REVSTAT – Statistical Journal
Volume 17, Number 2, April 2019, 223-244

ACCURACY MEASURES FOR BINARY CLASSIFICATION BASED ON A QUANTITATIVE VARIABLE

Authors: Rui Santos
– School of Technology and Management, Polytechnic Institute of Leiria
CEAUL – Centre of Statistics and Applications, Portugal
rui.santos@ipleiria.pt

Miguel Felgueiras
– School of Technology and Management, Polytechnic Institute of Leiria
CEAUL – Centre of Statistics and Applications
Centre of Applied Research in Management and Economics, Portugal
mfelg@ipleiria.pt

João Paulo Martins
– School of Technology and Management, Polytechnic Institute of Leiria
CEAUL – Centre of Statistics and Applications, Portugal
jpmartins@ipleiria.pt

Liliana Ferreira
– School of Technology and Management, Polytechnic Institute of Leiria
CMAFCIO – Centre for Mathematics, Fundamental Applications and Operations Research, Portugal
liliana.ferreira@ipleiria.pt

Received: October 2018 Revised: January 2019 Accepted: March 2019

Abstract:

- The identification of the right methodology to perform binary classification based on an observed quantitative variable is usually a complex choice. Thus, the use of appropriate accuracy measures is crucial. In fact, the ROC curve reveals a lot of information about the accuracy of the applied methodology for all the possible values of the cut-point. In particular, the integral and partial areas under the ROC curve are widely used. The ϕ index, in which sensitivity equals specificity, may also be applied. Nevertheless, the accuracy at one specific cut-point may be sufficient to assess the accuracy in some applications. Therefore, different ways to define the optimal cut-point may be applied, such as the maximization of the Youden index, the maximization of the concordance probability or the minimization of the distance to the point with absence of misclassification. To compare the adequacy of these measures, a simulation study was performed under different scenarios. The results highlight the advantages and disadvantages of each procedure and advise the use of the ϕ index.

Key-Words:

- *Binary classification; cut-point; ROC curve; sensitivity; specificity; simulation.*

AMS Subject Classification:

- 62P10, 92D30.

1. INTRODUCTION

Assume that an infection with prevalence rate p is affecting a population with N individuals. Let X_i, with $i = 1, \cdots, N$, be N independent Bernoulli trials ($X_i \sim \text{Ber}(p)$) with probability p, where the random variable (r.v.) X_i denotes the presence ($X_i = 1$) or the absence ($X_i = 0$) of the infection in the i-th individual. In addition, let Y_i represents the value of a diagnostic test performed by the i-th individual, characterized by the distribution D_0 with parameter vector θ_0 if $X_i = 0$ and by the distribution D_1 with parameter vector θ_1 if $X_i = 1$, for $i = 1, \ldots, N$. Finally, let t be the cut-point of the binary classification (healthy versus infected) based on the observation of the r.v. Y_i. Under these conditions we can define the following classification rule:

- If $Y_i \leq t \Rightarrow X_i^-$ (a negative result, i.e. the individual is classified as healthy);
- If $Y_i > t \Rightarrow X_i^+$ (a positive result, i.e. the individual is classified as infected).

As a matter of fact, the opposite inequalities can also be applied. Nevertheless, the reasoning is exactly the same and, therefore, we will restrict this presentation to the previously described situation.

The intention is to perform a diagnostic test to achieve a binary classification (e.g. healthy versus infected) based on the observed value of the quantitative variable Y_i. Nonetheless, almost all tests may result in misclassification due the occurrence of false negative or false positive results. Thus, it is essential to assess the performance of the applied binary classification procedure. The most common measure to evaluate the performance is the area under the Receiver Operating Characteristic (ROC) curve (AUC) [32], but it evaluates all possible cut-points, even those that are clinically unsuitable [7]. The partial AUC (pAUC) has been attracting the attention in medical issues [1, 2, 10] as well as in decision making and machine learning applications [16, 17] since it focus on a suitable range of interest for the true positive (or negative) rate [14]. Nevertheless, the partial AUC has some limitations in the application on ROC curves that cross the diagonal line, which are quite frequent in practice. Thus, there are still some contraindications for its widespread, regardless of some new proposals to overcome this problem (e.g., [30, 33]). And, to the best of our knowledge, there is no simulation study that allows to identify the existence of regions in which the computation of pAUC is suitable even in those cases. Moreover, despite its advantages over AUC, the pAUC continues to be unknown to many who apply binary classification procedures based on a quantitative variable. Hence, the main goal of this paper is to compare the usual measures of accuracy in binary classification in order to identify the most appropriate, completing the works already presented in [26, 25].

The main accuracy measures for binary classification based on a quantitative variable are presented in Section 2. In Section 3, a simulation study is

performed in order to compare those measures under different scenarios. All results were computed by the R software using different distributions as well as diverse sample sizes. Finally, the main conclusions are outlined in Section 4.

2. DIAGNOSTIC ACCURACY MEASURES

The usual accuracy measures for classification can be computed for each possible value for the cut-point t, namely the specificity φ_e (or true negative fraction) that corresponds to the probability of obtaining a negative result in a healthy individual, i.e.

$$\mathrm{P}\left(X_i^-|X_i=0\right)=\mathrm{P}\left(Y_i\leq t|X_i=0\right)=F_{\mathrm{D}_0}(t),$$

where F_{D} denotes the distribution function of the distribution D. Similarly, the sensitivity φ_s (true positive fraction) corresponds to the probability of getting a positive result in an infected individual, i.e.

$$\mathrm{P}\left(X_i^+|X_i=1\right)=\mathrm{P}\left(Y_i>t|X_i=1\right)=1-F_{\mathrm{D}_1}(t)=\overline{F}_{\mathrm{D}_1}(t),$$

where $\overline{F}_{\mathrm{D}}$ denotes the survival function of the distribution D.

Note that the probabilities φ_s and φ_e depend on the value t considered for the cut-point and are inversely correlated, since increasing one of them implies decreasing the other when the same classification test is performed. Figure 1 uses the densities of the healthy (distribution D_0) and of the infected (distribution D_1) individuals to emphasize the changes in the sensitivity and specificity when different values for the cut-point are applied. The three graphs show the decreasing sensitivity (shaded area represented on the right of the cut-point) and the increasing specificity (shaded area represented on the left of the cut-point) as the value of the cut-point increases.

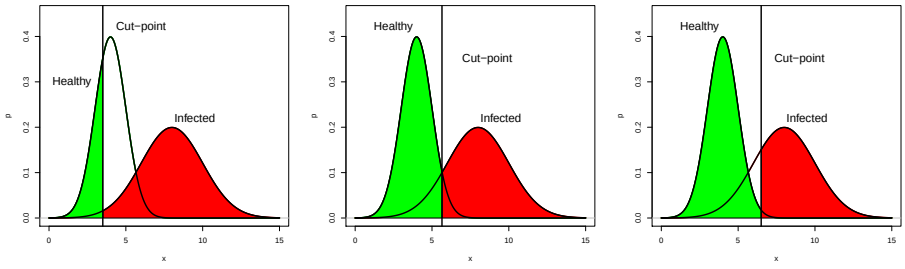

Figure 1: Sensitivity versus specificity on the use of different cut-points.

2.1. The receiver operating characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve allows to visualize the evolution of φ_s and φ_e when the cut-point goes through all possible values, from the point in which all individuals are classified as infected to the other extreme where all individuals are classified as healthy. Therefore, this curve reveals all pairs $(1-\varphi_e, \varphi_s)$ which are also usually denoted by $(x, \mathrm{ROC}(x))$. For this reason, the ROC curve is often used to identify the optimal cut-point of a binary classification methodology, as well as to compare the performance of different methodologies [5, 6, 9, 15, 18, 32, 35]. The first graph of Figure 2 displays an example of a ROC curve.

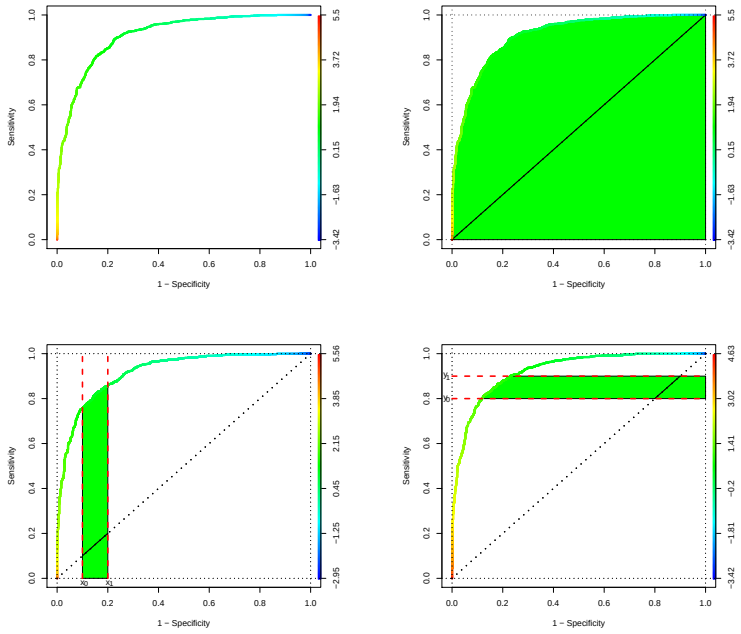

Figure 2: The ROC curve (top left) and the integral (top right) and partial (bottom) areas under the curve.

All ROC curves start in the point $(0,0)$ where all individuals are classified as healthy, and therefore $\varphi_e = 1$ and $\varphi_s = 0$; and finish on the opposite situation, i.e. where all individuals are classified as infected, $\varphi_e = 0$ and $\varphi_s = 1$. The segment $1-\varphi_e = \varphi_s$ connecting these two points represents a random classification without using the information of Y_i, where the probability of classifying any individual as infected is equal to φ_s. Note that the accuracy in any point below this segment would increase if the classification of all individuals were simply changed. The other two vertices of the ROC plane correspond to the remaining extreme cases, the ideal point $(0,1)$ with absence of misclassification $\varphi_e = \varphi_s = 1$; and the point $(1,0)$ in which every individual is misclassified $\varphi_e = \varphi_s = 0$.

2.1.1. The entire area under the ROC curve — AUC

The most widely used measure of accuracy is the area under the ROC curve (AUC). The second graph of Figure 2 shows the integral area under the ROC curve. It represents the mean value of φ_s for all possible values of φ_e. It can also be interpreted as the probability of correctly classifying a pair when the r.v. Y_i is continuous, where 0.5 means unreliability, as in a random classification, and 1 corresponds to the perfect classification (absence of misclassification). The value of the area is also related to the Wilcoxon-Mann-Whitney statistic, allowing to make inference about the ROC curve [18, 35]. The AUC is possibly the most commonly used measure to assess the diagnostic accuracy of a binary classification methodology [6, 18, 32]. However, this measure takes into account all possible values for the cut-point, even those that are unsuitable in practice because it generates very low specificity or sensitivity levels. This is the main drawback of this measure, although it summarizes the entire ROC curve it includes values which are not clinically relevant. In fact, these values should be neglected, otherwise they may interfere in the choice of the best methodology. Moreover, usually only a specific cut-point is applied.

2.1.2. The standardized partial area under the ROC curve — spAUC

The partial area under the ROC curve (pAUC) can be used to evaluate the performance at the interest cut-point values, for which the methodology performs satisfactorily [3, 8, 14, 13, 31, 35]. These values usually correspond to high specificity values, but can also be applied to high sensitivity values. The pAUC over the high specificity range $[1 - x_1, 1 - x_0]$ can be defined as

$$\mathrm{pAUC}(x_0, x_1) = \int_{x_0}^{x_1} \mathrm{ROC}(x)\ \mathrm{d}x,$$

which corresponds to the area of the shaded region in the bottom left chart of Figure 2. It analyses the φ_s when we fix the φ_e in a range of interest. However, in some applications the goal is to evaluate the φ_e when the φ_s is significant. In this cases the area is on the right of the ROC curve (see bottom right graph of Figure 2). Thus, we can compute the pAUC over the high sensitive range $[y_0, y_1]$ using

$$\mathrm{pAUC}(y_0, y_1) = \int_{y_0}^{y_1} 1 - \mathrm{ROC}^{-1}(y)\ \mathrm{d}y,$$

where ROC^{-1} denotes the generalized inverse function of the function ROC. The $\mathrm{pAUC}(y_0, y_1)$ corresponds to the area of the shaded region in the fourth chart of Figure 2. This latter case does not correspond properly to the area below the curve and perhaps the most appropriate designation would be the area on the right of the curve instead of the area under the curve.

In both cases, the pAUC verifies $\text{pAUC}(0,1) = \text{AUC}$ and

$$\frac{1}{2}\left(x_1^2 - x_0^2\right) \leq \text{pAUC}(x_0, x_1) \leq x_1 - x_0, \quad 0 \leq x_0 \leq x_1 \leq 1.$$

Hence, in order to be interpreted analogously to AUC, pAUC can be standardized by

$$\text{spAUC}(x_0, x_1) = \frac{1}{2}\left(1 + \frac{\text{pAUC}(x_0, x_1) - \frac{1}{2}\left(x_1^2 - x_0^2\right)}{x_1 - x_0 - \frac{1}{2}\left(x_1^2 - x_0^2\right)}\right).$$

Thus, spAUC varies between 0.5 (random classification) and 1 (absence of misclassification). Nevertheless, the use of pAUC or spAUC requires the definition of the range of interest $[x_0, x_1]$ or $[y_0, y_1]$. Usually $[x_0, x_1]$ corresponds to $[0, x_1]$, i.e. the highest values for specificity, and spAUC can be seen as (approximately) the average of φ_s when specificity ranges in $[1 - x_1, 1]$. Similarly, $[y_0, y_1]$ commonly corresponds to the highest values for the sensitivity, i.e. $[y_0, 1]$, and spAUC can be seen as (approximately) the average of φ_e when sensitivity ranges in $[y_0, 1]$.

Note that the use of spAUC introduces some difficulties to solve issues related to the arbitrariness of choosing the range of interest. Furthermore, some authors highlight the loss of information, claiming a loss of statistical precision as compared with inferences based on the entire AUC [7, 35].

2.2. The ϕ index

The use of the probability ϕ, which verifies $\varphi_s = \varphi_e = \phi$ for some cut-point, to measure the performance of diagnostic tests in the context of compound tests is advised in [23, 24]. In fact, it corresponds to the intersection of the ROC curve with the straight line $\varphi_s = \varphi_e$, as Figure 3 illustrates. If this value does not exist, as in the use of count distributions or small samples, the distance between φ_s and φ_e shall be minimized and $\phi = \frac{\varphi_s + \varphi_e}{2}$.

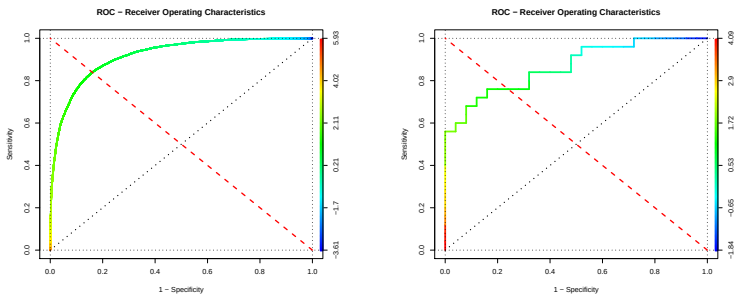

Figure 3: The ϕ index.

In the simulations performed in Section 3, computations of spAUC over the range $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$ for both specificity and sensitivity are used.

The idea was to consider not only significant values for both measures but also to use a small range (the largest range is equal to 0.1).

2.3. The optimal cut-point

In practical issues only a single cut-point is usually applied. Thus, the knowledge of the accuracy at this specific point may be sufficient to assess the classification methodology. Therefore, the selection of the optimal cut-point is a complex decision that depends on several factors. For example, the severity of the infection and the risk of not diagnosing the infection may clearly encourage the choice of a high sensitivity and somehow neglect the specificity level. In the opposite direction, the side effects of the treatment and the treatment cost may favour the use of a high specificity and disregard the sensitivity level. Hence, it may be important to decide between sensitivity or specificity in the selection of the cut-point, because its determination implies a compromise between these two measures. Nonetheless, in the absence of clinical factors that lead to the choice of one of these measures over the other, some criterion of optimization can be applied to choose the optimal cut-point. In fact, there are several available methodologies in the literature to obtain the optimal cut-point value [4, 11, 19, 22, 29, 34, 36], such as the maximization of the Youden index, the minimization of the distance to the point with absence of misclassification and the maximization of the concordance probability.

2.3.1. The Youden index — YI

One way of determining the cut-point is to choose the point that maximizes the Youden index (YI) defined by [4, 11, 20, 27, 34]

$$\mathrm{YI} = \varphi_e + \varphi_s - 1 = F_{\mathrm{D}_0}(t) - F_{\mathrm{D}_1}(t).$$

Geometrically, it corresponds to the point on the ROC curve in which the vertical distance is greater from the line $1 - \varphi_e = \varphi_s$, i.e. the difference between φ_s and $1 - \varphi_e$, as the first chart of Figure 4 shows. It also corresponds to the point t which maximizes the sum $\varphi_e + \varphi_s$ and, thus, maximizes the distance between $F_{\mathrm{D}_0}(t)$ (true positive rate) and $F_{\mathrm{D}_1}(t)$ (false positive rate).

2.3.2. The closest-to-$(0,1)$ criteria — DI

As previously stated, the point $(0, 1)$ corresponds to the perfect classification procedure where all individuals are well classified. Therefore, we intend to

be as close as possible to this situation. Hence, the minimization of the Euclidean distance to the ideal point $(0,1)$, with $\varphi_e = \varphi_s = 1$, is another criteria to choose the best cut-point [11, 19, 27, 29], i.e. minimizing

$$\mathrm{D} = \sqrt{(1-\varphi_e)^2 + (1-\varphi_s)^2} = \sqrt{\overline{F}^2_{\mathrm{D}_0}(t) + F^2_{\mathrm{D}_1}(t)}.$$

The second chart of Figure 4 illustrates this procedure. However, in order to compare with the other measures, in the simulations performed in Section 3 it will be used the maximization of

$$\mathrm{DI} = 1 - \mathrm{D} = 1 - \sqrt{\overline{F}^2_{\mathrm{D}_0}(t) + F^2_{\mathrm{D}_1}(t)},$$

which corresponds to the minimization of D and provides the point on the ROC curve that is the closest to the ideal case $(0,1)$. With this transformation all the measures to select the cut-point take values in the range $[0,1]$ and increase with the improvement of the accuracy of classification.

2.3.3. The concordance probability method — CP

When the r.v. Y_i is continuous, the AUC can be interpreted as the concordance probability. But, when Y_i is not continuous (discrete or ordinal) [11, 12] advocate the use of the concordance probability for a quantitative variable given by the product of sensitivity and specificity, i.e.

$$\mathrm{CP} = \varphi_e\, \varphi_s = F_{\mathrm{D}_0}(t)\, \overline{F}_{\mathrm{D}_1}(t).$$

The maximization of the CP can be used to define the cut-point. The third chart of Figure 4 shows the area of the rectangle which corresponds to the CP value. Thus, covering all the points on the ROC curve as the upper left vertex of the rectangle, we intend to determine the rectangle with maximum area.

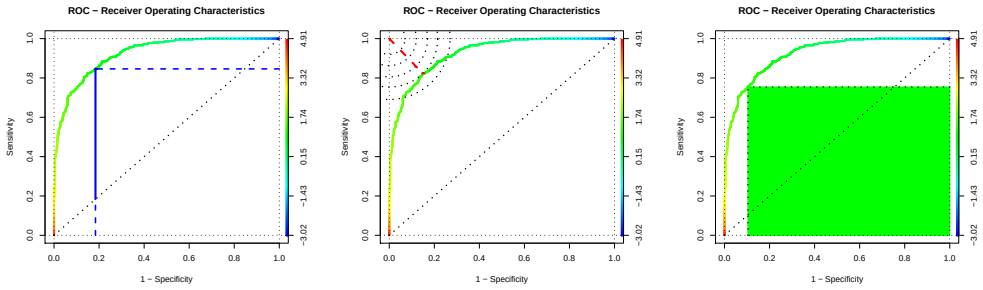

Figure 4: The optimal cut-point using the Youden index (left), the distance to the ideal point (center) and the concordance probability (right).

3. AN ACCURACY COMPARISON BY SIMULATION

In most cases, the focus in the application of a diagnostic test is the evaluation of the accuracy for a single cut-point, which shall be the best one for our purposes. Thus, it is indeed critical to compare the differences between the area AUC and the partial area spAUC under the ROC curve as well as the index ϕ, and to realize if a greater value in these accuracy measures is sufficient to ensure a good accuracy in the selected cut-point, considering the cut-points obtained by the application of the three procedures provided in Subsection 2.3.

Hence, a simulation study was performed through the R software using the `ROCR` and `pROC` packages [21, 28]. All scenarios were analysed using 10^3 replicas and the following accuracy measures were computed:

- AUC – the entire area under the ROC curve;
- SP_{90}, SP_{75}, SP_{50} – spAUC computed over the specificity range $[0.9, 1]$, $[0.75, 1]$ and $[0.5, 1]$, respectively;
- SE_{90}, SE_{75}, SE_{50} – spAUC computed over the sensitivity range $[0.9, 1]$, $[0.75, 1]$ and $[0.5, 1]$, respectively;
- ϕ (or Phi) – the ϕ index;
- SP_{ϕ} (or SPPhi) – spAUC computed over the specificity range $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$;
- SE_{ϕ} (or SEPhi) – spAUC computed over the sensitivity range $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$;
- YI – the maximum Youden index;
- DI – the maximum of 1-D where D denotes the distance to the ideal point $(0, 1)$;
- CP – the maximum of the concordance probability.

In order to compare the obtained results in these measures, the Spearman's rank correlation coefficients were computed to assess monotonic relationships between them. Therefore, these correlations evaluate if the rank of the accuracy in each model is made in the same way using different measures. Note that all those measures vary in $[0, 1]$ and increase with the improvement of the accuracy.

For the test design, diverse sample sizes were applied using equal number of infected and healthy individuals, i.e. $n_0 = n_1 = n \in \{50, 100, 250, 500, 1000\}$. The restriction $n_0 = n_1$ only aims to achieve the same accuracy in the estimation of the sensitivity (only infected individuals are analysed) and specificity (only

healthy individuals are used). Besides, different distributions for the characterization of the infected and healthy individuals were considered in the simulations, both discrete and continuous. In order to simplify the presentation of the obtained results, we will restrict to the cases where the two subpopulations have the same distribution $D_0 = D_1$ but with different values for the parameter vectors, i.e. $\theta_0 \neq \theta_1$. This restriction aims to simplify the interpretation of the results. Moreover, to minimize and simplify the discussion of the main conclusions, only the results obtained with some of the most applied distributions will be shown since they include the most usual shapes of ROC curves. In particular, the following distributions were used:

- Normal, with $\mu_0 = 0$, $\sigma_0 = 1$, $\mu_1 = 2$ and $\sigma_1 \in \{2/3, 1, 1.5, 2, 3\}$;
- Gamma, with $\alpha_0 = 2$, $\beta_0 = 1$, $\alpha_1 \in \{6, 9, 12\}$ and $\beta_1 \in \{1, 3\}$;
- Binomial, with $p_0 = 0.25$ and $p_1 \in \{0.3, 0.4, 0.5\}$;
- Geometric, with $p_0 = 0.2$ and $p_1 \in \{0.1, 0.02\}$.

The main goal is to evaluate the association between those accuracy measures and, therefore, to assess whether those measures are able to evaluate the same criterion of accuracy.

3.1. The sample dimension

Let us consider that the r.v. Y_i has Normal distribution with standard deviation $\sigma = 1$ and mean $\mu_0 = 0$ in a healthy individual and $\mu_1 = 2$ in an infected individual. Figure 5 contains the boxplot for different sample sizes $n \in \{50, 100, 250, 500, 1000\}$ of the applied diagnostic accuracy measures which do not depend on the cut-off value. As expected, the median seems to be always the same, but the range of variation and the interquartile range decrease with the increasing of the sample size. Besides, due to the symmetry of the ROC curves around the line $\varphi_e = \varphi_s$, the partial areas over the specificity have the same behaviour as the partial areas over the sensitivity, converging to the AUC when the range of interest increases. In the last chart some ROC curves obtained with different sample sizes are plotted to illustrate that the ROC curve becomes smoother as n increases.

Table 1 provides the Spearman's rank correlation coefficients between all the computed measures when the sample dimension is $n = 1000$ (upper triangular matrix) and when the sample dimension is $n = 50$ (lower triangular matrix). The results do not seem to have significant differences between the values obtained with $n = 50$ and $n = 1000$. The correlation between the partial areas and the entire area seems to increase when the interval of interest increases and converges

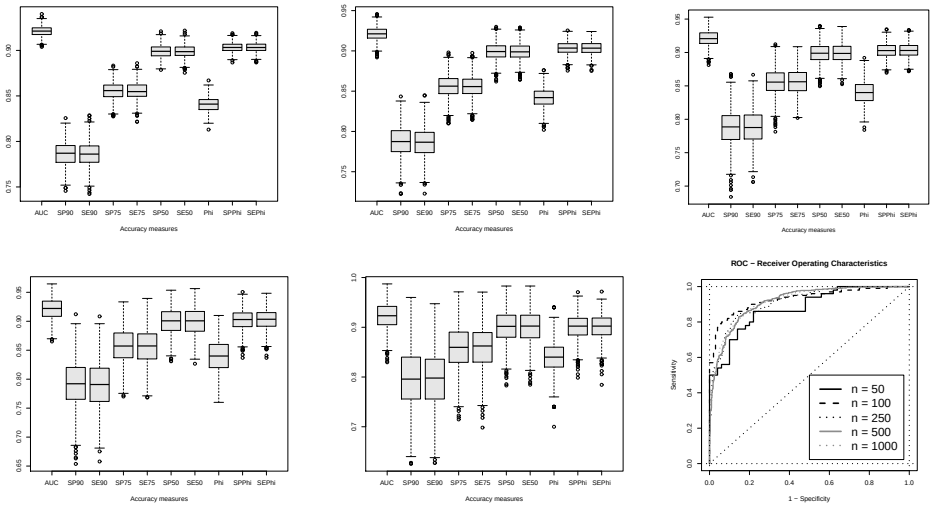

Figure 5: $\mathrm{D}_0(\theta_0) = \mathrm{N}(0,1)$ versus $\mathrm{D}_1(\theta_1) = \mathrm{N}(2,1)$ with different sample dimensions: $n = 1000$ (top left), $n = 500$ (top middle), $n = 250$ (top right), $n = 100$ (bottom left), $n = 50$ (bottom middle), and ROC curves (bottom right).

to all of the support $[0, 1]$ as expected. The partial areas SP_ϕ and SE_ϕ exhibit significant correlation with AUC albeit having lower range in its computation. Moreover, the ϕ index clearly reveals higher correlation with the measures YI, DI and CP used to set the best cut-point. Besides, the measures YI, DI and CP are strongly correlated with each other.

	AUC	SP_{90}	SE_{90}	SP_{75}	SE_{75}	SP_{50}	SE_{50}	ϕ	SP_ϕ	SE_ϕ	YI	DI	CP
AUC	1	.784	.784	.928	.927	.988	.988	.815	.869	.864	.841	.828	.837
SP_{90}	.791	1	.326	.912	.515	.818	.713	.550	.628	.575	.571	.560	.567
SE_{90}	.780	.329	1	.516	.910	.712	.817	.561	.584	.628	.585	.572	.581
SP_{75}	.935	.912	.531	1	.758	.959	.893	.810	.874	.841	.834	.823	.830
SE_{75}	.925	.525	.915	.766	1	.892	.957	.818	.848	.872	.844	.831	.840
SP_{50}	.991	.818	.718	.960	.896	1	.972	.831	.890	.880	.857	.844	.852
SE_{50}	.989	.722	.817	.902	.957	.975	1	.834	.885	.885	.862	.848	.858
ϕ	.823	.558	.592	.811	.830	.838	.842	1	.894	.890	.958	.977	.965
SP_ϕ	.866	.668	.576	.870	.816	.884	.876	.810	1	.981	.948	.931	.944
SE_ϕ	.862	.559	.690	.807	.881	.869	.885	.818	.882	1	.944	.927	.940
YI	.884	.634	.654	.865	.877	.896	.902	.913	.887	.890	1	.987	.998
DI	.863	.594	.629	.848	.868	.877	.884	.957	.888	.894	.975	1	.993
CP	.878	.619	.646	.860	.876	.891	.897	.937	.887	.891	.994	.989	1

Table 1: Spearman's rank correlation coefficient, with $n = 1000$ (upper triangular matrix) versus $n = 50$ (lower triangular matrix).

3.2. Normal distribution with different standard deviation

Let us now consider that the r.v. Y_i has Normal distribution with standard deviation $\sigma_0 = 1$ and mean $\mu_0 = 0$ in a healthy individual and $\mu_1 = 2$ in an infected individual. The standard deviation in an infected individual varies in $\sigma_1 \in \{2/3, 1, 1.5, 2, 3\}$ and we are collecting samples with size $n = 1000$. Obviously, the accuracy will get worse with the increase of σ_1. For $\sigma_1 \in \{2/3, 1\}$ the partial areas over the sensitivity are similar to the partial areas over the specificity (see Figure 6). Nevertheless, for $\sigma_1 \in \{1.5, 2\}$ the boxplots are quite different and for $\sigma_1 = 3$ the boxplot of SE_{90} is not even shown. Hence, this case reveals problems on the computation of the partial area over a range of high sensitivity. If we observe the last chart of Figure 6, for the worst plotted ROC curve the spAUC computed over the sensitivity range $[0.9, 1]$ would be lower (even after standardization) than 0.5 and, therefore, it is even worse than the random classification. Consequently, this measure is not shown. Note, also, that the worst ROC curves are not symmetric around $\varphi_e = \varphi_s$ and consequently the partial areas over the specificity have different behaviour comparing with the partial areas over the sensitivity. However, the partial areas over a neighbourhood of ϕ do not seem to have any problems in assessing accuracy.

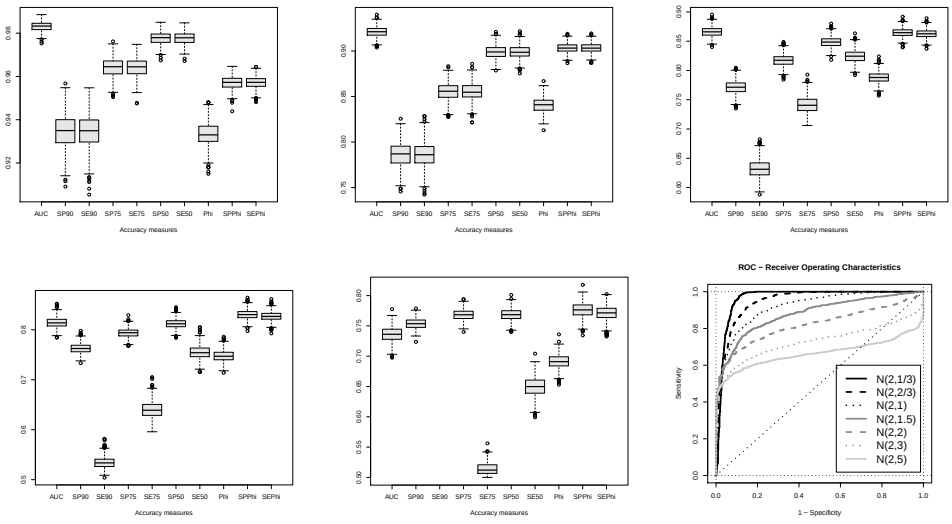

Figure 6: $\mathrm{D}_0(\theta_0) = \mathrm{N}(0, 1)$ versus $\mathrm{D}_1(\theta_1) = \mathrm{N}(2, 2/3)$ (top left), $\mathrm{N}(2, 1)$ (top middle), $\mathrm{N}(2, 1.5)$ (top right), $\mathrm{N}(2, 2)$ (bottom left), $\mathrm{N}(2, 3)$ (bottom middle), and ROC curves (bottom right), with $n = 1000$.

Table 2 displays the Spearman's rank correlation coefficients between all the computed measures when the r.v. Y_i is characterized by $\mathrm{N}(0, 1)$ for a healthy individual and characterized by $\mathrm{N}(2, 2/3)$ (upper triangular matrix) and $\mathrm{N}(2, 2)$ (lower triangular matrix) for an infected individual. There seems to be some

differences between the results obtained in these two situations, but the main conclusions appear to be the same. The correlation between the partial areas and the entire area continues to increase when the interval of interest increases to the support $[0, 1]$ and the ϕ index continues to reveal quite strong correlations with the measures YI, DI and CP. Even though, when $\sigma_1 = 2$ these correlations are lower.

	AUC	SP_{90}	SE_{90}	SP_{75}	SE_{75}	SP_{50}	SE_{50}	ϕ	SP_{ϕ}	SE_{ϕ}	YI	DI	CP
AUC	1	.890	.876	.969	.970	.997	.997	.771	.616	.599	.798	.786	.796
SP_{90}	.690	1	.597	.953	.801	.904	.878	.736	715.	.402	.703	.751	.761
SE_{90}	.721	.289	1	.779	.942	.860	.888	.753	.359	.656	.776	.767	.775
SP_{75}	.847	.906	.392	1	.919	.980	.963	.794	.673	.578	.824	.811	.823
SE_{75}	.952	.484	.825	.655	1	.963	.980	.807	.573	.637	.835	.823	.833
SP_{50}	.947	.793	.505	.945	.821	1	.993	.777	.629	.604	.805	.793	.803
SE_{50}	.998	.659	.730	.830	.960	.939	1	.780	.616	.605	.807	.795	.805
ϕ	.857	.579	.451	.815	.765	.909	.861	1	.286	.282	.959	.979	.963
SP_{ϕ}	.885	.603	.473	.842	.790	.936	.888	.982	1	.595	.403	.354	.394
SE_{ϕ}	.896	.606	.510	.813	.822	.924	.898	.843	.911	1	.405	.354	.396
YI	.793	.821	.367	.947	.608	.887	.783	.757	.785	.770	1	.990	.999
DI	.849	.656	.432	.897	.711	.919	.849	.908	.934	.886	.859	1	.992
CP	.825	.735	.397	.935	.657	.909	.820	.831	.859	.835	.949	.942	1

Table 2: Spearman's rank correlation coefficient, with $n = 1000$, $N(0, 1)$ versus $N(2, 2/3)$ (upper triangular matrix) and $N(2, 2)$ (lower triangular matrix).

3.3. Gamma **distribution**

Figure 7 and Table 3 show the results when the r.v. Y_i has Gamma distribution with $\alpha_0 = 2$, $\beta_0 = 1$ for a healthy individual, and $\alpha_1 \in \{6, 9, 12\}$ and $\beta_1 \in \{1, 3\}$ for an infected individual. The boxplots of the partial areas SP_{90} and SP_{75} relative to $D_1(\theta_1) = \text{Gamma}(6, 3)$ are not shown in Figure 7. It reveals problems on the computation of the partial area over a range of high specificity. If we observe the graph with the ROC curves, the curve with the worst performance is below the line of random classification in the high specificity values. Thus, the standardized partial area under the ROC curve would be lower than 0.5 (accuracy worse than in random classification). As in previous case, some of the ROC curves are not symmetric around $\varphi_e = \varphi_s$ and, therefore, the partial areas over the specificity are quite different from the partial areas over the sensitivity. Moreover, the partial areas over a neighbourhood of ϕ seem to continue to assess accuracy without revealing any problem, regardless of whether they are being computed over the sensitivity or over the specificity.

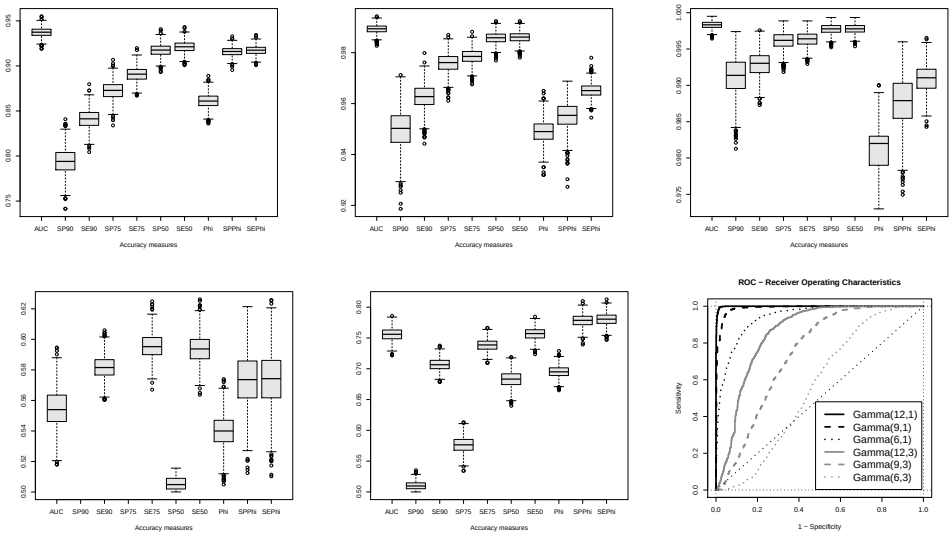

Figure 7: $\mathrm{D}_0(\theta_0) = \mathrm{Gamma}(2,1)$ versus $\mathrm{D}_1(\theta_1) = \mathrm{Gamma}(6,1)$ (top left), $\mathrm{Gamma}(9,1)$ (top middle), $\mathrm{Gamma}(12,1)$ (top right), $\mathrm{Gamma}(6,3)$ (bottom left), $\mathrm{Gamma}(9,3)$ (bottom middle), and ROC curves (bottom right), with $n = 1000$.

Table 3 displays the Spearman's rank correlation coefficients between all the computed measures when the r.v. Y_i is characterized by $\mathrm{Gamma}(2,1)$ for a healthy individual and characterized by $\mathrm{Gamma}(12,1)$ (upper triangular matrix) and $\mathrm{Gamma}(9,3)$ (lower triangular matrix) for an infected individual. In the $\mathrm{Gamma}(12,1)$ case, the AUC and the different spAUC are strongly correlated, but the rank correlations between AUC or any of the spAUC and the measures YI, DI and CP are not that significant. In fact, the index ϕ is the only accuracy measure that revels strong correlations with these indexes to select the optimal cut-point, albeit these correlations are not so significant in the $\mathrm{Gamma}(9,3)$ case.

3.4. Discrete distributions

In these last scenarios, two count distributions are analysed, the Binomial with n trials and success probability p ($\mathrm{B}(n,p)$) and the Geometric distribution with probability p ($\mathrm{G}(p)$). Hence, in the first scenario let the r.v. Y_i have $\mathrm{B}(20,p)$ with $p_0 = 0.25$ for a healthy individual and $p_1 \in \{0.5, 0.4, 0.3\}$ for an infected individual. In the second scenario, the r.v. Y_i is characterized by $\mathrm{G}(p)$ where $p_0 = 0.2$ for a healthy individual and $p_1 \in \{0.1, 0.02\}$ for an infected individual. The results in both scenarios do not reveal any problem in the calculation of any

	AUC	SP_{90}	SE_{90}	SP_{75}	SE_{75}	SP_{50}	SE_{50}	ϕ	SP_{ϕ}	SE_{ϕ}	YI	DI	CP
AUC	1	.994	.858	1.000	.966	1.000	.997	.638	.979	.802	.654	.656	.655
SP_{90}	.592	1	.836	.995	.956	.994	.990	.639	.990	.772	.652	.656	.653
SE_{90}	.657	.183	1	.856	.933	.858	.870	.811	.770	.984	.832	.830	.833
SP_{75}	.875	.778	.335	1	.965	1.000	.997	.639	.981	.800	.655	.657	.656
SE_{75}	.809	.250	.910	.465	1	.966	.975	.709	.923	.881	.727	.727	.728
SP_{50}	.993	.619	.571	.905	.751	1	.997	.638	.979	.802	.654	.656	.655
SE_{50}	.934	.357	.784	.660	.935	.904	1	.648	.974	.814	.666	.666	.667
ϕ	.859	.351	.540	.650	.736	.859	.892	1	.545	.794	.924	.958	.936
SP_{ϕ}	.899	.380	.552	.707	.759	.900	.916	.867	1	.697	.566	.566	.566
SE_{ϕ}	.886	.359	.558	.671	.761	.886	.920	.979	.935	1	.832	.822	.831
YI	.769	.230	.790	.442	.947	.730	.890	.709	.733	.733	1	.981	.998
DI	.849	.311	.627	.578	.860	.837	.925	.878	.901	.909	.841	1	.989
CP	.821	.276	.693	.518	.916	.798	.921	.800	.833	.829	.927	.948	1

Table 3: Spearman's rank correlation coefficient, with $n = 1000$, Gamma$(2, 1)$ versus Gamma$(12, 1)$ (upper triangular matrix) and Gamma$(9, 3)$ (lower triangular matrix).

of the spAUC, despite some of the ROC curves being asymmetric around the line $\varphi_e = \varphi_s$. Thus, in some cases the partial areas over the specificity assume different values when compared with the partial areas over the sensitivity, but all measures were computed in the analysed cases.

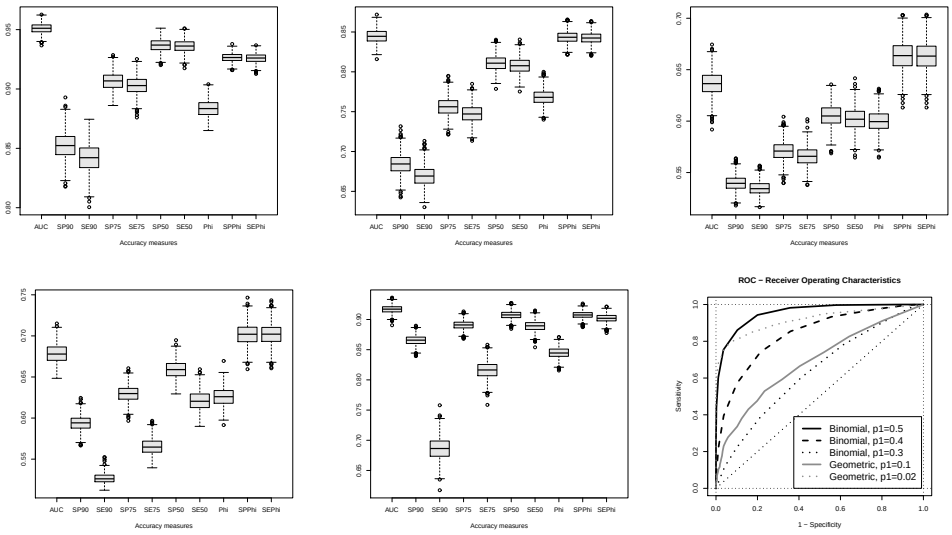

Figure 8: Binomial and Geometric distributions with $n = 1000$: $B(20, p), p_0 = \frac{1}{4}, p_1 = .5$ (top left), $B(20, p), p_0 = \frac{1}{4}, p_1 = .4$ (top middle), $B(20, p), p_0 = \frac{1}{4}, p_1 = .3$ (top right), $G(p), p_0 = .2, p_1 = .1$ (bottom left), $G(p), p_0 = .2, p_1 = .02$ (bottom middle), and ROC curves (bottom right).

Table 4 shows the Spearman's rank correlation coefficients between all the computed measures when the r.v. Y_i is characterized by $B(20, 0.25)$ for a healthy individual and by $B(20, 0.5)$ for an infected individual (upper triangular matrix), and Geometric(0.2) for a healthy individual versus Geometric(0.1) for an infected individual (lower triangular matrix). The results using count distributions appear to be similar to those previously obtained with the use of continuous distributions. Thus, the ϕ index continues to present very significant correlations with the measures YI, DI and CP, higher in the Binomial case than in the Geometric case.

	AUC	SP_{90}	SE_{90}	SP_{75}	SE_{75}	SP_{50}	SE_{50}	ϕ	SP_{ϕ}	SE_{ϕ}	YI	DI	CP
AUC	1	.822	.837	.941	.954	.991	.994	.798	.905	.870	.799	.799	.798
SP_{90}	.571	1	.410	.934	.641	.851	.783	.703	.799	.599	.704	.672	.699
SE_{90}	.598	.126	1	.628	.928	.791	.860	.562	.670	.810	.563	.576	.566
SP_{75}	.786	.860	.235	1	.839	.966	.923	.853	.942	.830	.854	.834	.851
SE_{75}	.786	.203	.885	.352	1	.934	.973	.801	.886	.916	.801	.802	.802
SP_{50}	.936	.689	.351	.909	.533	1	.983	.823	.929	.881	.823	.812	.822
SE_{50}	.955	.357	.687	.582	.883	.816	1	.810	.915	.889	.811	.803	.810
ϕ	.818	.318	.375	.556	.564	.806	.830	1	.920	.804	1.00	.990	1.00
SP_{ϕ}	.893	.397	.387	.641	.596	.892	.889	.924	1	.910	.921	.897	.918
SE_{ϕ}	.888	.414	.378	.649	.589	.890	.877	.847	.983	1	.805	.801	.805
YI	.854	.541	.310	.844	.462	.933	.745	.744	.829	.829	1	.990	1.00
DI	.862	.405	.344	.668	.532	.891	.835	.881	.960	.952	.869	1	.993
CP	.865	.422	.339	.697	.522	.904	.827	.863	.947	.941	.901	.990	1

Table 4: Spearman's rank correlation coefficient, with $n = 1000$, $B(20, 0.25)$ versus $B(20, 0.5)$ (upper triangular matrix) and Geometric(0.2) versus Geometric(0.1) (lower triangular matrix).

3.5. Sensitivity and specificity on the optimal cut-point

The first quartile q_1 and the third quartile q_3 of the sensitivity φ_s and of the specificity φ_e on the cut-points selected by the application of the YI, DI and CP criteria are displayed on Table 5. It is also shown q_1 and q_3 of the ϕ index, in which $\varphi_s = \varphi_e$ or, at least, its distance is minimized and $\phi = \frac{\varphi_s + \varphi_e}{2}$. The results clearly stand out the diverge accuracy levels obtained when the cut-points are set by YI, DI, and CP. Moreover, the results suggest that these differences may occur in any sense, i.e. none of these measures gives priority to sensitivity or to specificity in relation to the others measures. For example, the cut-point selected by the YI generates better sensitivity (consequently worse specificity) in the cases Gamma(. , 3) but generates worse sensitivity (and better specificity) in the $N(2, 1.5)$ or $G(0.10)$ cases. On the other hand, the cut-point selected by

the DI criterion generates better sensitivity (consequently worse specificity) in the cases N(2, .) but generates worse sensitivity (and better specificity) in the Gamma(. , 3) cases. Therefore, the accuracy of the cut-points selected through theses procedures must be evaluated and compared in each application.

Let us also point out that the ϕ index can also be used to select the cut-point in any application, as the results displayed in Table 5 prove. In this case the balance between sensitivity and specificity is a priority (these measures are the same, or at least very close), with a clear reduction in the variation of these measurements (as we can ascertain by comparing the interquartile range).

	YI				DI				CP				ϕ	
	φ_s		φ_e		φ_s		φ_e		φ_s		φ_e		$\varphi_s = \varphi_e$	
	q_1	q_3	q_1	q_3	q_1	q_3	q_1	q_3	q_1	q_3	q_1	q_3	q_1	q_3
n=50	.800	.900	.840	.920	.820	.885	.840	.900	.820	.900	.840	.905	.820	.860
n=100	.810	.890	.840	.910	.830	.880	.830	.880	.820	.890	.830	.890	.820	.860
n=250	.816	.876	.828	.888	.832	.864	.832	.868	.828	.872	.828	.876	.828	.852
n=500	.822	.868	.828	.874	.834	.860	.834	.860	.830	.866	.828	.866	.834	.850
n=1000	.825	.861	.829	.865	.833	.855	.834	.855	.830	.860	.830	.860	.835	.846
N(2, 2/3)	.927	.944	.928	.946	.931	.942	.929	.941	.929	.944	.927	.944	.930	.937
N(2, 1.5)	.710	.751	.849	.887	.758	.782	.808	.836	.733	.768	.829	.865	.782	.794
N(2, 2)	.631	.672	.880	.914	.701	.725	.803	.833	.667	.700	.843	.879	.740	.755
N(2, 3)	.554	.587	.923	.949	.631	.655	.810	.843	.592	.621	.875	.910	.684	.699
Ga(6, 1)	.869	.898	.832	.864	.864	.882	.847	.866	.869	.895	.835	.864	.856	.866
Ga(9, 1)	.952	.967	.937	.952	.950	.961	.942	.953	.953	.966	.938	.952	.946	.952
Ga(12, 1)	.982	.989	.977	.985	.982	.987	.979	.984	.982	.989	.977	.984	.979	.983
Ga(6, 3)	.854	.906	.283	.338	.642	.693	.454	.489	.668	.726	.434	.474	.533	.547
Ga(9, 3)	.852	.894	.555	.601	.756	.792	.641	.668	.797	.840	.608	.642	.689	.701
Ga(12, 3)	.889	.922	.718	.755	.840	.867	.766	.787	.875	.905	.736	.766	.795	.808
B(50, .5)	.861	.877	.892	.904	.861	.876	.892	.904	.861	.876	.892	.904	.861	.876
B(50, .4)	.741	.759	.778	.795	.741	.759	.778	.795	.741	.759	.778	.795	.741	.759
B(50, .3)	.571	.597	.606	.631	.572	.593	.606	.627	.572	.593	.606	.627	.572	.593
G(.10)	.470	.546	.733	.802	.581	.606	.658	.683	.570	.602	.663	.696	.608	.655
G(.02)	.774	.804	.918	.943	.815	.833	.875	.898	.787	.813	.906	.932	.839	.852

Table 5: First and third quartiles of sensitivity and specificity on the optimal cut-point.

4. CONCLUSION — FINAL REMARKS

In most situations AUC, spAUC and ϕ are strongly correlated and, therefore, seem to be able to evaluate the same criterion of accuracy. Nevertheless, AUC shows less variability than spAUC, mainly on small samples and in cases with worse accuracy. Moreover, spAUC with sensitivity or specificity over $[\phi - 0.05, \min\{\phi + 0.05, 1\}]$ shows less variability than over $[0.9, 1]$, $[0.75, 1]$ or even $[0.5, 1]$, albeit assessing a smaller range. In some cases, it is not possible

to compute the spAUC over a range of sensitivity using the Normal distribution and over a range of specificity using the Gamma distribution. Actually, in some situations the spAUC seems to provide better results when computed over a range of specificity (rather than sensitivity), but the opposite may also occur in other cases. However, the partial areas computed over a neighbourhood of ϕ do not seem to have any problem in assessing accuracy even when the ROC curve crosses the diagonal line and, therefore, it enables to overcome the main drawback usually identified in the application of the spAUC. Furthermore, the ϕ index has higher correlation with YI, DI, CP than AUC or any of the computed spAUC. In fact, the ϕ index seems to be the measure with higher rank correlation with the sensitivity and specificity of the optimal cut-point selected by the use of any of the analysed optimization criteria. Additionally, this index can also be applied to select the optimal cut-point, ensuring a balance between sensitivity and specificity. Accordingly, this index seems to perform better in the evaluation of the most appropriate model as well as in the selection of the optimal cut-point. Finally, it is equally important to point out that the cut-points set by YI, DI, and CP can, in some cases, be quite different and generate significantly distinct accuracy measures. Hence, in each application their performances should be evaluated and the selected cut-points compared.

In fact, the variability of the diagnostic accuracy measures in simulations under the same scenario is quite high and, therefore, the obtained estimates do not always reveal the true accuracy of the applied classification procedure. Hence, new estimation techniques for these measures (or other measures) must be investigated in order to minimize this variability and to achieve more robust estimates, for example applying bootstrap or other resampling techniques.

ACKNOWLEDGMENTS

This work has been funded by FCT - Fundação Nacional para a Ciência e Tecnologia, Portugal, through the projects UID/MAT/00006/2013, UID/MAT/04561/2013, UID/MAT/00006/2019 and UID/MAT/04561/2019.

REFERENCES

[1] Allen, F.; Behan, F.; Khodak, A.; Iorio, F.; Yusa, K.; Garnett, M. and Parts, L. (2019). JACKS: joint analysis of CRISPR/Cas9 knock-out screens, *Genome Research*, available online since January 23, 2019.

[2] Chen, M.L.; Davit, B.; Lionberger R.; Wahba Z.; Ahn H.Y. and Yu L.X. (2011). Using partial area for evaluation of bioavailability and bioequivalence, *Pharmaceutical Research*, **28**, 8, 1939–1947.

[3] Dodd, L.E. and Pepe, M.S. (2003). Partial AUC estimation and regression, *Biometrics*, **59**, 3, 614–623.

[4] Fluss, R.; Faraggi, D. and Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point, *Biometrical Journal*, **47**, 4, 458–472.

[5] Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, *Caspian Journal of Internal Medicine*, **4**, 2, 627–635.

[6] Hanley, J.A. and McNeil, J.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29–36.

[7] Hanley, J.A. (1989). Receiver operating characteristic (ROC) methodology: state of the art, *Critical Reviews in Diagnostic Imaging*, **29**, 307–335.

[8] Jiang, Y.; Metz, C.E. and Nishikawa, R.M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests, *Radiology*, **201**, 745–750.

[9] Krazanowski, W.J. and Hand, D.J. (2009). *ROC Curves for Continuous Data*, CRC press, New York.

[10] Lee, L.H.N.; Choi, C.; Gershkovich, P.; Barr, A.M.; Honer, W.G. and Procyshyn, R.M. (2016). Proposing the use of partial AUC as an adjunctive measure in establishing bioequivalence between deltoid and gluteal administration of long-acting injectable antipsychotics, *European Journal of Drug Metabolism and Pharmacokinetics*, **41**, 659–664.

[11] Liu, X. (2012). Classification accuracy and cut point selection, *Statistics in Medicine*, **31**, 2676–2686.

[12] Liu, X. and Jin Z. (2007). Item reduction in a scale for screening, *Statistics in Medicine*, **26**, 23, 4311–4327.

[13] Ma, H.; Bandos, A. and Gur, D. (2015). On the use of partial area under the ROC curve for comparison of two diagnostic tests, *Biometrical Journal*, **57**, 304–320.

[14] Ma, H.; Bandos, A.; Rockette, H. and Gur, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance, *Statistics in Medicine*, **32**, 3449–3458.

[15] Metz, C.E. (2008). ROC analysis in medical imaging: A tutorial review of the literature, *Radiological Physics and Technology*, **1**, 2–12.

[16] Narasimhan, H. and Agarwal, S. (2013). A structural SVM based approach for optimizing partial AUC, *Proceedings of the 30th International Conference on Machine Learning*, **28**.

[17] Narasimhan, H. and Agarwal, S. (2017). Support vector algorithms for optimizing the partial area under the ROC curve, *Neural Computation*, **29**, 7, 1919–1963.

[18] Pepe. M.S. (2003). *Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford.

[19] Perkins, N.J. and Schisterman, E.F. (2006). The inconsistency of "optimal" cut-points using two ROC based criteria, *American Journal of Epidemiology*, **53**, 670–675.

[20] Powers, D. (2011). Evaluation: from precision, recall and F-score to ROC, informedness, markedness & correlation, *Journal of Machine Learning Research*, **2**, 37–63.

[21] Robin, X.; Turck, N; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C. and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves, *Bioinformatics*, **12**, 1–8.

[22] Rota, M. and Antolini, L. (2014). Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers, *Computational Statistics and Data Analysis*, **69**, 1–14.

[23] Santos, R.; Martins, J.P. and Felgueiras, M. (2015). *An overview of quantitative continuous compound tests*. In "Dynamics, Games and Science" (J.P. Bourguignon, R. Jeltsch, A. Pinto and M. Viana, Eds.), CIM Series in Mathematical Sciences **1**, 627–641.

[24] Santos, R.; Felgueiras, M. and Martins, J.P. (2015). *Discrete compound tests and Dorfman's methodology in the presence of misclassification*. In "Theory and Practice of Risk Assessment" (C.P. Kitsos, T. Oliveira, A. Rigas and S. Gulati, Eds.), Springer Proceedings in Mathematics & Statistics **136**, 85–98.

[25] Santos, R.; Martins, J.P.; Felgueiras, M. and Ferreira, L. (2017). *Binary classification based on a quantitative variable – an accuracy comparison by simulation*. In "Proceedings of 17th International Conference Computational and Mathematical Methods in Science and Engineering" (J. Vigo-Aguiar, Eds.), 1883–1886.

[26] Santos, R.; Martins, J.P.; Felgueiras, M. and Ferreira, L. (2018). *Medidas de fiabilidade de classificação binária com base numa variável quantitativa – uma comparação via simulação*. In "Livro de Atas do III Encontro Luso-Galaico de Biometria" (M. Monteiro, A. Freitas, L. Teixeira and M. Costa, Eds.), Sociedade Portuguesa de Estatística, 86–89.

[27] Schisterman, E.F.; Perkins N.J.; Liu A. and Bondell H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples, *Epidemiology*, **16**, 1, 73–81.

[28] Sing, T.; Sander, O.; Beerenwinkel, N.; and Lengauer, T. (2005). ROCR: visualizing classifier performance in R, *Bioinformatics*, **21**, 20, 3940–3941.

[29] Unal, I. (2017). Defining an optimal cut-point value in ROC analysis: an alternative approach, *Computational and Mathematical Methods in Medicine*, Article ID 3762651, 14 pages.

[30] Vivo, J.-M.; Franco, M. and Vicari, D. (2018). Rethinking an ROC partial area index for evaluating the classification performance at a high specificity range, *Advances in Data Analysis and Classification*, **12**, 683–704.

[31] Walter, S.D. (2005). The partial area under the summary ROC curve, *Statistics in Medicine*, **53**, 2025–2040.

[32] Witten, E. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, *Caspian Journal of Internal Medicine*, **4**, 627–635.

[33] Yang H.; Lu K.; Lyu, X. and Hu, F. (2019). Two-way partial AUC and its properties, *Statistical Methods in Medical Research*, **28**, 1, 184–195.

[34] Youden, W.J. (1950). Index for rating diagnostic tests, *Cancer*, **3**, 32–35.

[35] Zhou, X.H.; Obuchowski, N.A. and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*, Wiley & Sons, New York.

[36] Zou, K.H.; Yu, C.R.; Liu, K.; Carlsson, M.O. and Cabrera, J. (2013). Optimal thresholds by maximizing or minimizing various metrics via ROC-type analysis, *Academic Radiology*, **20**, 7, 807–815.