lambda_n \rightarrow_n \lambda$, we have that

$$\begin{aligned}\sqrt{n}\cdot\{\hat{\mathcal{A}}-\mathcal{A}\} &= \sqrt{n}\cdot\left\{\int \hat{F}_m(Y_m,t)d\hat{F}_n(X_n,t) - \int F_\xi(t)dF_\chi(t)\right\}\\ &= \sqrt{\lambda_n}\cdot\sqrt{m}\cdot\left\{\int \hat{F}_m(Y_m,t)d\hat{F}_n(X_n,t) - \int F_\xi(t)d\hat{F}_n(X_n,t)\right\}\\ &+ \sqrt{n}\cdot\left\{\int F_\xi(t)d\hat{F}_n(X_n,t) - \int F_\xi(t)dF_\chi(t)\right\}\\ &= \sqrt{\lambda_n}\cdot\left\{\int \sqrt{m}\cdot\left[\hat{F}_m(Y_m,t)-F_\xi(t)\right]d\hat{F}_n(X_n,t)\right\}\\ &+ \int F_\xi(t)d\sqrt{n}\cdot\left[\hat{F}_n(X_n,t)-F_\chi(t)\right].\end{aligned}$$

The Hungarian embeddings (see, for instance, van der Vaart (1998)) guarantees that there exists a probability space on which one can define a sequence of a Brownian bridge version, $\mathcal{B}^m$, such that

$$\sqrt{m}\cdot\left[\hat{F}_m(Y_m,t)-F_\xi(t)\right] = \mathcal{B}^m\{F_\xi(t)\} + O\left(m^{-1/2}\cdot\log(m)\right) \quad \text{a.s.}$$

then,

$$\sqrt{n}\cdot\{\hat{\mathcal{A}}-\mathcal{A}\} = \sqrt{\lambda}\cdot\int \mathcal{B}_1^m\{F_\xi\}dF_\chi(t) - \int \mathcal{B}_2^n\{F_\chi\}dF_\xi(t) + O\left(n^{-1/2}\cdot lo\,g(n)\right). \tag{15}$$

Brownian bridge properties and the independence of $\mathcal{B}_1^m$ and $\mathcal{B}_2^n$ guarantee the asymptotic normality and an expected value of zero. On the other hand, if $\mathbb{V}[\cdot]$ stands for the variance operator,

$$\begin{aligned}\mathbb{V}\left[\sqrt{n}\cdot\{\hat{\mathcal{A}}-\mathcal{A}\}\right] &= \lambda\cdot\mathbb{V}\left[\int \mathcal{B}_1^m\{F_\xi(t)\}dF_\chi(t)\right] + \mathbb{V}\left[\int \mathcal{B}_2^n\{F_\chi(t)\}dF_\xi(t)\right]\\ &+ O\left(n^{-1/2}\cdot\log(n)\right)\\ &= \lambda\cdot\mathbb{E}\left[\left(\int \mathcal{B}_1^m\{F_\xi(t)\}dF_\chi(t)\right)^2\right]\\ &+ \mathbb{E}\left[\left(\int \mathcal{B}_2^n\{F_\chi(t)\}dF_\xi(t)\right)^2\right] + O\left(n^{-1/2}\cdot\log(n)\right)\\ &= \lambda\cdot\mathbb{E}\left[{S_1}^2\right] + \mathbb{E}\left[{S_2}^2\right] + O\left(n^{-1/2}\cdot\log(n)\right).\end{aligned} \tag{16}$$

Now, from the Brownian bridge properties, we derive

$$\begin{aligned}\mathbb{E}\left[{S_1}^2\right] &= \iint \mathbb{E}\left[\mathcal{B}_1^m\{F_\xi(t)\}\cdot\mathcal{B}_1^m\{F_\xi(t)\}\right]dF_\chi(t)dF_\chi(s)\\ &= \iint\left[F_\xi(t)\}\wedge F_\xi(t)\right]dF_\chi(t)dF_\chi(s) - \left(\int F_\xi(t)dF_\chi(t)\right)^2\\ &= \int\left[\int_{-\infty}^{s} F_\xi(t)dF_\chi(t) + \int_s^{\infty} F_\xi(s)dF_\chi(t)\right]dF_\chi(s)\\ &- \left(\int F_\xi(t)dF_\chi(t)\right)^2 = \int F_\chi^2(t)dF_\xi(t) - \left(\int F_\chi(t)dF_\xi(t)\right)^2\\ &= \left|\left|F_\chi\left(F_\xi^{-1}\right)\right|\right|. \end{aligned} \quad (17)$$

Arguing in the same way, $\mathbb{E}\left[{S_2}^2\right] = \left|\left|F_\xi\left(F_\chi^{-1}\right)\right|\right|$; then, the asymptotic expression for the variance is directly derived from (2) and (3). The proof concludes by taking into account (1), (2) and (3).

□

Theorem 3's Proof:

Directly, we have the equality

$$\begin{aligned}\sqrt{n}\cdot\{\hat{\mathcal{A}}_{\mathcal{N}} - \mathcal{A}\} &= \sqrt{\mathrm{n}}\cdot\left\{\Phi\left(\frac{\overline{\mathrm{Y_m}} - \overline{\mathrm{X_n}}}{\sqrt{\hat{S}_{\xi,m}^2 + \hat{S}_{\chi,n}^2}}\right) - \Phi\left(\frac{\mu_\xi - \mu_\chi}{\sqrt{\sigma_\xi^2 + \sigma_\chi^2}}\right)\right\}\\ &= \sqrt{n}\cdot\left\{\Phi\left(\frac{\overline{Y_m} - \overline{X_n}}{\sqrt{\hat{S}_{\xi,m}^2 + \hat{S}_{\chi,n}^2}}\right) - \Phi\left(\frac{\overline{Y_m} - \overline{X_n}}{\sqrt{\sigma_\xi^2 + \sigma_\chi^2}}\right)\right\}\\ &+\sqrt{n}\cdot\left\{\Phi\left(\frac{\overline{Y_m} - \overline{X_n}}{\sqrt{\sigma_\xi^2 + \sigma_\chi^2}}\right) - \Phi\left(\frac{\mu_\xi - \mu_\chi}{\sqrt{\sigma_\xi^2 + \sigma_\chi^2}}\right)\right\} = \sqrt{n}\cdot A + \sqrt{n}\cdot B. \end{aligned} \quad (18)$$

Applying one-term Taylor expansion on the first summand, it holds:

$$\begin{aligned}\sqrt{n}\cdot A &= \sqrt{n}\cdot\varphi(q_n)\left\{\frac{\overline{Y_m} - \overline{X_n}}{\sqrt{\hat{S}_{\xi,m}^2 + \hat{S}_{\chi,n}^2}} - \frac{\overline{Y_m} - \overline{X_n}}{\sqrt{\sigma_\xi^2 + \sigma_\chi^2}}\right\}\\ &= \varphi(q_n)\cdot\left\{\frac{\sqrt{n}\cdot(\overline{Y_m} - \overline{X_n})\cdot\left[\sigma_\xi^2 - \hat{\mathrm{S}}_{\xi,\mathrm{m}}^2 + \sigma_\chi^2 - \hat{\mathrm{S}}_{\chi,\mathrm{n}}^2\right]}{\sqrt{\left(\sigma_\xi^2 + \sigma_\chi^2\right)\cdot\left(\hat{S}_{\xi,m}^2 + \hat{S}_{\chi,n}^2\right)}\cdot\left[\sqrt{\sigma_\xi^2 + \sigma_\chi^2} + \sqrt{\hat{S}_{\xi,m}^2 + \hat{S}_{\chi,n}^2}\right]}\right\} + O_P\left(n^{-1/2}\right) \end{aligned} \quad (19)$$

where q_n is a point within the interval determined by the points $(\overline{Y_m} - \overline{X_n}) / \sqrt{\hat{S}^2_{\xi,m} + \hat{S}^2_{\chi,n}}$ and $(\overline{Y_m} - \overline{X_n}) / \sqrt{\sigma^2_\xi + \sigma^2_\chi}$. Arguing similarly for the second summand, it holds:

$$\sqrt{n} \cdot B = \sqrt{n} \cdot \varphi(p_n) \left\{ \frac{\overline{Y_m} - \overline{X_n}}{\sqrt{\sigma^2_\xi + \sigma^2_\chi}} - \frac{\mu_\xi - \mu_\chi}{\sqrt{\sigma^2_\xi + \sigma^2_\chi}} \right\}$$

$$= \frac{\varphi(p_n)}{\sqrt{\sigma^2_\xi + \sigma^2_\chi}} \cdot \left\{ \sqrt{\lambda_n} \cdot \sqrt{m} \cdot \{\overline{Y_m} - \mu_\xi\} + \sqrt{n} \cdot \{\mu_\chi - \overline{X_n}\} \right\}, \quad (20)$$

where p_n is within the interval determined by the points $a_n = (\mu_\xi - \mu_\chi) / \sqrt{\sigma^2_\xi + \sigma^2_\chi}$ and $b_n = \overline{(Y_m} - \overline{X_n}) / \sqrt{\sigma^2_\xi + \sigma^2_\chi}$. Due to $\lambda_n \to \lambda$, $\varphi(\cdot)$ is a continuous function, and that both χ and ξ are normal distributed, it is had the convergence

$$\sqrt{n} \cdot B \longrightarrow_{\mathrm{n}} \mathcal{N}\left(0, \sqrt{\frac{\lambda \cdot \sigma^2_\xi + \sigma^2_\chi}{\sigma^2_\xi + \sigma^2_\chi}}\right) \quad (21)$$

Since $|a_n - b_n| \to_n 0$ (a.s) and $p_n \to_n p$ (a.s). Theorem 3 is directly derived from (4), (5) and (7).

□

APPENDIX B:
SIMULATION RESULTS

Table 2 shows the results obtained in a Monte Carlo simulation study. Particularly, we report mean ± standard deviation for the observed errors in both the AUC and the variance estimations (given in Section 4) in 10,000 Monte Carlo iterations. Samples were drawn from normal distributions.

Table 2
Mean ± standard deviation for the observed errors in both the AUC and the variance estimations (given in section 4) in 10,000 Monte Carlo iterations. Samples were drawn from normal distributions

$\mathcal{A}$	n	λ	$\sigma_\xi^2 = 1$		$\sigma_\xi^2 = 4$	
			$\lvert\widehat{\mathcal{A}_N} - \mathcal{A}\rvert$	$\sqrt{n} \cdot \lvert\widehat{\mathbb{S}} - \mathbb{S}\rvert$	$\lvert\widehat{\mathcal{A}_N} - \mathcal{A}\rvert$	$\sqrt{n} \cdot \lvert\widehat{\mathbb{S}} - \mathbb{S}\rvert$
0.6	50	1/2	0.038 ± 0.03	0.011 ± 0.01	0.034 ± 0.03	0.008 ± 0.01
		1	0.044 ± 0.03	0.009 ± 0.01	0.044 ± 0.03	0.009 ± 0.01
		2	0.054 ± 0.04	0.024 ± 0.02	0.059 ± 0.04	0.021 ± 0.02
	100	1/2	0.026 ± 0.02	0.008 ± 0.01	0.024 ± 0.02	0.006 ± 0.01
		1	0.031 ± 0.02	0.006 ± 0.01	0.031 ± 0.02	0.006 ± 0.01
		2	0.038 ± 0.03	0.016 ± 0.01	0.042 ± 0.03	0.015 ± 0.01
0.7	50	1/2	0.035 ± 0.02	0.016 ± 0.01	0.031 ± 0.02	0.012 ± 0.01
		1	0.041 ± 0.03	0.018 ± 0.01	0.040 ± 0.03	0.018 ± 0.01
		2	0.050 ± 0.04	0.033 ± 0.03	0.056 ± 0.04	0.038 ± 0.03
	100	1/2	0.025 ± 0.02	0.012 ± 0.01	0.022 ± 0.02	0.008 ± 0.01
		1	0.029 ± 0.02	0.013 ± 0.01	0.029 ± 0.02	0.013 ± 0.01
		2	0.035 ± 0.03	0.023 ± 0.02	0.040 ± 0.03	0.026 ± 0.02
0.8	50	1/2	0.029 ± 0.02	0.022 ± 0.02	0.027 ± 0.02	0.017 ± 0.01
		1	0.034 ± 0.03	0.027 ± 0.02	0.035 ± 0.03	0.028 ± 0.02
		2	0.041 ± 0.03	0.046 ± 0.04	0.047 ± 0.03	0.055 ± 0.04
	100	1/2	0.021 ± 0.02	0.016 ± 0.01	0.019 ± 0.01	0.012 ± 0.01
		1	0.024 ± 0.02	0.020 ± 0.02	0.025 ± 0.02	0.020 ± 0.02
		2	0.030 ± 0.02	0.032 ± 0.03	0.033 ± 0.02	0.038 ± 0.03
0.9	50	1/2	0.020 ± 0.02	0.029 ± 0.02	0.019 ± 0.01	0.022 ± 0.02
		1	0.023 ± 0.02	0.036 ± 0.03	0.024 ± 0.02	0.038 ± 0.03
		2	0.029 ± 0.02	0.058 ± 0.04	0.033 ± 0.03	0.074 ± 0.05
	100	1/2	0.014 ± 0.01	0.020 ± 0.02	0.013 ± 0.01	0.016 ± 0.01
		1	0.017 ± 0.01	0.025 ± 0.02	0.017 ± 0.01	0.027 ± 0.02
		2	0.020 ± 0.02	0.041 ± 0.03	0.024 ± 0.02	0.052 ± 0.04

Finally, Table 3 shows the values for the AUCs, ORs, RRs and χ^2 for different theoretical situations. Considered sample sizes, relevant for RRs and χ^2 were n=50=m.

Table 3
Values for the AUCs, ORs, RRs and χ^2 for different theoretical situations. Considered sample size (relevant for RRs and χ^2 were n = 50 = m).

$\mathcal{A}_2$	$p_{1,N}$	OR	RR	$n \cdot \chi_1^2$
0.6	0.3	1.890	2.250	0.125
	0.5	2.333	1.555	0.083
	0.7	1.500	1.500	0.083
	0.9	3.857	1.933	0.122
0.7	0.5	9.000	3.857	0.381
	0.7	5.444	2.333	0.320
	0.9	9.000	2.333	0.381
0.8	0.7	21.000	6.000	0.750
	0.9	21.000	6.000	0.750
0.9	0.9	81.000	9.000	1.280

REFERENCES

1. Agresti, A. (2012). *Categorical data analysis*, Third Edition, Wiley, New Jersey.
2. Cleves, M.A, (2002). Comparative assessment of three common algorithms for estimating the variance of the area under the nonparametric receiver operating characteristic curve. *Stata Journal*, 2(3), 280-289.
3. Csörgő, M. (1983). *Quantile processes with statistical application*. SIAM, Philadelphia.
4. Csörgő, M. and Révész, P. (1981). *Strong approximations in probability and statistics*, Academic Press, New York.
5. Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, 21(20), 3093-3106.
6. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.
7. Green, D.M. and Swets, J.A. (1966). *Signal detection theory and psychophysics*, Wiley, New York.
8. Heagerty, P., Lumley, T. and Pepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56, 337-344.
9. Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making*, 11(2), 95-101.

10. Hsieh, F. and Turnbull, B. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, 24(1), 25-40.
11. Ma, H., Bandos, A.I., Rockettel H.E. and Gur, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in Medicine*, 32(20), 3449-3458.
12. Martínez-Camblor, P. (2017). Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical Methods in Medical Research*, 26(1), 5-20.
13. Martínez-Camblor, P. Corral, N., Rey, C., Pascual, J. and Cernuda, E. (2017). Receiver operating characteristic curve generalization for non-monotone relationships. *Statistical Methods in Medical Research*, 26(1), 113-123.
14. Martínez-Camblor, P., Carleos, C. and Corral, N. (2011). Powerful nonparametric statistics to compare *k*-independent ROC curves. *Journal of Applied Statistics*, 38(7), 1317-1332.
15. Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, 19(1), 78-89.
16. Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, Oxford.
17. Santos-Juanes, J., Coto-Segura, P., Fernández-Vega, I., Armesto, S. and Martínez-Camblor, P. (2015). Psoriasis vulgaris with or without arthritis and independent of disease severity or duration is a risk factor for hypercholesterolemia. *Dermatology*, 230, 170-176.
18. Van der Vaart, A.W. (1998). *Asymptotic statistics.* Cambridge University Press, Cambridge.
19. Welch, B.L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1-2), 28-35.
20. Zou, K.H., Hall, W.J. and Shapiro, D.E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16(19), 2143-2156.
21. Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2002). *Statistical methods in diagnostic medicine*, Wiley & Sons, New York.