Received: 15 January 2025 | Revised: 20 May 2025 | Accepted: 22 July 2025

DOI: 10.1111/sjos.70010

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Distorted distributions and ROC curves

Marco Capaldo[1] | **Jorge M. Arevalillo**[2] | **Jorge Navarro**[3]

[1]Dipartimento di Matematica, Università degli Studi di Salerno, Fisciano, Italy

[2]Department of Statistics and Operational Research, University Nacional Educación a Distancia (UNED), Madrid, Spain

[3]Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Murcia, Murcia, Spain

Correspondence
Marco Capaldo, Institute of Statistics, RWTH Aachen University, 52056 Aachen, Germany.
Email: capaldo@isw.rwth-aachen.de

Funding information
"European Union Next Generation EU" through MUR-PRIN 2022 "Anomalous Phenomena on Regular and Irregular Domains: Approximating Complexity for the Applied Sciences", Grant/Award Number: 2022XZSAFN; MUR-PRIN 2022 PNRR "Stochastic Models in Biomathematics and Applications", Grant/Award Number: P2022XSF5H; Ministerio de Ciencia e Innovacion of Spain (MCIN/AEI), Grant/Award Number: PID2022-137396NBI00

Abstract

We consider new distortion functions based on the receiver operating characteristic (ROC) and ordinal dominance (OD) curves. Various stochastic orders and aging properties are characterized by using the interpretation of ROC and OD distortions as cumulative distribution functions of suitable relative random variables. Connections with Lorenz curve and Gini's index are examined too. We also define further distortion functions based on the concept of partial area under the ROC curve and partial area under the OD curve, by pointing out their connections with recent univariate skewed models and with the equilibrium distribution. Our theoretical findings are illustrated in different settings through real data applications of the distortion-based approach by using a semiparametric estimation of the ROC curve.

KEYWORDS

area under the ROC curve, binary classification, distortion function, equilibrium distribution, ROC curve, skewed model, stochastic order

1 | INTRODUCTION

The receiver operating characteristic (ROC) curve is a plot of the true positive rate (TPR, sensitivity or probability of detection) against the false positive rate (FPR, probability of false

alarm or error of type I) at each threshold setting in a binary diagnosis (classification) test, see Bamber (1975) and Calì & Longobardi (2015). ROC analysis has been historically linked to radar detection theory developed by engineers for studying signal to noise relations (this justify its name), aiming to detect enemy objects in battlefields during World War II. In spite of this historical commonly accepted background, to the best of the authors knowledge, the first work dealing with the ROC curve in the form in which it is known nowadays did not appear until the early fifties in a technical report by Peterson & Birdsall (1953). Since then ROC analysis has expanded its scope by deploying applications to problems in many fields such as medical diagnostics (Bradley, 1997; Junge & Dettori, 2018), psychology and psychiatry (Lung & Lee, 2008; Mickes et al., 2012; Weinstein et al., 1989), social studies (Ewanation et al., 2023), environmental studies (Jurdao et al., 2012; Pontius & Schneider, 2001), and business analytics (Gigliarano et al., 2014; Huang & Kechadi, 2013; Moro et al., 2014; Siddiqi, 2012; Vanneschi et al., 2018) to name a few. The reader interested in more applications and developments of ROC analysis is referred to the monograph Krzanowski & Hand (2009) and the references therein.

A dual approach to the ROC analysis is described by the ordinal dominance (OD) curve, introduced by Bamber (1975). Indeed, the OD curve is a plot of the true negative rate (specificity or selectivity) against false negative rate (miss rate or error of type II) at each threshold setting in a binary diagnosis test. For a detailed description of ROC and OD curves with other detecting difference curves, we refer the reader to Ledwina & Zagdański (2024).

The concept of distortion function was introduced by Yaari (1987) in the context of theory of choice under risk, aiming to change an initial risk distribution with a distorted risk distribution. Nowadays distortion functions are applied in several fields, such as Reliability Theory (cf. Section 2.4 of Navarro, 2022), Order Statistics (see, for instance, Burkschat & Navarro, 2018), and Insurance (cf., for example, Sordo & Suárez-Llorens, 2011). In Capaldo et al. (2025), the authors introduced a new distortion function based on the ROC curve, and they used it to specify the dependence (described by copulas) of some pairs of reliability systems with shared components. For connections between ROC curves and distortion risk measures, see Schumacher (2018).

In this work, we interpret ROC and OD curves as distortion functions. After definitions and main properties, by considering some examples, we point out that ROC and OD distortion curves represent respectively the distortion function that connects two suitable survival functions or cumulative distribution functions when distortion-based model hypothesis are unknown. This can be also used in the context of weighted distributions. Indeed, the Lorenz curve of a given random variable X can be expressed as the OD distortion curve between X and its length biased version. Then, the Gini's index is related with the area under the ROC (AUROC) and the area under the OD (AUOD) between a given baseline distribution and its length biased version. Many stochastic comparisons and aging results regarding ROC and OD distortion curves are investigated as well, by taking into account that such new distortions are cumulative distribution functions of suitable relative random variables. For further stochastic comparisons related to Lorenz and ROC curves, see respectively Lando & Legramanti (2025) and Ramos et al. (2019).

We also define two new distortion functions based on the concept of partial AUROC and partial AUOD, which are related respectively with the left-skewed and right-skewed models proposed in Navarro & Arevalillo (2023). They are also linked with the equilibrium distribution of the relative random variables mentioned above. Some stochastic ordering and aging results are exhibited as well.

Finally, we apply the distortion functions introduced above to evaluate the performance of some Machine Learning (ML) classifiers, when used for churn prediction modeling, and the performance for medical diagnosis of several biomarkers measured after an aneurysmal

subarachnoid hemorrhage episode. For this aim, we consider two empirical versions of the ROC respectively based on empirical distributions and Gaussian kernel density functions. In addition, we propose two semiparametric estimation approaches related to the proportional hazard rate (PHR) Cox model (see Cox, 1972) and the proportional odds (PO) model (see Sankaran & Jayakumar, 2008). This allows us to provide a real data application of some theoretical results proved throughout the present work. For estimation of the ROC curve from the Lehmann family, see Jokiel-Rokita & Topolnicki (2020).

The rest of the paper is organized as follows. Some useful basic notions are recalled in Section 2. In Section 3, we introduce and study the ROC and OD distortion curves, by considering various related interpretations and examples. In Section 4, we provide many stochastic comparisons and aging results. Section 5 is devoted to define and study two new distortion functions related to the concepts of partial AUROC and partial AUOD. Our theoretical findings are applied to real data cases for the sake of illustration in Section 6. Finally, some concluding remarks are given in Section 7.

2 | PRELIMINARY NOTIONS

In this section, by fixing the notation, we recall some useful notions, such as stochastic orders, aging properties, weighted and distorted distributions. Throughout the paper the terms increasing and decreasing are used respectively as "non-decreasing" and "non-increasing", while "iff" denotes "if and only if".

For a given random variable X, we denote its cumulative distribution function (CDF) as $F(t) = \mathsf{Pr}(X \le t)$, for $t \in \mathbb{R}$, and its survival function (SF) as $\overline{F} = 1 - F$. For $u \in [0,1]$, the function $F^{-1}(u) = \sup\{t : F(t) \le u\}$ represents the right-continuous version of the inverse of F, namely quantile function. Note that $F^{-1}(1) = +\infty$, for all F. In addition, we denote by $\overline{F}^{-1}(u) = F^{-1}(1-u)$ for all $u \in [0,1]$, that is left-continuous with $\overline{F}^{-1}(0) = +\infty$. The mean or expected value of X can be expressed as

$$\mu = \mathsf{E}(X) = -\int_{-\infty}^{0} F(t)\mathrm{d}t + \int_{0}^{+\infty} \overline{F}(t)\mathrm{d}t$$

and we assume that μ is finite. If X has an absolutely continuous distribution, then its probability density function (PDF) is defined as $f = F'$. Moreover, the hazard rate (HR) of X is $\lambda(t) = f(t)/\overline{F}(t)$ for all t such that $\overline{F}(t) > 0$, while the reversed hazard rate (RHR) of X is $\tau(t) = f(t)/F(t)$ for all t such that $F(t) > 0$.

2.1 | Stochastic orders and aging properties

Now we recall some stochastic orders related with the quantities defined above. Here, $a/0$ is taken to be equal to $+\infty$ whenever $a > 0$, while the subscript refers to the random variables.

Definition 2.1. We say that X is smaller than Y in the

- usual stochastic order, denoted by $X \le_{st} Y$, if $\overline{F}_X(t) \le \overline{F}_Y(t)$ holds for all t. If there is equality in law, then we write $X =_{st} Y$;
- hazard rate order, denoted by $X \le_{hr} Y$, if $\overline{F}_Y(t)/\overline{F}_X(t)$ is increasing in t;

TABLE 1 Relationships among the stochastic orders introduced in Definition 2.1.

$X \leq_{lr} Y$	$\Rightarrow$	$X \leq_{hr} Y$
$\Downarrow$		$\Downarrow$
$X \leq_{rhr} Y$	$\Rightarrow$	$X \leq_{st} Y$

TABLE 2 Relationships among the aging classes introduced in Definition 2.2.

		DRHR	$\leftarrow$	ILR	$\rightarrow$	IHR		
	$\nearrow$						$\nwarrow$	
DD								ID
	$\nwarrow$						$\nearrow$	
		DHR	$\leftarrow^{*}$	DLR	$\rightarrow^{**}$	IRHR		

Note: Here, "$A \rightarrow B$" stands "A implies B", where $\leftarrow^{*}$ holds only if $-\infty < l < r = +\infty$, while $\rightarrow^{**}$ holds only if $-\infty = l < r < +\infty$, with $l, r \in \mathbb{R}$ denoting lower and upper limits for the support of the involved random variable.

- reversed hazard rate order, denoted by $X \leq_{rhr} Y$, if $F_Y(t)/F_X(t)$ is increasing in t; and
- likelihood ratio order, denoted by $X \leq_{lr} Y$, if $f_Y(t)/f_X(t)$ is increasing in t in the union of their supports, provided that X and Y have absolutely continuous distributions.

In the absolutely continuous case, one also has $X \leq_{hr} Y$ iff $\lambda_X(t) \geq \lambda_Y(t)$ for all t, while $X \leq_{rhr} Y$ iff $\tau_X(t) \leq \tau_Y(t)$ for all t. The relationships among the stochastic orders introduced in Definition 2.1 are summarized in Table 1. Note that the reverse implications do not hold in general. For a detailed description about stochastic orders, we refer the reader to Belzunce et al. (2016), Müller & Stoyan (2002), and Shaked & Shanthikumar (2007).

Next we define some useful aging properties.

Definition 2.2. A random variable X with an absolutely continuous distribution is said to have the

- increasing hazard rate (IHR) property if its HR λ is increasing. If λ is decreasing, then X is said to have the decreasing hazard rate (DHR) property;
- increasing reversed hazard rate (IRHR) property if its RHR τ is increasing. If τ is decreasing, then X is said to have the decreasing reversed hazard rate (DRHR) property;
- increasing in likelihood ratio (ILR) property if its PDF f is log-concave. If f is log-convex, then X is said to have the decreasing in likelihood ratio (DLR) property;
- increasing density (ID) property if f is increasing. If f is decreasing, then X is said to have the decreasing density (DD) property.

The relationships among the aging properties introduced in Definition 2.2 are shown in Table 2. Note that the reverse implications are not true in general. More details about aging

classes can be found in Belzunce et al. (2016), Section 4.1 of Navarro (2022) and Shaked & Shanthikumar (2007).

2.2 | Weighted and distorted distributions

Weighted or biased distributions are used to perturb an original distribution by multiplying a baseline PDF with a suitable weight function (cf. Shaked & Shanthikumar 2007, p. 28). In particular, for a given $w : \mathbb{R} \to \mathbb{R}$ non-negative weight function, if X has an absolutely continuous distribution with PDF f, then the weighted random variable X_w has PDF defined as

$$f_w(t) = \frac{w(t)f(t)}{\mu_w}, \qquad t \in \mathbb{R}, \tag{2.1}$$

by assuming that $\mu_w = \int_{\mathbb{R}} w(t)f(t)\mathrm{d}t$ is finite and non-zero. We remark that the weight function w appearing in Equation (2.1) is used to change the sampling probabilities of the original distribution.

Some particular choices of w in Equation (2.1) may conduct to relevant models. For example, if X is non-negative with mean $0 < \mu < +\infty$ and HR λ, then $w = 1/\lambda$ in Equation (2.1) leads to

$$f^e(t) = \frac{\overline{F}(t)}{\mu}, \qquad t \geq 0, \tag{2.2}$$

that is the PDF of X^e, namely the equilibrium random variable of X (cf. Shaked & Shanthikumar, 2007, p. 15). It is also known as stationary renewal distribution since it arises as the limiting distribution of the forward recurrence time in a renewal process.

Other alternatives for the weight w in Equation (2.1) have been considered in Navarro & Arevalillo (2023), who proposed a novel skewing strategy of a given baseline PDF. In particular, let us assume that X has an absolutely continuous distribution with PDF f, while Y has CDF G and SF $\overline{G}$, respectively. If X and Y are independent, then the left-skewed random variable X_L has PDF

$$f_L(t) = \frac{\overline{G}(t)f(t)}{\mathsf{Pr}(Y > X)}, \qquad t \in \mathbb{R}, \tag{2.3}$$

while the right-skewed random variable X_R has PDF

$$f_R(t) = \frac{G(t)f(t)}{\mathsf{Pr}(Y \leq X)}, \qquad t \in \mathbb{R}. \tag{2.4}$$

It can be noted that $X_L \leq_{lr} X \leq_{lr} X_R$ (cf. Proposition 2.2 in Navarro & Arevalillo, 2023).

An alternative way for the perturbation of a given distribution is provided by the concept of distortion function introduced by Yaari (1987). The formal definition is stated as follows.

> **Definition 2.3.** A distortion function is an increasing continuous function $q : [0, 1] \to [0, 1]$, such that $q(0) = 0$ and $q(1) = 1$.

Hence, the distorted CDF from F through q is given by $F_q = q(F)$. In addition, the dual distortion function with respect to q is defined as $\tilde{q}(u) = 1 - q(1 - u)$ for $u \in [0, 1]$. Then,

$\overline{F}_q = \tilde{q}(\overline{F})$ is the distorted SF of F_q. For further details about distortion functions, see Section 2.4 in Navarro (2022).

3 | ROC AND OD DISTORTION CURVES

In this section, we interpret the ROC and OD curves as distortion functions, by providing some basic properties and examples. We show how such new distortions can be connected with weighted models. We emphasize their relation with the Lorenz curve and the Gini's index. In addition, ROC and OD distortion curves represent CDFs of suitable relative random variables.

The formal definition of the ROC distortion curve is stated as follows.

Definition 3.1. Let X and Y be random variables with absolutely continuous SFs $\overline{F}$ and $\overline{G}$, respectively. We assume that X and Y have an interval support with the same initial point. Then, the ROC distortion curve between X and Y is defined as

$$ROC_{\overline{G},\overline{F}}(u) = \overline{G}(\overline{F}^{-1}(u)), \qquad u \in [0,1]. \tag{3.1}$$

Note that the ROC distortion curve in Equation (3.1) is increasing and continuous in $u \in [0,1]$, with $ROC_{\overline{G},\overline{F}}(0) = 0$ and $ROC_{\overline{G},\overline{F}}(1) = 1$, and in this sense it is a proper distortion function. It allows us to switch from $\overline{F}$ to $\overline{G}$, i.e., $ROC_{\overline{G},\overline{F}}(\overline{F}(t)) = \overline{G}(t)$ for all t.

Let us now provide the formal definition of the OD distortion curve.

Definition 3.2. Let X and Y be random variables with absolutely continuous CDFs F and G, respectively. We assume that X and Y have an interval support with the same initial point. Then, the OD distortion curve between X and Y is defined as

$$OD_{G,F}(u) = G(F^{-1}(u)), \qquad u \in [0,1]. \tag{3.2}$$

We remark that the OD distortion curve in Equation (3.2) is a proper distortion function since it is increasing and continuous in $u \in [0,1]$, with $OD_{G,F}(0) = 0$ and $OD_{G,F}(1) = 1$. It allows us to switch from F to G since $OD_{G,F}(F(t)) = G(t)$ for all t.

It is noteworthy that the distortion functions defined in Equations (3.1) and (3.2) are one from each other dual versions. Indeed, the dual ROC distortion curve for $u \in [0,1]$ is

$$\widetilde{ROC}_{\overline{G},\overline{F}}(u) = 1 - ROC_{\overline{G},\overline{F}}(1-u) = 1 - \overline{G}(\overline{F}^{-1}(1-u)) = OD_{G,F}(u), \tag{3.3}$$

while the dual OD distortion curve for $u \in [0,1]$ is

$$\widetilde{OD}_{G,F}(u) = 1 - OD_{G,F}(1-u) = 1 - G(F^{-1}(1-u)) = ROC_{\overline{G},\overline{F}}(u), \tag{3.4}$$

and we get $ROC_{\overline{G},\overline{F}}(u) + OD_{G,F}(1-u) = 1$ for all $u \in [0,1]$. In addition, if $X =_{st} Y$, then $ROC_{\overline{G},\overline{F}}(u) = OD_{G,F}(u) = u$ for all $u \in [0,1]$.

A useful index to summarize the ROC curve is the AUROC which is defined by

$$AUROC_{\overline{G},\overline{F}} = \int_0^1 ROC_{\overline{G},\overline{F}}(u)\mathrm{d}u. \tag{3.5}$$

TABLE 3 Summary of concepts and acronyms defined in Section 3.

Acronym	Concept	Formula
ROC	Receiver Operating Characteristic	(3.1)
OD	Ordinal Dominance	(3.2)
AUROC	Area Under the Receiver Operating Characteristic	(3.5)
AUOD	Area Under the Ordinal Dominance	(3.6)

Similarly, we define the AUOD as

$$AUOD_{G,F} = \int_0^1 OD_{G,F}(u)\mathrm{d}u \tag{3.6}$$

and clearly one has $AUROC_{\overline{G},\overline{F}} + AUOD_{G,F} = 1$. We have the following immediate result.

Proposition 3.1. *If X and Y are independent, then*

$$AUROC_{\overline{G},\overline{F}} = \Pr(Y > X), \qquad AUOD_{G,F} = \Pr(Y \leq X).$$

In Table 3, we summarize the concepts and the acronyms stated above. Note that, for a given distortion function q with dual distortion $\tilde{q}$, if $G = q(F)$ (and, thus, $\overline{G} = \tilde{q}(\overline{F})$), then from Equations (3.1) and (3.2) it respectively follows $ROC_{\overline{G},\overline{F}}(u) = \tilde{q}(u)$ and $OD_{G,F}(u) = q(u)$ for all $u \in [0,1]$. Conversely, when distortion-based model assumptions are unknown, then $ROC_{\overline{G},\overline{F}}$ can be used to describe the distortion function which connects $\overline{F}$ to $\overline{G}$, and, similarly, $OD_{G,F}$ provides the distortion to get G from F. Let us consider two examples.

Example 3.1. Let X be exponentially distributed with SF $\overline{F}(t) = e^{-\lambda t}$, for $t \geq 0$ and $\lambda > 0$. Let Y be Weibull distributed with SF $\overline{G}(t) = e^{-(\lambda t)^{\alpha}}$, for $t \geq 0$, $\lambda > 0$ and $\alpha > 0$. Then, from Equation (3.1), we get

$$ROC_{\overline{G},\overline{F}}(u) = e^{-(-\ln(u))^{\alpha}}, \qquad u \in [0,1],\ \alpha > 0, \tag{3.7}$$

that is plotted in the left-hand side of Figure 1 for $\alpha = 1/3, 1/2, 1, 2$. For such choices of α one respectively has $AUROC_{\overline{G},\overline{F}} = 0.4311101, 0.4543586, 0.5, 0.5456414$. In addition, from Equation (3.2), we have

$$OD_{G,F}(u) = 1 - e^{-(-\ln(1-u))^{\alpha}}, \qquad u \in [0,1],\ \alpha > 0, \tag{3.8}$$

which is plotted in the right-hand side of Figure 1 for $\alpha = 1/3, 1/2, 1, 2$. For such choices of α one respectively gets $AUOD_{G,F} = 0.5688899, 0.5456414, 0.5, 0.4543586$.

Example 3.2. Let X be a random variable distributed as Pareto type II having SF $\overline{F}(t) = (1+t)^{-\beta_1}$, for $t \geq 0$ and $\beta_1 > 0$. Let Y be Pareto type III distributed with SF $\overline{G}(t) = (1 + t^{\beta_2})^{-1}$, for $t \geq 0$ and $\beta_2 > 0$. We assume $\beta_1 = \beta_2 = \beta$. Then, from Equation (3.1), we get

$$ROC_{\overline{G},\overline{F}}(u) = \frac{u}{u + \left(1 - u^{\frac{1}{\beta}}\right)^{\beta}}, \qquad u \in [0,1],\ \beta > 0, \tag{3.9}$$

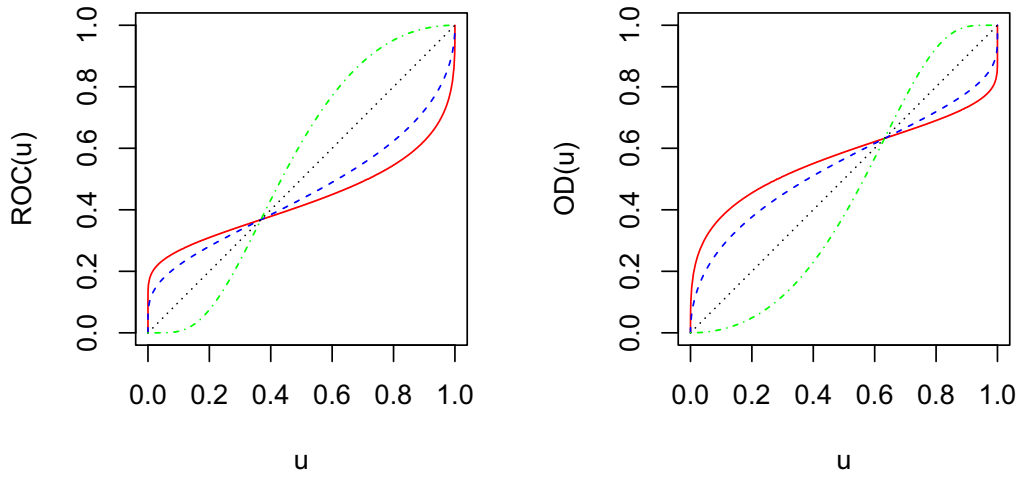

FIGURE 1 Plots of the ROC distortion curve given in Equation (3.7) (left) and of the OD distortion curve given in Equation (3.8) (right), for $u \in [0,1]$ and $\alpha = 1/3, 1/2, 1, 2$ (full, dashed, dotted and dotdashed, respectively).

that is plotted in the left-hand side of Figure 2 for $\beta = 1/4, 1/2, 1, 3/2$. For such choices of β one respectively has $AUROC_{\overline{G},\overline{F}} = 0.3280934, 0.3767747, 0.5, 0.6175911$. Moreover, from Equation (3.2), we obtain

$$OD_{G,F}(u) = \frac{\left(1-(1-u)^{\frac{1}{\beta}}\right)^{\beta}}{1-u+\left(1-(1-u)^{\frac{1}{\beta}}\right)^{\beta}}, \qquad u \in [0,1],\ \beta > 0, \tag{3.10}$$

which is plotted in the right-hand side of Figure 2 for $\beta = 1/4, 1/2, 1, 3/2$, respectively leading to $AUOD_{G,F} = 0.6719066, 0.6232253, 0.5, 0.3824089$.

A similar approach can be adopted in the context of weighted distributions, namely the ROC and OD distortion curves allow us to specify the distortion-based model between a baseline random variable and its weighted version. In particular, given a random variable X having absolutely continuous CDF F, SF $\overline{F}$, and PDF f, let us consider the weighted random variable X_w, with CDF F_w, SF $\overline{F}_w$, and PDF f_w as in Equation (2.1), according to a non-negative weight function w. Then, from Equation (3.1), the ROC distortion curve between X and X_w is

$$ROC_{\overline{F}_w,\overline{F}}(u) = \frac{1}{\mu_w}\int_{\overline{F}^{-1}(u)}^{+\infty} w(x)f(x)\mathrm{d}x = \frac{1}{\mu_w}\int_0^u w(\overline{F}^{-1}(v))\mathrm{d}v, \qquad u \in [0,1]$$

and, from Equation (3.2), the OD distortion curve between X and X_w is

$$L_w(u) = OD_{F_w,F}(u) = \frac{1}{\mu_w}\int_{-\infty}^{F^{-1}(u)} w(x)f(x)\mathrm{d}x = \frac{1}{\mu_w}\int_0^u w(F^{-1}(v))\mathrm{d}v, \qquad u \in [0,1]. \tag{3.11}$$

Remark 3.1. If X is non-negative and X_* is the corresponding length biased random variable (obtained from Equation (2.1) by setting $w(t) = t$, for all $t \geq 0$), respectively

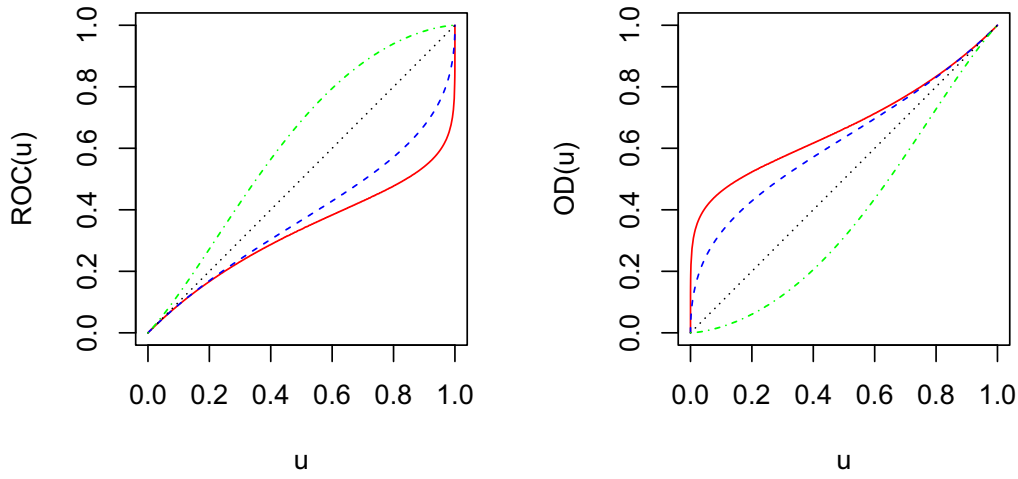

FIGURE 2 Plots of the ROC distortion curve given in Equation (3.9) (left) and of the OD distortion curve given in Equation (3.10) (right), for $u \in [0, 1]$ and $\beta = 1/4, 1/2, 1, 3/2$ (full, dashed, dotted, and dotdashed, respectively).

having CDFs F and F_*, then $\mu_w = \mu$ and, from Equation (3.11), we get

$$L(u) = OD_{F_*,F}(u) = \frac{1}{\mu}\int_0^u F^{-1}(v)\mathrm{d}v, \qquad u \in [0, 1], \tag{3.12}$$

which is the Lorenz curve of X (cf. Arnold & Sarabia, 2018, p. 25). Thus, the Lorenz curve can be interpreted as a distortion function, since it comes from the OD distortion between a non-negative random variable X and its length biased version X_*, that is $L(F) = F_*$. Hence, along this line, the distortion function given in Equation (3.11) could be interpreted as a weighted extension of the Lorenz curve, that satisfies $L_w(F) = F_w$, see also Theorem 1 in Bartoszewicz & Skolimowska (2006). For further modified Lorenz curves, we refer the reader to Section 6.4 in Arnold & Sarabia (2018) and Sordo et al. (2014).

If X' is an independent copy of a non-negative random variable X with $0 < \mathsf{E}(X) < +\infty$, then the Gini's index of X is defined as

$$\mathrm{Gini}(X) = \frac{\mathsf{E}|X - X'|}{2\mathsf{E}(X)} = 1 - 2\int_0^1 L(u)\mathrm{d}u. \tag{3.13}$$

It represents a measure of income inequality in a population related with the Lorenz curve L (cf. Arnold & Sarabia, 2018, p. 51). It is also a dispersion measure of X. We remark that in ML, particularly in the context of credit scoring models, practitioners use the so-called 'Gini coefficient' (equivalent to Somers' D statistic) to assess the quality of credit scoring models. It is derived from the AUROC curve, where the curve shows the FPR against the TPR for various score cut-off levels, that is $\mathcal{G} = 2 \times AUROC - 1$. This credit scoring 'Gini coefficient' is distinct from the Gini's dispersion index given in Equation (3.13).

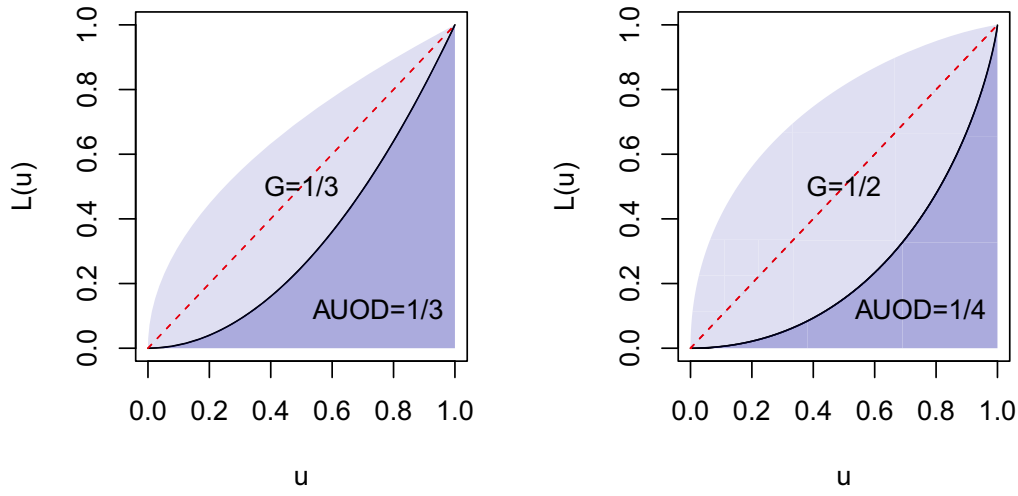

FIGURE 3 Plots of egalitarian line (red dashed line), Lorenz curve (black line) with related *AUOD* (blue area) and Gini's index denoted by G (grey area) corresponding to the uniform distribution (left) and the exponential distribution (right), respectively.

Hence, by recalling Equations (3.5) and (3.6), from Equations (3.12) and (3.13) one has

$$\mathsf{Gini}(X) = 1 - 2AUOD_{F_*,F} = 2AUROC_{\overline{F}_*,\overline{F}} - 1, \tag{3.14}$$

where F and $\overline{F}$ denote respectively the CDF and SF of X, while F_* and $\overline{F}_*$ denote respectively the CDF and SF of the corresponding length biased X_*. Note that, from Equation (3.14), we also have $\mathsf{Gini}(X) + AUOD_{F_*,F} = AUROC_{\overline{F}_*,\overline{F}}$. As example, for X uniformly distributed, in the left-hand side of Figure 3 we plot its Lorenz curve $L(u) = u^2$ for $u \in [0,1]$ (black line). The blue area corresponds to $AUOD_{F_*,F}$ equal to $1/3$, which from Equation (3.14) leads to $\mathsf{Gini}(X) = 1/3$ (grey area), and thus $AUROC_{\overline{F}_*,\overline{F}} = 2/3$ (grey+blue areas). Analogously, for X exponentially distributed, in the right-hand side of Figure 3 we plot its Lorenz curve $L(u) = u + (1-u)\ln(1-u)$ for $u \in [0,1]$ (black line). The blue area corresponds to $AUOD_{F_*,F}$ equal to $1/4$. From Equation (3.14) we obtain $\mathsf{Gini}(X) = 1/2$ (grey area) and so $AUROC_{\overline{F}_*,\overline{F}} = 3/4$ (grey+blue areas).

All distortion functions are CDFs with support in $[0,1]$. Indeed, we point out that the ROC and OD distortion curves defined respectively in Equations (3.1) and (3.2) can be interpreted as CDFs of suitable transformations of the baseline random variables. In particular, from Equation (3.1), we remark that $ROC_{\overline{G},\overline{F}}$ is the CDF of the relative random variable $S = \overline{F}(Y)$, while $ROC_{\overline{F},\overline{G}}$ is the CDF of the relative random variable $T = \overline{G}(X)$. Therefore, from Equation (3.6), it follows $\mathsf{E}(S) = AUOD_{G,F}$ and $\mathsf{E}(T) = AUOD_{F,G}$. If $X =_{st} Y$, then S and T have standard uniform distributions.

Similar considerations can be applied to the OD distortion curve. Indeed, from Equation (3.2), the relative random variable $V = F(Y)$ has CDF $OD_{G,F}$, while the relative random variable $W = G(X)$ has CDF $OD_{F,G}$. Hence, from Equation (3.5), one has $\mathsf{E}(V) = AUROC_{\overline{G},\overline{F}} = 1 - \mathsf{E}(S)$ and $\mathsf{E}(W) = AUROC_{\overline{F},\overline{G}} = 1 - \mathsf{E}(T)$. If $X =_{st} Y$, then V and W follow a standard uniform distribution. If X has a standard uniform distribution, then $V = Y$, that is $OD_{G,F}(u) = G(u)$ for all $u \in [0,1]$. Conversely, if Y has a standard uniform distribution, then $W = X$, that is $OD_{F,G}(u) = F(u)$ for all $u \in [0,1]$. Note that from Equation (3.12) the Lorenz curve L of X is the CDF of $F(X_*)$.

TABLE 4 CDFs, SFs, quantile functions and means of relative random variables having support $[0, 1]$.

Random variable	$V = F(Y)$	$S = \overline{F}(Y)$	$W = G(X)$	$T = \overline{G}(X)$
CDF	$OD_{G,F}(u)$	$ROC_{\overline{G},\overline{F}}(u)$	$OD_{F,G}(u)$	$ROC_{\overline{F},\overline{G}}(u)$
SF	$ROC_{\overline{G},\overline{F}}(1-u)$	$OD_{G,F}(1-u)$	$ROC_{\overline{F},\overline{G}}(1-u)$	$OD_{F,G}(1-u)$
Quantile function	$OD_{F,G}(u)$	$ROC_{\overline{F},\overline{G}}(u)$	$OD_{G,F}(u)$	$ROC_{\overline{G},\overline{F}}(u)$
Mean	$AUROC_{\overline{G},\overline{F}}$	$AUOD_{G,F}$	$AUROC_{\overline{F},\overline{G}}$	$AUOD_{F,G}$

Note: Here, X and Y have CDFs F and G and SFs $\overline{F}$ and $\overline{G}$, respectively.

Table 4 contains such mentioned relative random variables with their CDFs, SFs, quantile functions and means, respectively. A straightforward result is stated below, its proof is omitted.

Proposition 3.2. *With the notation of Table 4 one has* $\mathsf{E}(T) = \mathsf{E}(V)$ *and* $\mathsf{E}(S) = \mathsf{E}(W)$.

The Gini's index given in Equation (3.13) can be connected with the AUOD in a different way from Equation (3.14), as the next proposition shows.

Proposition 3.3. *Let X be a non-negative random variable with absolutely continuous CDF F and finite non-zero mean. Then*

$$\mathsf{Gini}(X) = AUOD_{F,F^e}$$

where F^e denotes the CDF of the equilibrium random variable X^e.

Proof. Let us denote by $\overline{F}^e$ the SF of X^e. From Equation (3.13), by using Remark 4.1 in Di Crescenzo & Longobardi (2009), one has

$$\mathsf{Gini}(X) = \mathsf{E}(\overline{F}^e(X))$$

and the thesis follows by assuming $Y =_{st} X^e$ in the last column of Table 4. ∎

4 | STOCHASTIC COMPARISONS AND AGING PROPERTIES

In this section, we characterize several stochastic comparisons and aging properties by taking into account the interpretation of ROC and OD distortion curves between X and Y as CDFs of the relative random variables specified in Table 4. Note that we state such propositions by considering V and W, even if they can be also exhibited by using the relative random variables S and T (cf. Table 4).

The first result involves the usual stochastic order.

Proposition 4.1. *Let X and Y be random variables with CDFs F and G, SFs $\overline{F}$ and $\overline{G}$, respectively. Let us denote with $V = F(Y)$ and $W = G(X)$ and let U be a random variable with a standard uniform distribution. The following statements are equivalent*:

(i) $X \leq_{st} Y$.
(ii) $U \leq_{st} V$.
(iii) $W \leq_{st} U$.

(iv) $OD_{G,F}(u) \leq u$ *for all* $u \in [0,1]$.
(v) $ROC_{\overline{G},\overline{F}}(u) \geq u$ *for all* $u \in [0,1]$.

Proof. Under the stated hypothesis, from Table 4, V has CDF $OD_{G,F}$. It follows $F(t) \geq G(t)$ for all t iff $OD_{G,F}(u) \leq u$ for all $u \in [0,1]$, that is $U \leq_{st} V$. This proves that (i), (ii), and (iv) are equivalent. The equivalences between (i), (iii), and (v) can be obtained in a similar way. ∎

As a corollary we obtain a well-known result.

Corollary 4.1. *If* X *and* Y *are independent and such that* $X \leq_{st} Y$, *then* $\Pr(Y > X) \geq 1/2$.

Along the same line we now characterize the hazard rate order.

Proposition 4.2. *Let* X *and* Y *be random variables with CDFs* F *and* G, *SFs* $\overline{F}$ *and* $\overline{G}$, *respectively. Let us denote with* $V = F(Y)$ *and* $W = G(X)$ *and let* U *be a random variable with a standard uniform distribution. The following statements are equivalent*:

(i) $X \leq_{hr} Y$.
(ii) $U \leq_{hr} V$.
(iii) $W \leq_{hr} U$.
(iv) $\frac{1-OD_{G,F}(u)}{1-u}$ *increasing for* $u \in [0,1]$.
(v) $\frac{ROC_{\overline{G},\overline{F}}(u)}{u}$ *decreasing for* $u \in [0,1]$.

Proof. Under the stated assumptions, from Table 4, V has SF $1 - OD_{G,F}$. Then

$$\frac{\overline{G}(t)}{\overline{F}(t)} \quad \text{increasing in } t \text{ iff} \quad \frac{1 - OD_{G,F}(u)}{1-u} \quad \text{increasing in } u,$$

where $1-u$ represents the SF of U, that proves the equivalences between (i), (ii), and (iv). The statements (iv) and (v) are equivalent due to Equation (3.4). The result is proved since the equivalence between (i) and (iii) can be obtained from Table 4 by taking into account that W has SF $ROC_{\overline{F},\overline{G}}(1-u)$ for $u \in [0,1]$. ∎

An analogous result for the reversed hazard rate order is shown below.

Proposition 4.3. *Let* X *and* Y *be random variables with CDFs* F *and* G, *SFs* $\overline{F}$ *and* $\overline{G}$, *respectively. Let us denote with* $V = F(Y)$ *and* $W = G(X)$ *and let* U *be a random variable with a standard uniform distribution. The following statements are equivalent*:

(i) $X \leq_{rhr} Y$.
(ii) $U \leq_{rhr} V$.
(iii) $W \leq_{rhr} U$.
(iv) $\frac{OD_{G,F}(u)}{u}$ *increasing for* $u \in [0,1]$.
(v) $\frac{1-ROC_{\overline{G},\overline{F}}(u)}{1-u}$ *decreasing for* $u \in [0,1]$.

Proof. Under the stated hypothesis, since from Table 4 V has CDF $OD_{G,F}$, then

$$\frac{G(t)}{F(t)} \quad \text{increasing in } t \text{ iff} \quad \frac{OD_{G,F}(u)}{u} \quad \text{increasing in } u$$

which proves the equivalences between (i), (ii) and (iv). The statements (iv) and (v) are equivalent due to Equation (3.4). The equivalence between (i) and (iii) follows similarly from Table 4. This completes the proof. ∎

In the following, we prove a result for the likelihood ratio order.

Proposition 4.4. *Let X and Y be random variables with absolutely continuous CDFs F and G, SFs $\overline{F}$ and $\overline{G}$ and PDFs f and g, respectively. Let us denote with $V = F(Y)$ and $W = G(X)$ and let U be a random variable with a standard uniform distribution. The following statements are equivalent*:

(i) $X \leq_{lr} Y$.
(ii) $U \leq_{lr} V$.
(iii) $W \leq_{lr} U$.
(iv) $OD_{G,F}(u)$ *convex for* $u \in [0,1]$.
(v) $ROC_{\overline{G},\overline{F}}(u)$ *concave for* $u \in [0,1]$.
(vi) V *is ID*.

Proof. Under the stated assumptions, $g(t)/f(t)$ is increasing in t iff the PDF of V

$$\frac{\mathrm{d}}{\mathrm{d}u} OD_{G,F}(u) = \frac{g(F^{-1}(u))}{f(F^{-1}(u))}, \qquad u \in [0,1],$$

is increasing in u, i.e., the $OD_{G,F}$ is convex in u, and thus $ROC_{\overline{G},\overline{F}}(u)$ is concave in u due to Equation (3.4). This proves the equivalences between (i), (ii), (iv), (v), and (vi). The equivalence between (i) and (iii) follows in a similar way from Table 4. ∎

Note that, if $G = q(F)$ for a given distortion q, then Propositions 4.1, 4.2, 4.3, and 4.4 are respectively related with the statements (i)-(iv) of Proposition 2 in Navarro et al. (2021).

We now provide conditions leading to IHR or DRHR property for $V = F(Y)$, respectively. Similar aging results can be proved for the other relative random variables given in Table 4.

Proposition 4.5. *Let X and Y be random variables with absolutely continuous CDFs F and G, SFs $\overline{F}$ and $\overline{G}$ and PDFs f and g, respectively. Anyone of the following statements implies that $V = F(Y)$ is IHR*:

(i) $X \leq_{lr} Y$.
(ii) X *is DHR and* Y *is IHR*.

Proof. For the statement (i), we use (vi) of Proposition 4.4 and Table 2. Conversely, under the assumptions given in (ii), the HR of V is increasing in $u \in [0,1]$ since

$$\lambda_V(u) = \frac{g(F^{-1}(u))}{f(F^{-1}(u))\overline{G}(F^{-1}(u))} = \frac{\lambda_Y(F^{-1}(u))}{(1-u)\lambda_X(F^{-1}(u))}, \qquad u \in [0,1],$$

where λ_X and λ_Y denote the HRs of X and Y, respectively. This concludes the proof. ∎

Proposition 4.6. *Let X and Y be random variables with absolutely continuous CDFs F and G, SFs $\overline{F}$ and $\overline{G}$ and PDFs f and g, respectively. Anyone of the following statements implies that $V = F(Y)$ is DRHR*:

(i) $Y \leq_{lr} X$.
(ii) *X is ID and Y is DRHR, provided that X has interval support $(-\infty, r_1)$ for $r_1 \in \mathbb{R}$, and that Y has interval support $(-\infty, r_2)$ for $r_2 \in \mathbb{R}$.*

Proof. For the statement (i), we recall Proposition 4.4. If $Y \leq_{lr} X$, then V is DD, and therefore is DRHR (cf. Table 2). Conversely, under the hypothesis specified in (ii), the RHR of V in decreasing in $u \in [0,1]$ since

$$\tau_V(u) = \frac{g(F^{-1}(u))}{f(F^{-1}(u))G(F^{-1}(u))} = \frac{\tau_Y(F^{-1}(u))}{f(F^{-1}(u))}, \qquad u \in [0,1],$$

where τ_Y denote the RHR of Y. This completes the proof. ∎

We can move a step further by providing ordering results between two different relative random variables V_1 and V_2 with CDFs defined by OD distortion curves between their respective baseline distributions. Similar propositions can be stated for the other relative random variables given in Table 4.

The first result contains conditions for the usual stochastic order.

Proposition 4.7. *Let X_1, Y_1, X_2 and Y_2 be random variables. Let us assume that $V_1 = F_1(Y_1)$ and $V_2 = F_2(Y_2)$, where F_1 and F_2 are the CDFs of X_1 and X_2, respectively. If $X_2 \leq_{st} X_1$ and $Y_1 \leq_{st} Y_2$, then $V_1 \leq_{st} V_2$.*

Proof. Since $X_2 \leq_{st} X_1$, then $F_2^{-1}(u) \leq F_1^{-1}(u)$ for all $u \in [0,1]$. Hence, by recalling Equation (3.2), one gets

$$OD_{G_1,F_1}(u) = G_1(F_1^{-1}(u)) \geq G_2(F_1^{-1}(u)) \geq G_2(F_2^{-1}(u)) = OD_{G_2,F_2}(u),$$

for all $u \in [0,1]$. From Table 4, the result follows by taking into account that V_1 has CDF OD_{G_1,F_1} while V_2 has CDF OD_{G_2,F_2}. ∎

As a corollary of Proposition 4.7, we now assume distortion-based models between two pairs of random variables. Suitable conditions on such distortions allow us to compare the corresponding relative random variables in the usual stochastic order, by avoiding establishing assumptions on the involved baseline distributions.

Corollary 4.2. *Let X and Y be random variables with CDFs F and G, respectively. Let X^* and Y^* be random variables having distorted CDFs $F^* = q_1(F)$ and $G^* = q_2(G)$ through the distortions q_1 and q_2, respectively. Let us denote with $V = F(Y)$ and $V^* = F^*(Y^*)$. If q_1 is strictly increasing and $q_1(u) \geq u \geq q_2(u)$ for all $u \in [0,1]$, then $V \leq_{st} V^*$.*

We now provide ordering and aging conditions on different pairs of baseline random variables leading to the hazard rate order between the corresponding relative ones.

Proposition 4.8. *Let X_1, Y_1, X_2 and Y_2 be random variables. Let us assume that $V_1 = F_1(Y_1)$ and $V_2 = F_2(Y_2)$, where F_1 and F_2 are the CDFs of X_1 and X_2, respectively. If $X_2 \leq_{hr} X_1$ and $Y_1 \leq_{hr} Y_2$, with X_1 DHR and Y_1 IHR, then $V_1 \leq_{hr} V_2$.*

Proof. Let us denote the HR of any random variable Z as λ_Z. Since $X_2 \leq_{hr} X_1$, then $X_2 \leq_{st} X_1$ and therefore $F_2^{-1}(u) \leq F_1^{-1}(u)$ for all $u \in [0,1]$. Under the stated assumptions, for all $u \in [0,1]$ one has

$$(1-u)\lambda_{V_1}(u) = \frac{\lambda_{Y_1}(F_1^{-1}(u))}{\lambda_{X_1}(F_1^{-1}(u))} \geq \frac{\lambda_{Y_1}(F_2^{-1}(u))}{\lambda_{X_1}(F_2^{-1}(u))} \geq \frac{\lambda_{Y_2}(F_2^{-1}(u))}{\lambda_{X_2}(F_2^{-1}(u))} = (1-u)\lambda_{V_2}(u),$$

where in the first inequality we have used that X_1 is DHR and Y_1 is IHR, while in the second one the hypothesis $X_2 \leq_{hr} X_1$ and $Y_1 \leq_{hr} Y_2$. This concludes the proof. ∎

An analogous result can be proved for the reversed hazard rate order, as shown below.

Proposition 4.9. *Let X_1, Y_1, X_2 and Y_2 be random variables having interval supports, with $(-\infty, r)$ for $r \in \mathbb{R}$ interval support of X_2, while $(-\infty, r^*)$ for $r^* \in \mathbb{R}$ interval support of Y_2. Let us assume that $V_1 = F_1(Y_1)$ and $V_2 = F_2(Y_2)$, where F_1 and F_2 are the CDFs of X_1 and X_2, respectively. If $X_2 \leq_{rhr} X_1$ and $Y_1 \leq_{rhr} Y_2$, with X_2 IRHR and Y_2 DRHR, then $V_1 \leq_{rhr} V_2$.*

Proof. Let us denote the RHR of any random variable Z as τ_Z. Since $X_2 \leq_{rhr} X_1$, then $X_2 \leq_{st} X_1$ and therefore $F_2^{-1}(u) \leq F_1^{-1}(u)$ for all $u \in [0,1]$. Under the stated assumptions, for all $u \in [0,1]$ one has

$$\tau_{V_1}(u)u = \frac{\tau_{Y_1}(F_1^{-1}(u))}{\tau_{X_1}(F_1^{-1}(u))} \leq \frac{\tau_{Y_2}(F_1^{-1}(u))}{\tau_{X_2}(F_1^{-1}(u))} \leq \frac{\tau_{Y_2}(F_2^{-1}(u))}{\tau_{X_2}(F_2^{-1}(u))} = \tau_{V_2}(u)u,$$

where in the first inequality we have used $X_2 \leq_{rhr} X_1$ and $Y_1 \leq_{rhr} Y_2$, while in the second one the hypothesis X_2 IRHR and Y_2 DRHR have been applied. The proof is completed. ∎

5 | RELATIVE AUROC AND RELATIVE AUOD DISTORTIONS

This section is devoted to define and study new distortion functions related with the partial AUROC and the partial AUOD, by describing their connections with left-skewed and right-skewed models defined in Equations (2.3) and (2.4), respectively. Moreover, the dual version of each distortion represents the equilibrium distribution of the corresponding relative random variable given in Table 4. Stochastic comparisons and aging results are considered as well.

Let X and Y be random variables with absolutely continuous SFs $\overline{F}$ and $\overline{G}$, respectively, having an interval support with the same initial point. By recalling Equation (3.1), the partial AUROC between X and Y for a given interval (u_1, u_2) is defined by

$$pAUROC_{\overline{G},\overline{F}}(u_1,u_2) = \int_{u_1}^{u_2} ROC_{\overline{G},\overline{F}}(v)\mathrm{d}v, \qquad 0 \leq u_1 \leq u_2 \leq 1, \tag{5.1}$$

which represents the AUROC between two given thresholds u_1 and u_2. Moreover, from Equations (3.5) and (5.1), the relative AUROC between X and Y for the interval (u_1, u_2) is defined by

$$rAUROC_{\overline{G},\overline{F}}(u_1, u_2) = \frac{pAUROC_{\overline{G},\overline{F}}(u_1, u_2)}{AUROC_{\overline{G},\overline{F}}}, \qquad 0 \le u_1 \le u_2 \le 1, \tag{5.2}$$

which can be interpreted as the proportion of the AUROC being captured between two given thresholds u_1 and u_2.

Proposition 5.1. *Let X be a random variable having absolutely continuous SF $\overline{F}$. Let Y be a random variable with SF $\overline{G}$, independent from X. Then*

$$rAUROC_{\overline{G},\overline{F}}(u_1, u_2) = ROC_{\overline{F}_L,\overline{F}}(u_2) - ROC_{\overline{F}_L,\overline{F}}(u_1), \qquad 0 \le u_1 \le u_2 \le 1,$$

where $\overline{F}_L$ is the SF of the random variable X_L with PDF given in Equation (2.3).

Proof. Let f denote the PDF of X. Since X and Y are independent, from Equation (5.2) it holds

$$\begin{aligned} rAUROC_{\overline{G},\overline{F}}(u_1, u_2) &= \frac{1}{\mathsf{Pr}(Y > X)} \int_{u_1}^{u_2} \overline{G}(\overline{F}^{-1}(v))\mathrm{d}v \\ &= \frac{1}{\mathsf{Pr}(Y > X)} \int_{\overline{F}^{-1}(u_2)}^{\overline{F}^{-1}(u_1)} \overline{G}(t) f(t)\mathrm{d}t \\ &= \int_{\overline{F}^{-1}(u_2)}^{\overline{F}^{-1}(u_1)} f_L(t)\mathrm{d}t \\ &= ROC_{\overline{F}_L,\overline{F}}(u_2) - ROC_{\overline{F}_L,\overline{F}}(u_1), \end{aligned}$$

where we have used Equation (2.3) and the proof is completed. ∎

As a consequence of Proposition 5.1 and $X_L \le_{lr} X$, the following function

$$rAUROC_{\overline{G},\overline{F}}(0, u) = ROC_{\overline{F}_L,\overline{F}}(u) = \overline{F}_L(\overline{F}^{-1}(u)), \qquad u \in [0, 1], \tag{5.3}$$

is a proper convex distortion function, namely the relative AUROC distortion between X and Y. Therefore, given $\overline{F}$ and $\overline{G}$, the corresponding relative AUROC distortion allows us to switch from $\overline{F}$ to its left-skewed version $\overline{F}_L$ in which the weight in the PDF is given by $\overline{G}$.

Similar considerations arise for the OD distortion curve defined in Equation (3.2). In particular, let us denote with F and G the CDFs of X and Y, respectively. The partial AUOD between X and Y for a given interval (u_1, u_2) is defined by

$$pAUOD_{G,F}(u_1, u_2) = \int_{u_1}^{u_2} G(F^{-1}(v))\mathrm{d}v, \qquad 0 \le u_1 \le u_2 \le 1, \tag{5.4}$$

quantifying the AUOD between two given thresholds u_1 and u_2. Moreover, from Equations (3.6) and (5.4), the relative AUOD between X and Y for the interval (u_1, u_2) is defined by

$$rAUOD_{G,F}(u_1, u_2) = \frac{pAUOD_{G,F}(u_1, u_2)}{AUOD_{G,F}}, \qquad 0 \le u_1 \le u_2 \le 1, \tag{5.5}$$

which represents the proportion of the AUOD being captured between two given thresholds u_1 and u_2.

Proposition 5.2. *Let X be a random variable having absolutely continuous CDF F. Let Y be a random variable with CDF G, independent from X. One has*

$$rAUOD_{G,F}(u_1, u_2) = OD_{F_R,F}(u_2) - OD_{F_R,F}(u_1), \qquad 0 \le u_1 \le u_2 \le 1,$$

where F_R is the CDF of the random variable X_R with PDF given in Equation (2.4).

Proof. Let f denote the PDF of X. Since X and Y are independent, from Equation (5.5) it follows

$$\begin{aligned} rAUOD_{G,F}(u_1, u_2) &= \frac{1}{\Pr(Y \le X)} \int_{u_1}^{u_2} G(F^{-1}(v))\mathrm{d}v \\ &= \frac{1}{\Pr(Y \le X)} \int_{F^{-1}(u_1)}^{F^{-1}(u_2)} G(t) f(t)\mathrm{d}t \\ &= \int_{F^{-1}(u_1)}^{F^{-1}(u_2)} f_R(t)\mathrm{d}t \\ &= OD_{F_R,F}(u_2) - OD_{F_R,F}(u_1), \end{aligned}$$

where we have used Equation (2.4) and the result is obtained. ∎

From Proposition 5.2 and $X \le_{lr} X_R$, the following function

$$rAUOD_{G,F}(0, u) = OD_{F_R,F}(u) = F_R(F^{-1}(u)), \qquad u \in [0, 1], \tag{5.6}$$

is a proper convex distortion function, namely the relative AUOD distortion between X and Y. Therefore, given F and G, the corresponding relative AUOD distortion allows us to switch from F to its right-skewed version F_R in which the weight in the PDF corresponds to G.

Remark 5.1. The relative AUOD is not the dual distortion of the relative AUROC. Indeed, by taking into account Equation (3.2), from Equation (5.3) the dual relative AUROC distortion is

$$\widetilde{rAUROC}_{\overline{G},\overline{F}}(0, u) = OD_{F_L,F}(u) \ge OD_{F_R,F}(u) = rAUOD_{G,F}(0, u), \qquad u \in [0, 1],$$

while, from Equation (5.6) the dual relative AUOD distortion is

$$\widetilde{rAUOD}_{G,F}(0, u) = ROC_{\overline{F}_R,\overline{F}}(u) \ge ROC_{\overline{F}_L,\overline{F}}(u) = rAUROC_{\overline{G},\overline{F}}(0, u), \qquad u \in [0, 1],$$

where both the inequalities are due to $F_L(t) \ge F_R(t)$ for all t (since $X_L \le_{lr} X_R$).

We now show several results which allow us to interpret the dual relative AUROC distortion and the dual relative AUOD distortion as CDFs of the equilibrium version of the random variables given in Table 4, respectively.

In the first result, we prove that the dual relative AUROC distortion between X and Y can be interpreted as the CDF of the equilibrium random variable of $V = F(Y)$, where F denotes the CDF of X.

Proposition 5.3. *Let X and Y be independent random variables with CDFs F and G, SFs $\overline{F}$ and $\overline{G}$, respectively. Let $V = F(Y)$. We assume that X has an absolutely continuous distribution and X_L denotes its left-skewed version through the weight $\overline{G}$, having CDF F_L. Then, the equilibrium random variable V^e has CDF given by $OD_{F_L,F}(u)$ for $u \in [0,1]$, and thus $V^e =_{st} F(X_L)$.*

Proof. Since X and Y are independent, from Equations (2.2) and (3.4), the equilibrium PDF of V can be expressed as

$$f_V^e(u) = \frac{1 - OD_{G,F}(u)}{\mathsf{E}(V)} = \frac{ROC_{\overline{G},\overline{F}}(1-u)}{\Pr(Y > X)}, \qquad u \in [0,1]. \tag{5.7}$$

Therefore, from Equation (5.3), V^e has CDF for all $u \in [0,1]$ given by

$$\begin{aligned} F_V^e(u) &= \frac{1}{\Pr(Y > X)} \int_0^u ROC_{\overline{G},\overline{F}}(1-v)\mathrm{d}v \\ &= \frac{1}{\Pr(Y > X)} \int_{1-u}^1 ROC_{\overline{G},\overline{F}}(v)\mathrm{d}v \\ &= 1 - ROC_{\overline{F}_L,\overline{F}}(1-u) \\ &= OD_{F_L,F}(u), \end{aligned} \tag{5.8}$$

where in the last equality we have used Equation (3.3). This completes the proof. ∎

Hereafter we prove some aging results related to V^e.

Proposition 5.4. *Under the assumptions of Proposition 5.3, it holds that V^e is IHR (DHR) iff $X_L \leq_{hr} (\geq_{hr}) Y$.*

Proof. From Equations (5.7) and (5.8), the HR of V^e can be expressed as

$$\lambda_V^e(u) = \frac{1}{\Pr(Y > X)} \frac{ROC_{\overline{G},\overline{F}}(1-u)}{ROC_{\overline{F}_L,\overline{F}}(1-u)} = \frac{1}{\Pr(Y > X)} \frac{\overline{G}(F^{-1}(u))}{\overline{F}_L(F^{-1}(u))},$$

which is increasing (decreasing) in $u \in [0,1]$ iff $\overline{G}/\overline{F}_L$ is increasing (decreasing) in its argument as we aimed to prove. ∎

Proposition 5.5. *Under the assumptions of Proposition 5.3, if X is DHR and Y is IHR, then V^e is ILR.*

Proof. It holds that V^e is ILR iff V is IHR. Then, the result is obtained from (ii) of Proposition 4.5. ∎

An analogous result to Proposition 5.3 for the equilibrium version of $W = G(X)$ is stated below. The proof is omitted since it follows similarly to Proposition 5.3.

Proposition 5.6. *Let X and Y be independent random variables with CDFs F and G, SFs $\overline{F}$ and $\overline{G}$, respectively. Let $W = G(X)$. We assume that Y has an absolutely continuous distribution and Y_L denotes its left-skewed version through the weight $\overline{F}$, having CDF G_L. Then, the equilibrium random variable W^e has CDF given by $OD_{G_L,G}(u)$ for $u \in [0,1]$, and thus $W^e =_{st} G(Y_L)$.*

Propositions 5.4 and 5.5 can be also stated in a similar way for W^e.

We now prove that the likelihood ratio order between V^e and W^e is equivalent to the hazard rate order between the baseline random variables X and Y.

Proposition 5.7. *Under the assumptions of Propositions 5.3 and 5.6, if $X \leq_{hr} Y$, then $W^e \leq_{lr} V^e$.*

Proof. One has $W^e \leq_{lr} V^e$ iff $W \leq_{hr} V$ (cf. Theorem 1.C.13 in Shaked & Shanthikumar, 2007). Then, the result follows from Proposition 4.2. ∎

Next we prove that the dual relative AUOD distortion between X and Y can be interpreted as the CDF of the equilibrium random variable of $S = \overline{F}(Y)$, with $\overline{F}$ denoting the SF of X.

Proposition 5.8. *Let X and Y be independent random variables with CDF F and G, SFs $\overline{F}$ and $\overline{G}$, respectively. Let $S = \overline{F}(Y)$. We assume that X has an absolutely continuous distribution and X_R denotes its right-skewed version through the weight G, having SF $\overline{F}_R$. Then, the equilibrium random variable S^e has CDF given by $ROC_{\overline{F}_R,\overline{F}}(u)$, for $u \in [0,1]$ and thus $S^e =_{st} \overline{F}(X_R)$.*

Proof. Since X and Y are independent, from Equations (2.2) and (3.3), the equilibrium PDF of S is given by

$$f_S^e(u) = \frac{1 - ROC_{\overline{G},\overline{F}}(u)}{\mathsf{E}(S)} = \frac{OD_{G,F}(1-u)}{\Pr(Y \leq X)}, \qquad u \in [0,1]. \tag{5.9}$$

Therefore, from Equation (5.6), S^e has CDF expressed for all $u \in [0,1]$ as

$$\begin{aligned} F_S^e(u) &= \frac{1}{\Pr(Y \leq X)} \int_0^u OD_{G,F}(1-v)\mathrm{d}v \\ &= \frac{1}{\Pr(Y \leq X)} \int_{1-u}^1 OD_{G,F}(v)\mathrm{d}v \\ &= 1 - OD_{F_R,F}(1-u) \\ &= ROC_{\overline{F}_R,\overline{F}}(u), \end{aligned} \tag{5.10}$$

where in the last equality we have used Equation (3.4). This completes the proof. ∎

Proposition 5.9. *Under the assumptions of Proposition 5.8, it follows that S^e is IHR (DHR) iff $Y \leq_{rhr} (\geq_{rhr}) X_R$.*

Proof. From Equations (5.9) and (5.10), the HR of S^e can be expressed as

$$\lambda_S^e(u) = \frac{1}{\Pr(Y \leq X)} \frac{OD_{G,F}(1-u)}{OD_{F_R,F}(1-u)} = \frac{1}{\Pr(Y \leq X)} \frac{G(\overline{F}^{-1}(u))}{F_R(\overline{F}^{-1}(u))},$$

TABLE 5 CDFs, SFs, quantile functions, and means of the equilibrium version of the relative random variables with support in [0, 1] given in Table 4.

Random variable	$V^e =_{st} F(X_L)$	$S^e =_{st} \overline{F}(X_R)$	$W^e =_{st} G(Y_L)$	$T^e =_{st} \overline{G}(Y_R)$
CDF	$OD_{F_L,F}(u)$	$ROC_{\overline{F}_R,\overline{F}}(u)$	$OD_{G_L,G}(u)$	$ROC_{\overline{G}_R,\overline{G}}(u)$
SF	$ROC_{\overline{F}_L,\overline{F}}(1-u)$	$OD_{F_R,F}(1-u)$	$ROC_{\overline{G}_L,\overline{G}}(1-u)$	$OD_{G_R,G}(1-u)$
Quantile function	$OD_{F,F_L}(u)$	$ROC_{\overline{F},\overline{F}_R}(u)$	$OD_{G,G_L}(u)$	$ROC_{\overline{G},\overline{G}_R}(u)$
Mean	$AUROC_{\overline{F}_L,\overline{F}}$	$AUOD_{F_R,F}$	$AUROC_{\overline{G}_L,\overline{G}}$	$AUOD_{G_R,G}$

Note: Here, X_L (X_R) denotes the left-skewed (right-skewed) version of X with weight $\overline{G}$ (G), while Y_L (Y_R) denotes the left-skewed (right-skewed) version of Y with weight $\overline{F}$ (F).

which is increasing (decreasing) in $u \in [0,1]$ iff G/F_R is decreasing (increasing) in its argument as we aimed to prove. ∎

A similar result to Proposition 5.8 for the equilibrium version of $T = \overline{G}(X)$ can be stated as follows. The proof is omitted since it is similar to the one of Proposition 5.8.

Proposition 5.10. *Let X and Y be independent random variables with CDF F and G, SFs $\overline{F}$ and $\overline{G}$, respectively. Let $T = \overline{G}(X)$. We assume that Y has an absolutely continuous distribution and Y_R denotes its right-skewed version through the weight F, having SF $\overline{G}_R$. Then, the equilibrium random variable T^e has CDF given by $ROC_{\overline{G}_R,\overline{G}}(u)$, for $u \in [0,1]$, and thus $T^e =_{st} \overline{G}(Y_R)$.*

Proposition 5.9 can be formulated in an analogous way for T^e.

In Table 5, we collect all the mentioned equilibrium relative random variables by specifying their CDFs, SFs, quantile functions and means. Note that, in regards to the relative random variables given in Table 4, the quantile function of $V^e =_{st} F(X_L)$ does not coincide with the CDF of $W^e =_{st} G(Y_L)$, and vice-versa. Similarly for $S^e =_{st} \overline{F}(X_R)$ and $T^e =_{st} \overline{G}(Y_R)$.

Remark 5.2. It is well-known that one can iteratively apply the equilibrium transformation in Equation (2.2) to a baseline random variable to obtain a sequence of equilibrium random variables, provided that the respective means are finite. Similarly, for $n = 1, 2, \ldots$ we can also define a sequence of ROC distortion functions as

$$ROC_{\overline{F}_L^{(n)},\overline{F}}(u) = \overline{F}_L^{(n)}(\overline{F}^{-1}(u)), \qquad u \in [0,1],$$

representing the CDF of $V_n = \overline{F}(X_L^{(n)})$, with $X_L^{(n)}$ independent from X and having SF

$$\overline{F}_L^{(n)}(t) = \int_t^{+\infty} \frac{\overline{F}_L^{(n-1)}(x) f(x)}{\Pr(X_L^{(n-1)} > X)} \mathrm{d}x, \qquad t \in \mathbb{R},$$

where $\overline{F}_L^{(0)}$ denotes the initial SF of $X^{(0)} =_{st} Y$, independent from X. Clearly, if $n = 1$, then we obtain Equation (5.3). The sequence of OD distortions can be derived similarly as

$$OD_{F_R^{(n)},F}(u) = F_R^{(n)}(F^{-1}(u)), \qquad u \in [0,1],$$

representing the CDF of $V_n = F(X_R^{(n)})$, with $X_R^{(n)}$ independent from X and having CDF

$$F_R^{(n)}(t) = \int_{-\infty}^{t} \frac{F_R^{(n-1)}(x) f(x)}{\Pr(X_R^{(n-1)} \leq X)} \mathrm{d}x, \qquad t \in \mathbb{R},$$

where $F_R^{(0)}$ denotes the CDF of $X^{(0)} =_{st} Y$, independent from X. Clearly, if $n = 1$, then one gets Equation (5.6).

6 | APPLICATIONS TO REAL CASES

In this section, we illustrate our theoretical findings by means of applications to real data. In particular, using the ideas previously discussed on ROC and OD distortion functions, we evaluate the performance of four representative ML binary classification methods, such as logistic regression (LR), support vector machine (SVM), multilayer perceptron (MLP) with a single hidden layer (which is a simple neural network), and random forest (RF) ensemble algorithm. We also apply the proposed ideas to illustrate the performance of biomarkers for medical diagnosis.

Let us denote by B the binary outcome, taking values 0 or 1 for each one of the categories. Let us denote by S a scoring diagnosis variable being determined by the outcomes of a ML classifier, the measures of a biomarker or by any other measurable variable aimed at predicting the target class $B = 1$ of the binary outcome. On the other hand, let us consider the conditional random variables $X =_{st} (S|B = 0)$ and $Y =_{st} (S|B = 1)$, with CDFs F and G, SFs $\overline{F}$ and $\overline{G}$, respectively. Then, the theoretical ROC associated with scoring variable S is given by Equation (3.1). The empirical estimation of the ROC can be calculated from the scores provided by S evaluated on sample data, that is

$$\hat{R}_{n,m}(u) = 1 - \hat{G}_n(\hat{F}_m^{-1}(1-u)), \qquad u \in [0,1],$$

where $\hat{F}_m$ and $\hat{G}_n$ are the empirical CDFs, estimating respectively F and G, with n and m denoting the respective sample sizes for the categories $B = 1$ and $B = 0$ in the sample data. The empirical estimation is the default facility implemented in the standard pROC from the statistical software R, see Robin et al. (2011).

An alternative approach is concerned with a kernel-based estimation which uses a smooth kernel PDF estimation of class $B = 0$ and $B = 1$ observations to calculate the ROC curve, see Zou et al. (1997). The `smooth.roc` function of the pROC R package (cf. Robin et al., 2011) provides a simple to use interface for applying this approach.

We now propose two semiparametric estimation approaches related with two different proportional models which rely on the theoretical ideas about the ROC distortion.

First, we consider the PHR Cox model (see Cox, 1972), described by $\overline{G}_\theta(x) = (\overline{F}(x))^\theta$ for $\theta > 0$, where $\overline{F}$ and $\overline{G}_\theta$ are the SFs of X and Y, respectively. In this case, one has

$$ROC_{\overline{G}_\theta, \overline{F}}(u) = \overline{G}_\theta(\overline{F}^{-1}(u)) = u^\theta, \qquad u \in [0,1], \tag{6.1}$$

which is a concave distortion function for $\theta \in (0,1)$ and a convex one for $\theta > 1$. From Equations (3.5) and (6.1), it follows that $AUROC = 1/(1+\theta)$, from which a semiparametric estimation of

the ROC can be calculated as

$$\hat{R}_{n,m}(u;\hat{\theta}_{PHR}) = u^{\hat{\theta}_{PHR}}, \qquad u \in [0,1], \tag{6.2}$$

where $\hat{\theta}_{PHR} = -1 + 1/MW$ is the estimated value of θ, with MW denoting the Mann-Whitney non-parametric estimation of the AUROC.

Another option concerns with the Marshall–Olkin parametric relationships between the respective CDFs F and G_θ, SFs $\overline{F}$ and $\overline{G}_\theta$ of X and Y, given by

$$G_\theta(x) = \frac{F(x)}{\theta + (1-\theta)F(x)}, \quad \overline{G}_\theta(x) = \frac{\theta\overline{F}(x)}{1 + (\theta-1)\overline{F}(x)}, \quad \theta > 0, \tag{6.3}$$

see Marshall & Olkin (1997). From Equation (6.3) it follows the PO model, which is described by the odds ratio proportionality

$$\frac{\overline{G}_\theta(x)}{G_\theta(x)} = \theta\frac{\overline{F}(x)}{F(x)}, \quad \theta > 0,$$

see, for instance, Sankaran & Jayakumar (2008). Therefore, under the PO model, one has

$$ROC_{\overline{G}_\theta,\overline{F}}(u) = \overline{G}_\theta(\overline{F}^{-1}(u)) = \frac{\theta u}{1 + (\theta-1)u}, \qquad u \in [0,1], \tag{6.4}$$

which is a concave distortion when $\theta > 1$ and a convex one for $\theta \in (0,1)$. From Equations (3.5) and (6.4), it turns out that

$$AUROC_{\overline{G}_\theta,\overline{F}} = \int_0^1 \frac{\theta u}{1 + (\theta-1)u}\mathrm{d}u = \frac{\theta^2 - \theta - \theta\log\theta}{(\theta-1)^2}, \tag{6.5}$$

which allows us to propose the following semiparametric estimation of the ROC

$$\hat{R}_{n,m}(u;\hat{\theta}_{MO}) = \frac{\hat{\theta}_{MO}\cdot u}{1 + (\hat{\theta}_{MO}-1)u}, \qquad u \in [0,1], \tag{6.6}$$

where $\hat{\theta}_{MO}$ is the estimated value of θ obtained by solving the following equation

$$\frac{\theta^2 - \theta - \theta\log\theta}{(\theta-1)^2} = MW.$$

The potential of resorting to parametric families of ROC distortions is that the ROC can be expressed analytically. As a result, we could easily obtain a closed expression for the relative AUROC distortion defined in Equation (5.3). In particular, for the PHR family we get

$$rAUROC_{\overline{G}_\theta,\overline{F}}(0,u) = (\theta+1)\int_0^u v^\theta dv = u^{\theta+1}, \qquad u \in [0,1], \tag{6.7}$$

that can be estimated by plugging the estimator $\hat{\theta}_{PHR}$ from Equation (6.2) in Equation (6.7). Similarly, for the Marshall–Olkin family in Equation (6.3) the partial AUROC can be calculated from

TABLE 6 Input variables collected and measured in the churn dataset.

Variable	Description
VMailMessage	Number of voice mail messages
DayMins	Service use during the day (in minutes)
DayCalls	Total number of calls during the day
EveMins	Service use during the evening (in minutes)
EveCalls	Total number of calls during the evening
NightMins	Service use during the night (in minutes)
NightCalls	Total number of calls during the night
IntlMins	Service use for international calls (in minutes)
IntlCalls	Total number of international calls
CustServCalls	Number of calls to the customer service

Equation (6.4) for $\theta \neq 1$ as follows

$$pAUROC_{\overline{G}_\theta,\overline{F}}(0,u) = \int_0^u \frac{\theta v}{1+(\theta-1)v}\mathrm{d}v = \frac{\theta}{\theta-1}\left[u - \frac{\log(1+(\theta-1)u)}{\theta-1}\right], \qquad u \in [0,1],$$

and, by taking into account Equation (6.5), the parametric expression for the relative AUROC is given by

$$rAUROC_{\overline{G}_\theta,\overline{F}}(0,u) = \frac{(\theta-1)u - \log(1+(\theta-1)u)}{\theta - 1 - \log\theta}, \qquad u \in [0,1]. \tag{6.8}$$

In this case, a natural estimator would be obtained by inserting the estimator $\hat{\theta}_{MO}$, which results from Equation (6.6), in Equation (6.8).

6.1 | Assessment of the performance of ML classifiers

We now apply the previously mentioned semiparametric approaches to assess the performance of LR, SVM, MLP and RF classifiers to predict customer churn in a telecommunications company. This dataset is a classical one used in the literature (see Larose & Larose, 2014). It consists of a sample of 3333 customers for whom several input business variables have been collected along with a binary outcome taking the value $B = 1$ if the customer has left the company (churn event) and the value $B = 0$ otherwise. The churn rate for this business case is 14.5%. A description of the input variables used to train the classifiers is provided in Table 6.

The ML classifiers are trained on a 2/3 of the data and evaluated on the remaining 1/3 testing sample data by using ROC curves. The empirical CDF, an approach based on the kernel estimation of the PDF, as well as the semiparametric approaches given in Equations (6.2) and (6.6) are used to obtain the ROC curves on the test sample dataset. The resulting curves are depicted in Figure 4. It can be noted that RF outperforms the other methods with LR giving the poorer outcomes. Unlike the empirical approaches from CDF or PDF kernel estimations, the semiparametric ones lead to concave curves which exhibit some differences in their shape. This is an

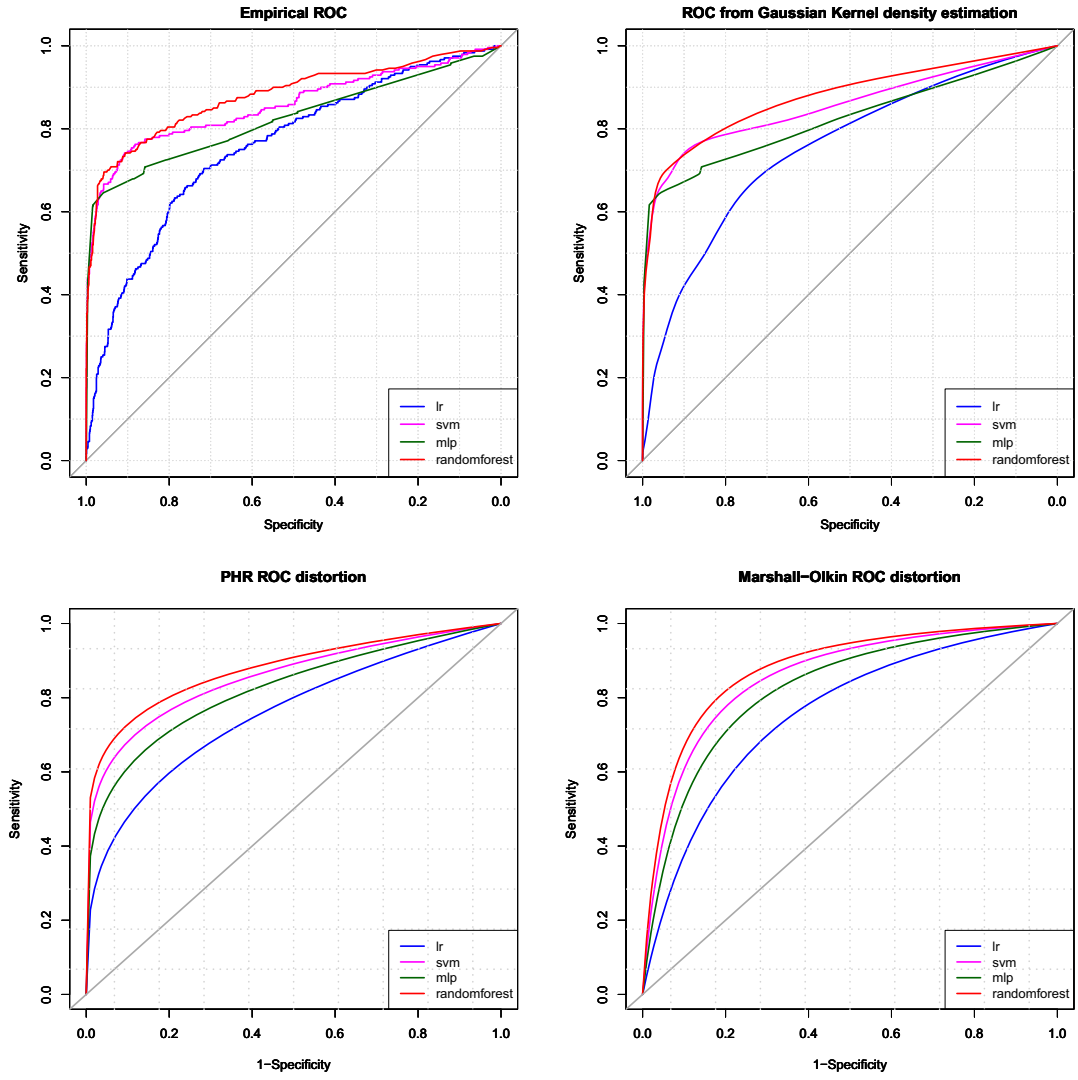

FIGURE 4 Plots for ROC comparisons applied to the assessment of LR, SMV, MLP and RF classifiers for churn prediction: Empirical ROC (top left), ROC from Gaussian kernel PDF estimation (top right), PHR ROC distortion (bottom left), and Marshall–Olkin ROC distortion (bottom right).

expected observation since these models lead to concave or convex curves and they would not give other shapes as those resulting from the empirical estimations. In this case, the concavity of the semiparametric estimation of the ROC reveals the likelihood ratio order between the class conditional scoring variables X and Y for all the ML classifiers, that is $X \leq_{lr} Y$, as stated in Proposition 4.4.

An implication of the ROC parametric based approach is that it allows us to quantify and compare the detection capacity (measured by the TPR) of ML classifiers in a straightforward manner because the estimation of the ROC is known in closed form (see Equations 6.2 and 6.6). The detection rates of LR, SMV, MLP and RF classifiers for given FPR thresholds of 10%, 20% and 30% are shown in Table 7. The resulting outcomes show the better performance of RF algorithm in all the cases.

TABLE 7 TPR outcomes for given values of the FPR (numbers in percentages).

	PHR ROC				Marshall–Olkin ROC			
FPR threshold	***LR***	***SVM***	***MLP***	***RF***	***LR***	***SVM***	***MLP***	***RF***
10	47.7	68.1	61.1	72.8	37.5	60.6	51.9	66.7
20	59.6	76.5	70.9	80.1	57.5	77.6	70.8	81.9
30	67.9	81.8	77.3	84.7	69.8	85.6	80.6	88.5

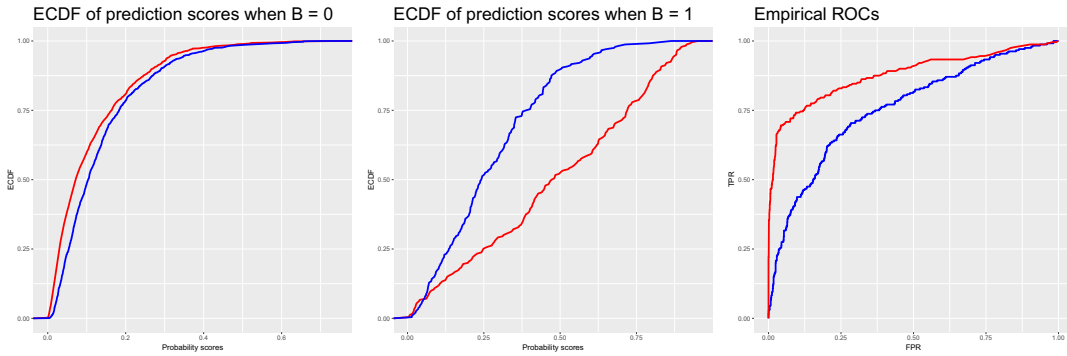

FIGURE 5 Empirical CDF comparisons of probability scores for $B = 0$ (left) and $B = 1$ (middle) as well the corresponding empirical ROC comparison of LR (blue) and RF (red) churn predictors (right).

Additionally, the theoretical findings of Proposition 4.7 are validated empirically by the plots depicted in Figure 5, which shows the empirical CDFs corresponding to the class conditional scores of LR and RF classifiers along with their ROC curves. Let us denote by X_{RF} and Y_{RF} the class conditional scoring variables of RF ensemble learning classifier and by X_{LR} and Y_{LR} the corresponding class scoring variables for LR classifier. The stochastic dominance between the empirical CDFs (see the left and middle plots in Figure 5) highlights that $X_{RF} \leq_{st} X_{LR}$ and $Y_{LR} \leq_{st} Y_{RF}$ hold. Thus, taking into account Proposition 4.7 together with the relationship given by expression (3.4), we would conclude that the ROC curve of RF ensemble learning algorithm should dominate the ROC curve resulting from LR, as pointed out by the empirical ROCs in the right plot of Figure 5.

Finally, the estimated relative AUROC curves are displayed in Figure 6 for both parametric families, the PHR and Marshall–Olkin. They provide a visual assessment of the performance of each classifier relative to its own AUROC. The visual inspection shows once again the superiority of RF algorithm for churn prediction.

6.2 | Assessment of biomarkers in medical studies

The second application is concerned with the assessment of biomarkers for predicting the clinical outcome following aneurysmal subarachnoid hemorrhage (aSAH). The experimental data come from a clinical assay comprising blood samples of 113 patients, collected 48 hours after aSAH (cf. Turck et al., 2010), from whom different clinical scores and biomarker measures were collected. After six months of the aSAH, these patients were classified in accordance to the

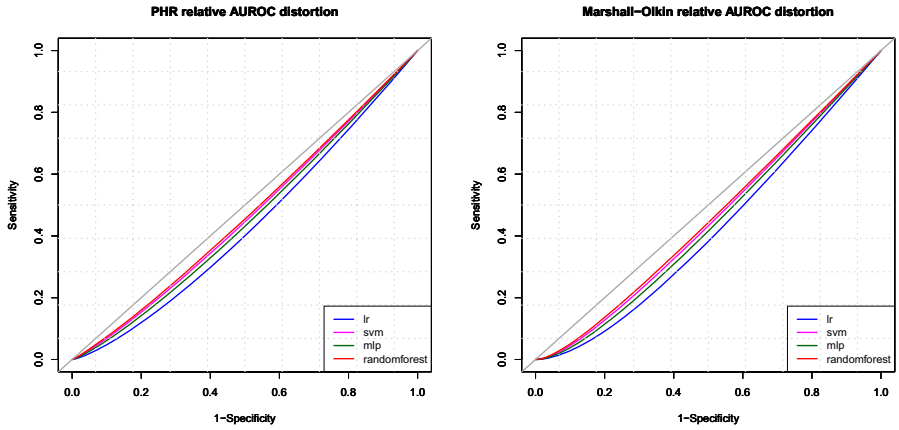

FIGURE 6 Plots of the relative AUROC distortions for PHR (left) and Marshall–Olkin (right) families.

standard neurological Glasgow outcome scale and then categorized as poor status (72 individuals) or good status (41 individuals). The dataset is available at the pROC R package in Robin et al. (2011).

Our goal is to illustrate how the predictive performance of the biomarkers $s100\beta$ protein and nucleotide diphosphate kinase A (*ndka*) can be assessed by means of the proposed PHR and Marshall–Olkin semiparametric ROC distortion-based approaches. The measures of both biomarkers are taken as the scoring values for outcome prediction. Their performance is assessed by means of ROC and relative AUROC curves estimated from the expressions (6.1), (6.4), (6.7) and (6.8) by means of the aforementioned plug-in estimations. The resulting curves are shown in Figure 7 which shows an overall better performance of $s100\beta$ biomarker.

Due to the closed analytical form of these parametric ROC distortions, we can easily derive specific performance indicators from ROC and relative AUROC curves. The outcomes of two key performance indicators are shown in Table 8. In this table, we summarize the sensitivity (that is TPR) and relative AUROC values resulting from given thresholds of 10%, 20% and 30% in the FPR. The results show that $s100\beta$ sensitivity is greater than *ndka* sensitivity with an average difference around 18.5%. The 19.7% highest performance difference corresponds to the PHR model at 10% FPR, while the lowest difference of 15% is observed for the same FPR in the Marshall–Olkin model. On the other hand, the relative AUROC outcomes also inform about the superiority of $s100\beta$ biomarker for both the PHR and Marshall–Olkin models.

7 | SUMMARY AND CONCLUDING REMARKS

In this paper, we study the ROC distortion curve. Analogously, the OD curve has been examined as a distortion too. This fact allows us to explicate the distortion function which connects two SFs (or, respectively, CDFs) when distortion-based model assumptions are unknown. Moreover, it is also helpful to interpret the Lorenz curve as a distortion function which results from the OD distortion between a non-negative random variable and its length biased version. Connections of Gini's index with AUROC and AUOD are obtained too. Several stochastic comparisons and

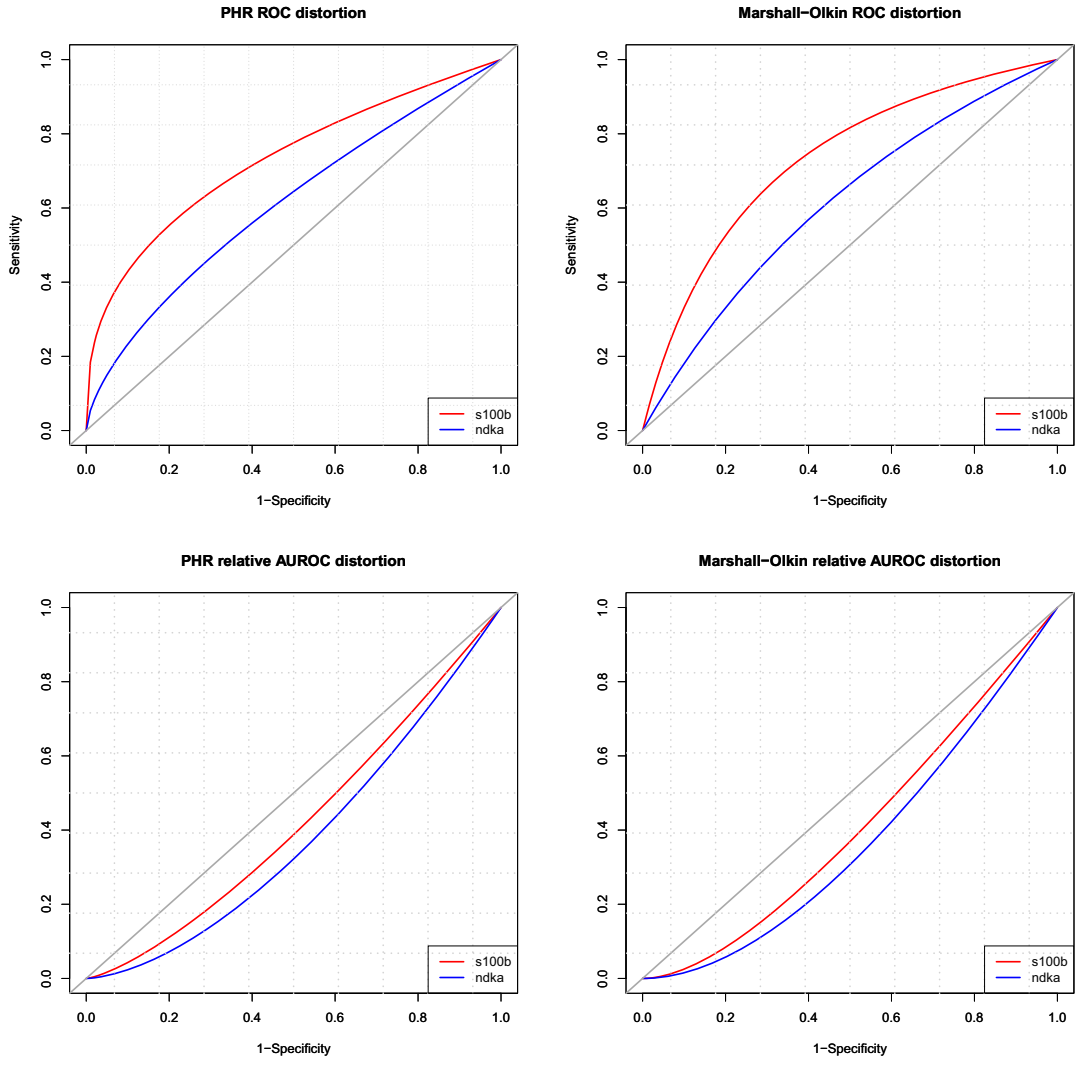

FIGURE 7 Plots of ROC and relative AUROC for the comparative assessment of *s*100*β* and *ndka* biomarkers: PHR and Marshall–Olkin ROC distortions (top), PHR and Marshall–Olkin relative AUROC distortions (bottom).

TABLE 8 Performance outcomes about sensitivity and relative AUROC indicators for biomarkers *s*100*β* and *ndka* (numbers in percentages).

	PHR ROC		PHR rel. AUROC		Marshall–Olkin ROC		Marshall–Olkin rel. AUROC	
FPR	***s*100*β***	***ndka***	***s*100*β***	***ndka***	***s*100*β***	***ndka***	***s*100*β***	***ndka***
10	42.9	23.2	4.3	2.3	33.0	18.0	2.5	1.5
20	55.4	36.0	11.1	7.2	52.6	33.1	8.4	5.7
30	64.3	46.6	19.3	14.0	65.5	45.9	16.6	12.2

aging results have been provided as well, by taking into account the interpretation of ROC and OD distortions as CDFs of suitable relative random variables.

Another relevant tool in data analysis is the partial AUROC. A relative version of the partial AUROC distortion has been studied, showing that it can be put as the ROC distortion between the original random variable and a specific left-skewed version of it determined by the distorted variable involved in the ROC. Similarly, its counterpart relative AUOD distortion can be described by the OD distortion between the original variable and a right-skewed version of it determined by the initial OD distortion. Both findings have close connections with recent skewing mechanisms defined in Navarro & Arevalillo (2023) and with the concept of equilibrium distribution. On the other hand, it also provides the theoretical basis to re-interpret partial AUROC and partial AUOD curves.

The applied part of the work is concerned with the use of the ROC distortion-based approach to two real cases that serve to illustrate our theoretical proposal. The first one shows how it works in the assessment of some well-known ML classifiers used to predict customer churn in a Telco company, whereas the second one illustrates its usefulness in the evaluation of the predictive performance of biomarkers in clinical studies. In both cases, PHR and Marshall–Olkin families of distortions have been used as parametric models that allow us to estimate the ROC and the relative AUROC in a semiparametric manner. Due to the analytical tractability of the parametric ROC and relative AUROC distortions, key performance indicators can be easily computed from them, so we advocate their use as alternatives to standard empirical approaches commonly used in ML ROC-based analysis. The outcomes resulting from the proposed distortion-based analysis have provided revealing insights which shed light on its potential and usefulness for the statistical practice.

Finally, this work opens the road for future research. Firstly, further stochastic comparisons and aging results could be investigated along the same line of the ones proved in Section 4. Moreover, the iterative scheme of Remark 5.2 suggests a study of the asymptotic behavior of sequences of equilibrium variables (cf. Table 5). We figure out that some of the ideas on iterated Lorenz curves given in Ignatov & Yordanov (2024) may help to address this issue. Another proposal for research advances is related with the development of data analysis tools that may enhance the relevance of our theoretical findings in the data science statistical practice. The ROC curves can be also helpful to perform fit tests for the stated semiparametric models. Further real examples can be considered in engineering and credit risk analysis. Alternative estimation methods based on ROC curves, such as Bayesian or non-parametric approaches, can be proposed as well.

ACKNOWLEDGEMENTS

M.C. is member of the Gruppo Nazionale Calcolo Scientifico-Istituto Nazionale di Alta Matematica (GNCS-INdAM). This work is partially funded by the "European Union—Next Generation EU" through MUR-PRIN 2022, project 2022XZSAFN "Anomalous Phenomena on Regular and Irregular Domains: Approximating Complexity for the Applied Sciences", and MUR-PRIN 2022 PNRR, project P2022XSF5H "Stochastic Models in Biomathematics and Applications". M.C. expresses his warmest thanks to the Departamento de Estadística e Investigación Operativa of Universidad de Murcia for the hospitality during a two-month visit carried out in 2024. J.N. thanks the support of Ministerio de Ciencia e Innovación of Spain in the project PID2022-137396NBI00, funded by MICIU/AEI/10.13039/501100011033 and by "ERDF A way of making Europe". Open access publishing facilitated by Università degli Studi di Salerno, as part of the Wiley - CRUI-CARE agreement.

CONFLICT OF INTEREST STATEMENT

None of the authors have a conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Marco Capaldo https://orcid.org/0000-0002-0255-8935
Jorge M. Arevalillo https://orcid.org/0000-0003-1944-3699
Jorge Navarro https://orcid.org/0000-0003-2822-915X

REFERENCES

Arnold, B. C., & Sarabia, J. M. (2018). *Majorization and the Lorenz order with applications in applied mathematics and economics*. In *Statistics for social and Behavioral sciences*. Springer.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Bartoszewicz, J., & Skolimowska, M. (2006). Preservation of classes of life distributions and stochastic orders under weighting. *Statistics & Probability Letters*, *76*, 587–596.

Belzunce, F., Martínez-Riquelme, C., & Mulero, J. (2016). *An introduction to stochastic orders*. Elsevier.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Burkschat, M., & Navarro, J. (2018). Stochastic comparisons of systems based on sequential order statistics via properties of distorted distributions. *Probability in the Engineering and Informational Sciences*, *32*, 246–274.

Calì, C., & Longobardi, M. (2015). Some mathematical properties of the ROC curve and their applications. *Ricerche di Matematica*, *64*, 391–402.

Capaldo, M., Di Crescenzo, A., & Pellerey, F. (2025). Mean distances and dependence structures for lifetimes of systems with shared components. *Applied Stochastic Models in Business and Industry*, *41*, e70002.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*, 187–202.

Di Crescenzo, A., & Longobardi, M. (2009). On cumulative entropies. *Journal of Statistical Planning and Inference*, *139*, 4072–4087.

Ewanation, L., Bennell, C., Tonkin, M., & Santtila, P. (2023). Receiver operating characteristic curves in the crime linkage context: Benefits, limitations, and recommendations. *Applied Cognitive Psychology*, *37*, 1277–1289.

Gigliarano, C., Figini, S., & Muliere, P. (2014). Making classifier performance comparisons when ROC curves intersect. *Computational Statistics & Data Analysis*, *77*, 300–312.

Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. *Expert Systems with Applications*, *40*, 5635–5647.

Ignatov, Z., & Yordanov, V. (2024). On iterated Lorenz curves with applications. arXiv:2401.13183.

Jokiel-Rokita, A., & Topolnicki, R. (2020). Estimation of the ROC curve from the Lehmann family. *Computational Statistics & Data Analysis*, *142*, 106820.

Junge, M. R., & Dettori, J. R. (2018). ROC solid: Receiver operator characteristic (ROC) curves as a foundation for better diagnostic tests. *Global Spine Journal*, *8*, 424–429.

Jurdao, S., Chuvieco, E., & Arevalillo, J. M. (2012). Modelling fire ignition probability from satellite estimates of live fuel moisture content. *Fire Ecology*, *8*, 77–97.

Krzanowski, W. J., & Hand, D. (2009). *ROC curves for continuous data*. Chapman & Hall.

Lando, T., & Legramanti, S. (2025). A new class of nonparametric tests for second-order stochastic dominance based on the Lorenz P-P plot. *Scandinavian Journal of Statistics*, *52*, 480–512.

Larose, D. T., & Larose, C. T. (2014). *Discovering knowledge in data: An introduction to data mining*. John Wiley & Sons.

Ledwina, T., & Zagdański, A. (2024). ODC and ROC curves, comparison curves and stochastic dominance. *International Statistical Review*, *92*, 431–454.

Lung, F. W., & Lee, M. B. (2008). The five-item brief-symptom rating scale as a suicide ideation screening instrument for psychiatric inpatients and community residents. *BMC Psychiatry*, *8*, 53.

Marshall, A. W., & Olkin, I. (1997). A new method for adding a parameter to a family of distribution with application to the exponential and Weibull families. *Biometrika*, *84*, 641–652.

Mickes, L., Flowem, H. D., & Wixtedm, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*, 361–376.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31.

Müller, A., & Stoyan, D. (2002). *Comparison methods for stochastic models and risks*. Wiley.

Navarro, J. (2022). *Introduction to system reliability theory*. Springer.

Navarro, J., & Arevalillo, J. M. (2023). On connections between skewed, weighted and distorted distributions: Applications to model extreme value distributions. *Test*, *32*, 1307–1335.

Navarro, J., Durante, F., & Fernández-Sánchez, J. (2021). Connecting copula properties with reliability properties of coherent systems. *Applied Stochastic Models in Business and Industry*, *37*, 496–512.

Peterson, W. W., & Birdsall, T. G. (1953). The theory of signal detectability. TR No. 13, Engineering Research Institute. University of Michigan.

Pontius, R. G., & Schneider, L. C. (2001). Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems & Environment*, *85*, 239–248.

Ramos, H. M., Ollero, J., & Suárez-Llorens, A. (2019). Two sensitivity orders applied to the comparison of ROC curves. *Communications in Statistics - Theory and Methods*, *50*, 1884–1896.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Sankaran, P. G., & Jayakumar, K. (2008). On proportional odds models. *Statistical Papers*, *49*, 779–789.

Schumacher, J. M. (2018). Distortion risk measures, ROC curves, and distortion divergence. *Statistics & Risk Modeling*, *35*, 35–50.

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. In *Springer Series in Statistics*. Springer.

Siddiqi, N. (2012). *Credit risk scorecards: Developing and implementing intelligent credit scoring*. John Wiley & Sons.

Sordo, M. A., Navarro, J., & Sarabia, J. M. (2014). Distorted Lorenz curves: Models and comparisons. *Social Choice and Welfare*, *42*, 761–780.

Sordo, M. A., & Suárez-Llorens, A. (2011). Stochastic comparisons of distorted variability measures. *Insurance: Mathematics and Economics*, *49*, 11–17.

Turck, N., Vutskits, L., Sanchez-Pena, P., et al. (2010). A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine*, *36*, 107–115.

Vanneschi, L., Horn, D. M., Castelli, M., & Popovič, A. (2018). An artificial intelligence system for predicting customer default in e-commerce. *Expert Systems with Applications*, *104*, 1–21.

Weinstein, M. C., Berwick, D. M., Goldman, P. A., Murphy, J. M., & Barsky, A. J. (1989). A comparison of three psychiatric screening tests using receiver operating characteristic (ROC) analysis. *Medical Care*, *27*, 593–607.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica*, *85*, 95–115.

Zou, K. H., Hall, W. J., & Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, *16*, 2143–2156.

How to cite this article: Capaldo, M., Arevalillo, J. M., & Navarro, J. (2025). Distorted distributions and ROC curves. *Scandinavian Journal of Statistics*, 1–30. https://doi.org/10.1111/sjos.70010