


# The area between ROC curves, a non-parametric method to evaluate a biomarker for patient treatment selection

Yoann Blangero<sup>1,2</sup>  | Muriel Rabilloud<sup>1,2</sup> | Pierre Laurent-Puig<sup>3,4,5</sup> | Karine Le Malicot<sup>6</sup> | Côme Lepage<sup>6,7,8</sup> | René Ecochard<sup>1,2</sup> | Julien Taieb<sup>3,9</sup> | Fabien Subtil<sup>1,2</sup>

<sup>1</sup>Service de Biostatistique, Pôle Santé Publique, Hospices Civils de Lyon, Lyon, France

<sup>2</sup>Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, Villeurbanne, France

<sup>3</sup>Université Paris Descartes, Sorbonne Paris Cité, Paris, France

<sup>4</sup>Service de génétique, Hôpital Européen Georges Pompidou, Paris, France

<sup>5</sup>INSERM UMR-S 1147, Paris, France

<sup>6</sup>Fédération Francophone de Cancérologie Digestive, Dijon, France

<sup>7</sup>Hépatogastroentérologie et cancérologie digestive, Centre hospitalier universitaire Dijon Bourgogne, Dijon, France

<sup>8</sup>INSERM U 866, Dijon, France

<sup>9</sup>Chirurgie digestive générale et cancérologique, Hôpital Européen Georges Pompidou, Paris, France

## Correspondence

Yoann Blangero, Service de Biostatistique, Pôle Santé Publique, Hospices Civils de Lyon, Lyon 69003, France.

Email: yoann.blangero@chu-lyon.fr

## Funding information

Fondation pour la Recherche Médicale, Grant/Award Number: ECO 20160736050; Fondation pour la Recherche Médicale, Grant/Award Number: ECO20160736050



## Reproducible Research

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## Abstract

Treatment selection markers are generally sought for when the benefit of an innovative treatment in comparison with a reference treatment is considered, and this benefit is suspected to vary according to the characteristics of the patients. Classically, such quantitative markers are detected through testing a marker-by-treatment interaction in a parametric regression model. Most alternative methods rely on modeling the risk of event occurrence in each treatment arm or the benefit of the innovative treatment over the marker values, but with assumptions that may be difficult to verify. Herein, a simple non-parametric approach is proposed to detect and assess the general capacity of a quantitative marker for treatment selection when no overall difference in efficacy could be demonstrated between two treatments in a clinical trial. This graphical method relies on the area between treatment-arm-specific receiver operating characteristic curves (ABC), which reflects the treatment selection capacity of the marker. A simulation study assessed the inference properties of the ABC estimator and compared them with other parametric and non-parametric indicators. The simulations showed that the estimate of the ABC had low bias, power comparable to parametric indicators, and that its confidence interval had a good coverage probability (better than the other non-parametric indicator in some cases). Thus, the ABC is a good alternative to parametric indicators. The ABC method was applied to data of the PETACC-8 trial that investigated FOLFOX4 versus FOLFOX4 + cetuximab in stage III colon adenocarcinoma. It enabled the detection of a treatment selection marker: the *DDR2* gene.

## KEYWORDS

clinical trial, predictive marker, quantitative marker, receiver operating characteristic curve, treatment selection

## 1 | INTRODUCTION

A major aim of precision medicine is to determine the best treatment for individual patients. It is therefore essential to identify and assess markers able to guide treatment decisions so as to avoid the occurrence of a given event (e.g., disease progression, recurrence, or death) in a given post-treatment interval. When comparing the efficacy of two treatments (an innovative vs. a reference treatment), such markers are expected to improve patient outcomes by selecting patients who would likely most benefit from the innovative treatment and avoid treating those who would not benefit from this. There is currently no consensus on the naming of such a marker; whereas Italiano (2011) and Ballman (2015) have used “predictive marker” Janes, Brown, Huang, and Pepe (2014a) used “treatment selection marker” that will be used herein.

A treatment selection marker is generally sought for when the overall risk of event occurrence is nearly the same with two different treatments; it is then expected that a subgroup of patients would get more benefit from one of the treatment than from the other. One example of treatment selection marker is the mutated KRAS gene in metastatic colorectal cancer. The presence of this mutated gene is a marker of benefit from chemotherapy alone as opposed to chemotherapy + epidermal growth factor receptor (EGFR) inhibitor; patients with tumors harboring mutated KRAS exon 2 are known to be resistant to EGFR inhibitors, whereas those with KRAS wild-type tumors do benefit from the combined treatment (De Roock et al., 2008; Di Fiore et al., 2007; Lièvre et al., 2008).

In the case of a quantitative marker, it is necessary to find a threshold value of the marker that determines the optimal treatment allocation for the patients with a marker value above or below this threshold (Blangero, Rabilloud, Ecochard, & Subtil, 2019; Janes et al., 2014a; Janes, Pepe, & Huang, 2014b; Vickers, Kattan, & Sargent, 2007). However, before defining a threshold, the first step in the assessment of a new promising quantitative marker is to quantify and test its overall performance for treatment selection. Various methods have been proposed to evaluate the overall performance of a marker for treatment selection. The classical approach consists in modeling the risk of event given the treatment options and the marker values, and then testing for a statistical interaction between these two variables, as proposed by Byar (1985), and applied in several studies (for some examples, see Skougaard, Nielsen, Jensen, Pfeiffer, and Hendel (2016), or Weidhaas et al. (2016)). One limit of this approach is that the interaction coefficient depends on the additive or multiplicative structure of the model; the interaction may be present in one type of model but not in the other, and conversely so (Byar, 1985). A marker is defined as a treatment selection marker when the difference in risk of event occurrence between the two treatment arms is inconstant over the marker values (Song & Pepe, 2004), which means that the additive scale should be used to assess treatment selection markers.

In addition, although the interaction approach is straightforward with binary markers, it is quite complex with quantitative markers because of the difficulty of verifying the adequacy of the functional form retained in modeling the interaction. One extension of the previous approach is the use of graphical tools such as “marker-by-treatment predictiveness curves” as proposed by Janes, Pepe, Bossuyt, and Barlow (2011). These graphs plot the risk of event in each treatment arm given the marker value versus the cumulative distribution function of the marker. The cumulative distribution function instead of real values enables the use of a single scale ranging between 0 and 1, allows marker-by-treatment predictiveness curve comparisons, and gives the proportions of patients who would receive each treatment according to the marker values, which is important in medical decision-making. Marker-by-treatment predictiveness curves allow visualization of the performance of a marker for treatment selection, but they rely on a good calibration of the risk modeling in each arm. Such a model often assumes a linear marker-by-treatment interaction on the linear predictor scale. Unfortunately, this assumption is not always valid and not easy to check. Moreover, the marker-by-treatment predictiveness curve is a graphical tool, but does not allow to quantify the performance of the marker for treatment selection.

Other methods assess the treatment selection capacity of a quantitative marker (Huang, Gilbert, & Janes, 2012; Zhang, Nie, Soon, & Liu, 2014) by measuring its ability to distinguish patients who would have a better outcome with the innovative treatment in comparison to the reference treatment, from patients who would have a worse outcome. However, this kind of approach needs to model the probability of having a better outcome with the innovative treatment compared with the reference treatment in each patient. Except in cross-over trials, this requires modeling using a potential outcomes framework with complex assumptions that may be very difficult to verify (Janes, Pepe, McShane, Sargent, & Heagerty, 2015b). For example, Huang et al. (2012) made the monotonicity assumption (one treatment is always at least as effective as the other one) to estimate the individual benefit, which is a strong assumption. Zhang et al. (2014) relaxed the latter assumption by assuming that the potential outcomes are independent given observed covariates. That means that the benefit of the innovative treatment for a patient may be calculated by comparing its outcome to the outcomes of the patients that had similar covariate values, but that received the reference treatment. This method assumes that the observed covariates are sufficient to explain the dependence between the two potential outcomes, which is an assumption difficult to verify.

In the present paper, a simple non-parametric method is proposed to investigate the capacity of a quantitative marker for treatment selection when the overall risk of event in each treatment arm is equal. The method relies on a special use of receiver operating characteristic (ROC) curves and provides a bounded indicator able to quantify and test the treatment selection capacity of the marker. The method is described, tested, and compared with other methods in a simulation study; it is then applied to a real dataset.

## 2 | METHODS

Throughout this article it is assumed that the marker under study (denoted  $V$ ) is measured before treatment allocation within the context of a parallel randomized controlled clinical trial with two treatment arms (innovative vs. reference) and that the outcome of interest is a binary event measured after a fixed duration of follow-up.

The binary event of interest is denoted by  $E$ , where  $E = 1$  indicates the presence of the event of interest, and  $E = 0$  its absence. Let us also denote the treatments under study by  $T$ , where  $T = 1$  indicates the innovative treatment and  $T = -1$  indicates the reference one.

Moreover, it is assumed that:

$$\rho_{(-1)} = \rho_{(1)} = \rho \quad (1)$$

where  $\rho_{(-1)} = P(E = 1|T = -1)$  and  $\rho_{(1)} = P(E = 1|T = 1)$  denote the overall risk of event in each arm, and  $\rho = P(E = 1)$  denotes the marginal risk of event in the trial. Assumption (1) means that no overall difference in efficacy could be demonstrated between the two treatment arms.

Song and Pepe (2004) proposed a mathematical definition of a treatment selection marker. A marker has no capacity for treatment selection if

$$\delta(v) = \rho_{(-1)}(v) - \rho_{(1)}(v) = \rho_{(-1)} - \rho_{(1)} \quad \forall v \quad (2)$$

where  $\rho_{(-1)}(v) = P(E = 1|T = -1, V = v)$  and  $\rho_{(1)}(v) = P(E = 1|T = 1, V = v)$  denote the risk of event in each treatment arm for a value  $v$  of the marker.

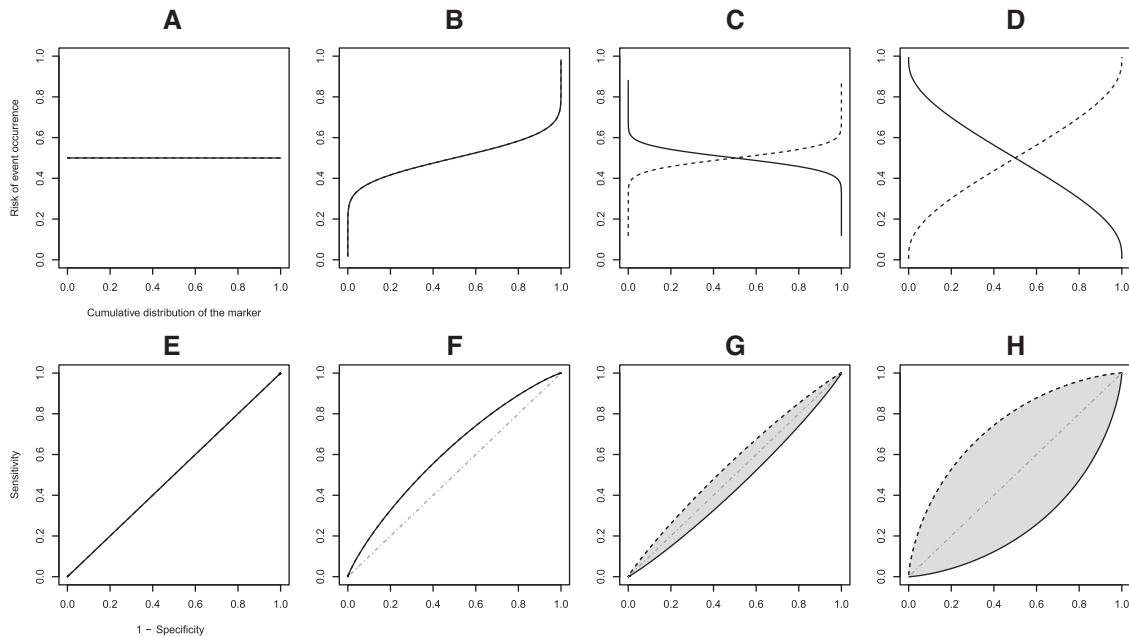
Conversely, a marker has a capacity for treatment selection when the difference in risks between the two treatment arms is dependent of the marker values. As Song and Pepe (2004) and Janes et al. (2014a) suggested, the difference in risk is the key point in treatment selection marker assessment. A marker is all the more interesting for treatment selection that the changes in risk differences are important according to the marker values.

### 2.1 | Marker-by-treatment predictiveness curves

Marker-by-treatment predictiveness curves are simple graphical tools that help in understanding the difference between a treatment selection marker and a simple prognostic marker.

In Figure 1, each panel presents two curves: one relative to the reference treatment and another relative to the innovative treatment.

- Panel A shows a case where the risk of event is independent of the marker value in each treatment arm. As the overall risk of event is the same in each treatment arm (assumption (1)), the marker-by-treatment predictiveness curves overlap; hence, the marker cannot be a treatment selection marker.
- Panel B shows a case where the risk changes with the marker value in both treatment arms. Thus, the marker may be called “prognostic marker” in each arm. However, the difference in risk between the two arms ( $\delta(V)$ ) is constant and equal to 0 in this case. Thus, there is no interaction between the treatment arm and the marker values, the marker cannot be a treatment selection marker.
- Panel C shows a case where the risk of event occurrence decreases with the marker value in the innovative arm but increases in the reference arm: the prognostic value is different between treatment arms.  $\delta(V)$  changes with the marker value: this marker is thus a treatment selection marker. In this case, the threshold of marker value that defines treatment allocation should be close to the marker value that corresponds to 50% of its cumulative distribution.



**FIGURE 1** Marker-by-treatment predictiveness curves of four markers (A, B, C, and D), and their corresponding ROC curves (E, F, G, and H). Dotted line: innovative treatment; solid line: reference treatment; shaded area: area between ROC curves

- Panel D shows another case where the risk of event occurrence decreases with the marker value in the innovative arm and increases in the reference arm, but the slopes are greater than in panel C: the prognostic value of the marker is stronger in the two treatment arms. This marker is also a treatment selection marker; furthermore, its capacity for treatment selection is greater than in panel C because of greater magnitude of changes in  $\delta(V)$ . The treatment selection capacity of a marker is all the more important that the changes in  $\delta(V)$  over the cumulative distribution function of the marker values are important.

Thus, a marker is a treatment selection marker when its prognostic ability is different between two treatment arms, which is the definition of a marker-by-treatment interaction. This is the basis for the development of the method presented hereafter.

## 2.2 | Notations and illustration of the "area between curves"

A simple and non-parametric method to estimate the prognostic ability of a marker in a single treatment arm relies on the area under the ROC curve (AUC,  $\theta$  in equations) that quantifies the ability of the marker to discriminate subjects who will experience the event in a given post-treatment interval from those who will not (Hanley & McNeil, 1982). We propose to estimate the treatment selection capacity of a marker by estimating the difference in prognostic ability between two treatment arms. This difference can be quantified by the area that separates the two treatment-arm-specific ROC curves, named "area between curves" (ABC). An alternative to estimate the prognostic ability of the marker would have been to use the area between the density curves of the marker in patients subgroups (e.g., between diseased and non-diseased patients) as proposed by Böhning, Hempfling, Schelp, and Schlattmann (1992) and Giacoletti and Heyse (2015). However, for treatment selection markers, this kind of approach would need the estimation of four density curves (diseased and non-diseased patients in each treatment arm).

A classical assumption in marker evaluation using ROC curves is that the risk of event in both arms is either monotonically increasing or monotonically decreasing over the marker values. Otherwise, the issue of improper ROC curves would arise (Metz & Pan, 1999).

When the two ROC curves do not intersect, the ABC can be measured by the difference between the two AUCs:  $\Delta_\theta = \theta_{(-1)} - \theta_{(1)}$ ,  $\theta_{(-1)}$  and  $\theta_{(1)}$  being, respectively, the AUCs of the marker for  $T = -1$  and  $T = 1$ . As both  $\theta_{(-1)}$  and  $\theta_{(1)}$  range between 0 and 1,  $\Delta_\theta$  ranges between  $-1$  and  $1$  (when  $\Delta_\theta$  is negative, the ABC is the absolute value of  $\Delta_\theta$ ).

The second row in Figure 1 presents the ROC curves that correspond to the marker-by-treatment predictiveness curves of the first row.

- Panel E shows two overlapping ROC curves on the diagonal; the marker has no prognostic ability in either arm,  $\Delta_\theta = 0$ .

- Panel F shows two overlapping ROC curves but distinct from the diagonal; the marker has the same prognostic ability in both arms but no capacity for treatment selection and  $\Delta_\theta = 0$ .
- Panel G shows two distinct ROC curves located on either side of the diagonal; the marker has a prognostic ability in both arms but the risk is increasing in the innovative arm and decreasing in the reference arm. The marker in Panel G has a capacity for treatment selection and  $\Delta_\theta = -0.11$ .
- Panel H shows two distinct ROC curves located on either side of the diagonal too. As shown by the marker-by-treatment predictiveness curves, the marker in panel H has a stronger capacity for treatment selection than the marker in panel G. This is reflected by a  $\Delta_\theta = -0.48$  further from zero than in panel G.

To summarize, the capacity of a marker for treatment selection increases with the ABC or the gap between the ROC curves; thus  $\Delta_\theta$  is different from 0. When  $\Delta_\theta = 0$  (overlapping ROC curves) the marker has no capacity for treatment selection. When  $\Delta_\theta$  is equal to  $-1$  or  $1$ , the marker is a perfect treatment selection marker; that is, the marker distinguishes perfectly patients with (or, alternatively, without) the event under the innovative treatment from those under the reference treatment (Appendix A.1 presents an illustration of the definition of a perfect marker). Furthermore,  $\Delta_\theta$  may be used to test whether a marker has a statistically significant treatment selection capacity (i.e., testing whether  $\Delta_\theta = 0$ ) and compare the capacity for treatment selection of several markers.

### 2.3 | Justification of the use of $\Delta_\theta$

The use of  $\Delta_\theta$  to quantify the treatment selection capacity of a marker is justified by its close connection with the difference in risk between the two treatment arms over the marker values. Viallon and Latouche (2011) demonstrated that the AUC in a single treatment arm could be written as a function of the predictiveness curve:

$$\theta_{(T)} = \frac{\int_{-\infty}^{+\infty} F(v)\rho_{(T)}(v) dF(v) - \frac{\rho_{(T)}^2}{2}}{\rho_{(T)}(1 - \rho_{(T)})}$$

where  $F(\cdot)$  is the cumulative distribution function of marker  $V$ . With this expression, it is easy to show that when the overall risks of event occurrence in the reference and innovative treatment arms are equal (i.e. when  $\rho_{(-1)} = \rho_{(1)} = \rho$ ),  $\Delta_\theta$  can be expressed as a function of  $\delta(V)$ :

$$\Delta_\theta = \theta_{(-1)} - \theta_{(1)} = \frac{\int_{-\infty}^{+\infty} F(v) \times \delta(v) dF(v)}{\rho(1 - \rho)}$$

$\Delta_\theta$  is greater when the variations in the risk difference  $\delta(V)$  are high on the range of marker values, hence when the marker has a greater capacity for treatment selection. According to Equation (2), it can be shown that when a marker has no capacity for treatment selection then  $\Delta_\theta = 0$ , and conversely (Appendix A.2).

### 2.4 | Connection between $\Delta_\theta$ and two other indicators

Hereafter, two indicators are presented in order to show their connection with  $\Delta_\theta$ : the total gain indicator of Janes et al. (2014b) and the  $\gamma$  indicator of Zhang, Ma, Nie, and Soon (2017).

#### 2.4.1 | The total gain

In their article, Janes et al. (2014a) proposed to evaluate the overall capacity of a marker for treatment selection using the total gain (TG) expressed as:

$$TG = \int |\delta(v) - (\rho_{(-1)} - \rho_{(1)})| dF_\delta$$

In this equation,  $F_\delta$  is the cumulative distribution function of  $\delta(V)$ .

The TG indicator measures the overall treatment selection capacity of a marker. When the marker has no treatment selection capacity, the TG equals 0, and conversely so. However, the maximum TG value depends on  $\rho_{(-1)}$  and  $\rho_{(1)}$ , and therefore the TG cannot be used to compare markers from different studies.

The TG and  $\Delta_\theta$  are two closely connected overall indicators of the treatment selection capacity of a marker. From the expressions of TG and  $\Delta_\theta$ , one may see that there is a monotone relationship between these two indicators and that the intensity of this relationship depends on the overall risk of event occurrence in each treatment arm. However, whereas the TG is based on risks,  $\Delta_\theta$  is based on ROC curves that measure the ability of the marker to separate two groups of patients. In fact,  $\Delta_\theta$  is an indicator of the ability of the marker to distinguish patients with (or, alternatively, without) the event under the innovative treatment from those under the reference treatment. Moreover,  $\Delta_\theta$  is non-parametric regarding the functional form of the interaction, and is always bounded between  $-1$  and  $1$ .

Finally, note that Janes et al. (2014a) did not propose an inference method for the TG except from bootstrap.

## 2.4.2 | The $\gamma$ concordance measure

In another article, Zhang et al. (2017) proposed a quantitative concordance measure for the assessment of the overall performance of treatment selection markers. This concordance measure is expressed as

$$\gamma = E(G_{ij})$$

where  $G_{ij} = \text{sgn}(V_i - V_j)[\delta(V_i) - \delta(V_j)]$ ,  $i$  and  $j$  are the indices of two independent patients, and  $\text{sgn}(\cdot)$  is the sign function.

As one may see, when  $\delta(V)$  is constant over the marker values then  $\gamma = 0$ , and the greater the variations in  $\delta(V)$  are, the greater  $\gamma$  is, and the greater the performance of the marker for treatment selection is.

There is a connection between this indicator and the ones described above as there are all functions of  $\delta(V)$  and that their value depends on the variations in  $\delta(V)$ .  $\gamma$  is estimated non-parametrically using pairwise comparisons of patient outcomes; since it is a U-statistic, the estimator converges to a normal distribution (Hoeffding, 1948). The variance of the estimator follows from asymptotic theory and may rely on a working model that predicts the risk of event occurrence in order to be more efficient. The variance is optimal when the working model for risk prediction includes all the covariates that impact the risk of event (see Zhang et al. (2017) for more details).

## 2.5 | Estimation and inference of $\Delta_\theta$

When estimated non-parametrically with the trapezoidal rule, the AUC estimate is asymptotically normally distributed with DeLong's variance (DeLong, DeLong, & Clarke-Pearson, 1988). As  $\Delta_\theta$  is a difference between two independent AUCs, its estimator is also asymptotically normally distributed, with variance equal to the sum of the two AUC variances. Thus, a symmetric confidence interval can be obtained using the normal approximation:

$$\left[ \hat{\Delta}_\theta \pm z_{1-\alpha/2} \times \sqrt{\text{Var}(\hat{\Delta}_\theta)} \right]$$

In this expression  $\hat{\Delta}_\theta$  denotes the estimator of  $\Delta_\theta$ , and  $z_{1-\alpha/2}$  is the  $(1 - \alpha/2)$  quantile of a standard normal distribution.

The symmetric confidence interval may indicate limits  $> 1$  or  $< -1$ , especially when  $\Delta_\theta$  is close to  $1$  or  $-1$ . To obtain asymmetric confidence limits between  $1$  and  $-1$ , these limits may be calculated on the inverse hyperbolic tangent scale of  $\Delta_\theta$  using the Delta method:

$$\left[ \text{arctanh}(\hat{\Delta}_\theta) \pm z_{1-\alpha/2} \times \frac{\sqrt{\text{Var}(\hat{\Delta}_\theta)}}{(1 + \hat{\Delta}_\theta)(1 - \hat{\Delta}_\theta)} \right]$$

The confidence interval on the inverse hyperbolic tangent scale of  $\Delta_\theta$  is then back-transformed to provide the asymmetric interval for  $\Delta_\theta$ .

The treatment selection capacity of a marker may be tested using the Wald statistic:

$$z = \frac{\hat{\Delta}_\theta}{\sqrt{\text{Var}(\hat{\Delta}_\theta)}}$$

In this formula,  $z$  follows asymptotically a standardized normal distribution when the marker has no capacity for treatment selection ( $\Delta_\theta = 0$ ).

The above estimation method is investigated hereafter in a simulation study and applied later to a real dataset.

### 3 | SIMULATION STUDY

#### 3.1 | Design

Simulation studies were designed to evaluate the performance of  $\hat{\Delta}_\theta$  to evaluate the overall capacity of a marker for treatment selection.

To evaluate the method performance, three scenarios for the risk of event occurrence were created. Scenarios 1, 2, and 3 considered that the overall risk in both  $T_{(-1)}$  and  $T_{(1)}$  was equal to 0.5, 0.25, and 0.1, respectively. Varying the overall risks is expected to affect the variance of  $\Delta_\theta$  because the number of events in each arm changes according to each arm-specific risk. Each scenario was evaluated with four theoretical values of  $\Delta_\theta$  (0.6, 0.4, 0.2, and 0.1; except for Scenario 3 where value 0.6 could not be considered), and five  $N$  sizes of 200, 400, 1,000, 1,600, and 2,000 subjects. This generated 55 different settings. Each was run 10,000 times.

As the marker is assessed in a clinical trial context, its values are supposed to have the same distribution in the two trial arms. This is defined as the randomization constraint that can be expressed as:

$$P(V \leq c|T = -1) = P(V \leq c|T = 1) \quad \forall c \quad (3)$$

In Scenario 1, the marker values were sampled from four Gaussian distributions with same variance; two distributions per arm, one for those who will experience the event and one for the others. The means and variances of these distributions were chosen so as to obtain the four theoretical  $\Delta_\theta$  values whilst fulfilling the randomization constraint given in Equation (3) (for details, see Supporting Information). The theoretical  $\Delta_\theta$  values were calculated using the analytical expression of AUC with Gaussian distributions (Pepe, 2003).

The randomization constraint (3) could not be fulfilled with four Gaussian distributions in Scenarios 2 and 3 (see Supporting Information). Hence, in these scenarios, the marker values were sampled from Gaussian distributions with same variance, except for patients without event in the innovative treatment arm; for these, the marker distribution was built from the distributions in the three other groups using the randomization constraint. A Metropolis algorithm (Gelman et al., 2013) was used to sample values from this compound marker distribution (see Supporting Information). The means of the three Gaussian distributions were chosen so as to obtain the four theoretical  $\Delta_\theta$  values. The theoretical  $\Delta_\theta$  values were calculated by numerical integration.

A first simulation study was performed to assess the coverage probability of the symmetric and asymmetric confidence intervals of  $\Delta_\theta$  for all scenarios,  $\Delta_\theta$  values, and  $N$  sizes defined previously. Additional simulations were performed with Scenario 1 and a theoretical  $\Delta_\theta = 0.95$  to assess the coverage probability of the symmetric and asymmetric confidence intervals when  $\Delta_\theta$  is close to 1.

A second simulation study was performed to estimate the mean relative bias in  $\hat{\Delta}_\theta$ , and its power for detection of treatment selection markers. The mean relative bias, the power, and the coverage probability of the asymmetric confidence interval for  $\hat{\Delta}_\theta$  were compared with those of the TG estimator (relative bias, and coverage probability of the percentile bootstrap confidence interval), denoted  $\widehat{TG}$ , the  $\gamma$  estimator (relative bias, coverage probability of confidence intervals, and power), denoted  $\hat{\gamma}$ , and the interaction coefficient estimator of a logistic regression model (power), denoted  $\hat{\beta}_3$ .

For  $\widehat{TG}$ ,  $\delta(V)$  was estimated using the predictions from a logistic regression model assuming a linear interaction between the marker value and the treatments:

$$\text{logit}[P(E = 1|T, V)] = \beta_0 + \beta_1 \times V + \beta_2 \times T + \beta_3 \times V \times T$$

The power of the interaction coefficient method was calculated using the maximum likelihood estimator of the coefficient  $\beta_3$  of this model. For the  $\gamma$  concordance measure, the estimation method provided by Zhang et al. (2017) was used based on the predictions of the following working model:

$$\text{logit}[P(E = 1|V)] = \eta_0 + \eta_1 \times V$$

This working model should lead to an optimal variance estimate as the occurrence of the event of interest only depends on the treatments (that should not be included in this working model) and on the marker under study.

The theoretical TGs and  $\gamma$ s were calculated by numerical integration in order to estimate the relative bias in  $\widehat{TG}$  and in  $\hat{\gamma}$ .

A third simulation was performed to verify the properties of the Wald test applied to  $\hat{\Delta}_\theta$  and to  $\hat{\beta}_3$  by checking the  $\alpha$ -risk value. This simulation was performed with Scenario 1,  $\Delta_\theta = 0$ , and on five  $N$  sizes of 40, 100, 200, 400, and 1,000 subjects.

**TABLE 1** Coverage probability and mean width of the 95% symmetric and asymmetric confidence intervals of  $\Delta_\theta$  (Scenario 1)

$\Delta_\theta$	$N$	Symmetric		Asymmetric	
		CP	WCI	CP	WCI
0.95	200	0.9081	0.07	0.9474	0.07
	400	0.9339	0.05	0.9519	0.05
	1,000	0.9416	0.03	0.9498	0.03
	1,600	0.9484	0.02	0.9517	0.02
	2,000	0.9441	0.02	0.9493	0.02
0.6	200	0.9439	0.24	0.9507	0.24
	400	0.9482	0.17	0.9519	0.17
	1,000	0.9471	0.11	0.9480	0.11
	1,600	0.9491	0.09	0.9500	0.09
	2,000	0.9515	0.08	0.9512	0.08
0.4	200	0.9466	0.29	0.9526	0.29
	400	0.9495	0.20	0.9526	0.20
	1,000	0.9462	0.13	0.9484	0.13
	1,600	0.9493	0.10	0.9495	0.10
	2,000	0.9503	0.09	0.9508	0.09
0.2	200	0.9489	0.31	0.9537	0.31
	400	0.9489	0.22	0.9509	0.22
	1,000	0.9461	0.14	0.9480	0.14
	1,600	0.9482	0.11	0.9483	0.11
	2,000	0.9500	0.10	0.9502	0.10
0.1	200	0.9493	0.32	0.9532	0.32
	400	0.9481	0.23	0.9499	0.23
	1,000	0.9481	0.14	0.9480	0.14
	1,600	0.9476	0.11	0.9483	0.11
	2,000	0.9498	0.10	0.9497	0.10

$\Delta_\theta$ , area between ROC curves;  $N$ , sample size; CP, coverage probability; WCI, mean width of the confidence interval.

A fourth simulation was performed to assess the impact of deviations from assumption (1) ( $\rho_{(-1)} = \rho_{(1)}$ ) on the estimation and inference properties of  $\hat{\Delta}_\theta$ .

### 3.2 | Results

For the first simulation study, in Scenario 1 when  $\Delta_\theta = 0.95$ , the symmetric confidence interval did not provide a good coverage probability; up to  $N = 400$  inclusive, the mean coverage probability was  $<93.5\%$ . With larger  $N$  (1,000, 1,600, and 2,000), the coverage probability of the symmetric confidence interval was close to 95%. The asymmetric confidence interval provided a better coverage probability; that is, nearly 95% irrespective of  $N$ . Moreover, the mean width of the asymmetric confidence interval was equal to that of the symmetric confidence interval irrespective of  $N$ , which means that with the asymmetric confidence interval the gain in coverage probability was not associated with a loss of precision (Table 1).

In other settings, the coverage probabilities of symmetric and asymmetric confidence intervals of  $\Delta_\theta$  were always close to 95% and their mean widths were close in almost all settings. In Scenario 3, the symmetric confidence interval gave coverage probabilities  $<94\%$  with  $N = 200$  (Tables 1–3).

For the second simulation study, the mean relative bias in  $\hat{\Delta}_\theta$  was  $< 3 \times 10^{-2}$  in all cases. The relative bias values did not meaningfully change with the risk scenarios (between  $-0.00417$  and  $0.03636$ ) but decreased when  $N$  increased. The mean relative bias in  $\hat{T}\hat{G}$  was also always close to 0; however, in most cases it was higher than the mean relative bias in  $\hat{\Delta}_\theta$ , especially when the true  $\Delta_\theta$  and  $N$  were equal to 0.1 and 200, respectively. The mean relative bias in  $\hat{\gamma}$  was always close to 0, and close to the mean relative bias in  $\hat{\Delta}_\theta$  (Tables 4–6).



**TABLE 2** Coverage probability and mean width of the 95% symmetric and asymmetric confidence intervals of  $\Delta_\theta$  (Scenario 2)

$\Delta_\theta$	$N$	Symmetric		Asymmetric	
		CP	WCI	CP	WCI
0.6	200	0.9414	0.24	0.9511	0.24
	400	0.9484	0.17	0.9526	0.17
	1,000	0.9486	0.11	0.9482	0.11
	1,600	0.9468	0.09	0.9481	0.09
	2,000	0.9462	0.08	0.9469	0.08
0.4	200	0.9487	0.32	0.9541	0.32
	400	0.9473	0.22	0.9512	0.22
	1,000	0.9505	0.14	0.9509	0.14
	1,600	0.9505	0.11	0.9504	0.11
	2,000	0.9490	0.10	0.9504	0.10
0.2	200	0.9481	0.36	0.9533	0.36
	400	0.9472	0.25	0.9496	0.25
	1,000	0.9507	0.16	0.9515	0.16
	1,600	0.9512	0.13	0.9519	0.13
	2,000	0.9514	0.11	0.9528	0.11
0.1	200	0.9500	0.37	0.9554	0.37
	400	0.9474	0.26	0.9503	0.26
	1,000	0.9542	0.16	0.9561	0.16
	1,600	0.9501	0.13	0.9501	0.13
	2,000	0.9490	0.12	0.9503	0.12

$\Delta_\theta$ , area between ROC curves;  $N$ , sample size; CP, coverage probability; WCI, mean width of the confidence interval.

The coverage probability of  $\gamma$  was close to 95% in all settings of Scenario 1 (between 93.67% and 95.07%). For Scenario 2, it was close to 95% except for  $\Delta_\theta = 0.6$  for which coverage probability was > 97.5%, and for  $\Delta_\theta = 0.4$  for which coverage probability was > 96% irrespective of the sample size. In Scenario 3, with  $\Delta_\theta = 0.4$ , the coverage probability was always > 98%, while it was close to 95% in other settings (Tables 4–6).

The coverage probability of the TG was always close to 95% in all settings of Scenario 1 except when  $\Delta_\theta = 0.1$ , and  $N \leq 400$  (> 96.9%). In Scenario 2, it was close to 95% in all settings, except when  $\Delta_\theta = 0.6$  for which the coverage probability of the bootstrap confidence interval decreased with the increase of  $N$  (92.86% with  $N = 200$ , and 64.45% with  $N = 2,000$ ). Note that for the latter setting, the TG estimator had a larger bias than in most of the other settings. Also, when  $\Delta_\theta = 0.1$  the coverage probability of the bootstrap confidence interval was > 96.7% with  $N \leq 400$ . In Scenario 3, it was close to 95% in all settings, except when  $\Delta_\theta = 0.1$  and  $N \leq 1,600$  (> 96.7%).

The power of  $\hat{\Delta}_\theta$  decreased along with the decrease of  $N$  and as  $\Delta_\theta$  approached 0. This power varied between risk scenarios but decreased along with the decrease of the risk in each arm. With  $N = 200$  and a theoretical  $\Delta_\theta = 0.2$ , the < 70% power was insufficient to demonstrate a significant predictive ability; with  $\Delta_\theta = 0.1$ , large  $N$  values were needed. In Scenarios 1, and 2,  $N = 1,600$  was necessary to ensure > 80% power. In Scenario 3,  $N = 2,000$  was insufficient to ensure > 80% power (power was equal to 65%). The power of  $\hat{\Delta}_\theta$  was almost always equal to the power of  $\hat{\gamma}$  and to the one of  $\hat{\beta}_3$  in all settings (Tables 4–6).

For the third simulation study, in Scenario 1 and  $\Delta_\theta = 0$  ( $H_0$ ), the  $\alpha$ -risk of rejecting  $H_0$  for  $\hat{\Delta}_\theta$  got closer to 5% as  $N$  increased; with  $N = 40$  and 1,000, the  $\alpha$ -risk was equal to 0.0603 and 0.0519, respectively. Similar results were obtained with the interaction coefficient estimation method (Table 7).

For the fourth simulation study, the results show that the  $\alpha$ -risk was always close to 5% in case of small differences between the overall risk of event in the two treatment arms, except when the overall risk of event in each treatment arm was low (0.0875 vs. 0.1125, and 0.075 vs. 0.125) where an inflation in the  $\alpha$ -risk was observed (e.g.,  $\alpha = 8.69\%$  with  $N = 5,000$  and  $\rho_{(-1)} = 0.075$  vs.  $\rho_{(1)} = 0.125$ ; Table S3 in Supporting Information). Source code to reproduce all the results is available as Supporting Information on the journal's web page.

**TABLE 3** Coverage probability and mean width of the 95% symmetric and asymmetric confidence intervals of  $\Delta_\theta$  (Scenario 3)

$\Delta_\theta$	$N$	Symmetric		Asymmetric	
		CP	WCI	CP	WCI
0.4	200	0.9325	0.43	0.9458	0.43
	400	0.9417	0.30	0.9477	0.30
	1,000	0.9482	0.19	0.9495	0.19
	1,600	0.9512	0.15	0.9518	0.15
	2,000	0.9505	0.14	0.9529	0.14
0.2	200	0.9334	0.51	0.9449	0.50
	400	0.9419	0.36	0.9459	0.36
	1,000	0.9468	0.23	0.9493	0.23
	1,600	0.9529	0.18	0.9536	0.18
	2,000	0.9485	0.16	0.9501	0.16
0.1	200	0.9335	0.53	0.9446	0.52
	400	0.9424	0.37	0.9478	0.37
	1,000	0.9483	0.24	0.9504	0.24
	1,600	0.9535	0.19	0.9547	0.19
	2,000	0.9496	0.17	0.9508	0.17

$\Delta_\theta$ , area between ROC curves;  $N$ , sample size; CP, coverage probability; WCI, mean width of the confidence interval.

#### 4 | APPLICATION TO THE PETACC-8 TRIAL

The PETACC-8 trial was led by the *Fédération Francophone de Cancérologie Digestive*. This trial was an open-label, randomized, controlled, multinational phase 3 study that included patients aged 18 to 75 years with pathologically confirmed and resected stage III colon adenocarcinoma (Taieb et al., 2014). The trial compared the efficacy of FOLFOX4 (oxaliplatin, fluorouracil, and leucovorin) versus FOLFOX4 + cetuximab (an inhibitor of EGFR). As patients with *KRAS* exon 2 mutated tumors were found resistant to EGFR antibodies in the metastatic setting (De Roock et al., 2008; Di Fiore et al., 2007; Lièvre et al., 2008), an amendment to the protocol restricted the enrollment to patients with *KRAS* wild-type tumors.

Adding cetuximab to FOLFOX4 did not improve disease-free survival in the intent-to-treat population (hazard ratio: 1.05; 95% CI [0.85; 1.29]) (Taieb et al., 2014). However, the heterogeneity of the response to FOLFOX4 + cetuximab led the investigators to assume that cetuximab could be effective in specific patient subgroups in the per protocol population ( $N = 1,432$ ). The study analyzed the amplification levels of two genes involved in cancer (*DDR2* and *FBXW7*) and considered them as potential treatment selection markers; this restricted the analysis to patients in the per protocol population that had measures of both markers ( $N = 1,068$ ). The capacities of these two genes for treatment selection were assessed in the per protocol population to restrict the analysis to the patients who actually received their treatment. The study outcome was cancer progression or death within 21 months. This delay was a good compromise between the number of censored data and a sufficient time to observe treatment effects (without censoring, the delay would have been too short). The number of censored data was 40; for these, outcomes were imputed (Dmitrienko, Molenberghs, Chuang-Stein, & Offen, 2005) (for more details on this imputation, see Supporting Information). At 21 months, the risk of event occurrence in the FOLFOX4 arm was 0.15 versus 0.16 in the FOLFOX4 + cetuximab arm.

Concerning the *DDR2* gene (Figure 2), the FOLFOX4 ROC curve was located below the diagonal; in this arm, the risk of event occurrence decreased with the increase of the amplification level of *DDR2*. The FOLFOX4 + cetuximab curve was located above the diagonal; in this arm, the risk of event occurrence increased with the amplification level of *DDR2*. The estimated  $\hat{\Delta}_\theta$  was  $-0.12$  [ $-0.22$ ;  $-0.03$ ] and the Wald test p-value = 0.012. The  $\hat{TG}$  was estimated to 0.04 and the interaction coefficient of the logistic regression model to 18.12 (p-value = 0.037). The estimated  $\hat{\gamma}$  was  $-0.06$  [ $-0.12$ ;  $-0.01$ ] and the Wald test p-value = 0.015.

Concerning gene *FBXW7* (Figure 3), the FOLFOX4 ROC curve was located above the diagonal, whereas the FOLFOX4 + cetuximab curve was close to the diagonal. Thus, the risk of event occurrence under FOLFOX4 alone increased with the increase of the amplification level of *FBXW7*. The estimated  $\hat{\Delta}_\theta$  was 0.07 [ $-0.03$ ; 0.16] and the p-value = 0.158. The  $\hat{TG}$  was estimated to 0.02 and the interaction coefficient of the logistic regression model to  $-15.22$  (p-value = 0.163). The estimated  $\hat{\gamma}$  was 0.04 [ $-0.01$ ; 0.09] and the Wald test p-value = 0.166.

**TABLE 4** Performance of the treatment selection indicators (Scenario 1)

$\Delta_\theta$	$N$	$\hat{\Delta}_\theta$			$\hat{\gamma}$			$\hat{TG}$		$\hat{\beta}_3$
		MRB	CP	Power	MRB	CP	Power	MRB	CP	Power
0.6	200	0.00102	0.9507	1	0.00041	0.9367	1	0.01242	0.9366	1
	400	-0.00040	0.9519	1	-0.00079	0.9436	1	0.00498	0.9442	1
	1,000	-0.00104	0.9480	1	-0.00120	0.9454	1	0.00105	0.9452	1
	1,600	-0.00063	0.9500	1	-0.00071	0.9486	1	0.00059	0.9473	1
	2,000	-0.00015	0.9512	1	-0.00023	0.9507	1	0.00084	0.9477	1
0.4	200	0.00246	0.9526	0.9990	0.00087	0.9424	0.9990	0.01626	0.9394	0.9991
	400	-0.00073	0.9526	1	-0.00153	0.9461	1	0.00605	0.9448	1
	1,000	-0.00190	0.9484	1	-0.00226	0.9460	1	0.00077	0.9468	1
	1,600	-0.00110	0.9495	1	-0.00130	0.9494	1	0.00050	0.9479	1
	2,000	-0.00031	0.9508	1	-0.00048	0.9502	1	0.00103	0.9474	1
0.2	200	0.00669	0.9537	0.7062	0.00469	0.9435	0.7195	0.02378	0.9473	0.7170
	400	-0.00116	0.9509	0.9377	-0.00225	0.9462	0.9378	0.00646	0.9462	0.9450
	1,000	-0.00417	0.9480	0.9996	-0.00463	0.9457	0.9996	-0.00104	0.9493	0.9999
	1,600	-0.00227	0.9483	1	-0.00259	0.9483	1	-0.00049	0.9479	1
	2,000	-0.00057	0.9502	1	-0.00087	0.9497	1	0.00101	0.9488	1
0.1	200	0.01491	0.9532	0.2450	0.01260	0.9432	0.2585	0.11453	0.9696	0.2473
	400	-0.00220	0.9499	0.4182	-0.00331	0.9462	0.4265	0.02345	0.9718	0.4267
	1,000	-0.00838	0.9480	0.7840	-0.00892	0.9469	0.7858	-0.00475	0.9502	0.8042
	1,600	-0.00463	0.9483	0.9390	-0.00498	0.9469	0.9400	-0.00283	0.9482	0.9456
	2,000	-0.00118	0.9497	0.9739	-0.00153	0.9494	0.9744	0.00063	0.9495	0.9795

$\Delta_\theta$ , area between ROC curves;  $N$ , sample size; MRB, mean relative bias; CP, coverage probability;  $\beta_3$ , interaction coefficient.

**TABLE 5** Performance of the treatment selection indicators (Scenario 2)

$\Delta_\theta$	$N$	$\hat{\Delta}_\theta$			$\hat{\gamma}$			$\hat{TG}$		$\hat{\beta}_3$
		MRB	CP	Power	MRB	CP	Power	MRB	CP	Power
0.6	200	0.00023	0.9511	1	-0.00158	0.9765	1	-0.03992	0.9286	0.9998
	400	-0.00088	0.9526	1	-0.00244	0.9819	1	-0.04663	0.9038	1
	1,000	-0.00075	0.9482	1	-0.00179	0.9788	1	-0.05005	0.7978	1
	1,600	0.00113	0.9481	1	0.00161	0.9821	1	-0.04830	0.7091	1
	2,000	0.00084	0.9469	1	0.00084	0.9824	1	-0.04890	0.6445	1
0.4	200	0.00168	0.9541	0.9971	-0.00078	0.9619	0.9970	-0.00747	0.9434	0.9980
	400	0.00206	0.9512	1	0.00046	0.9651	1	-0.01416	0.9479	1
	1,000	-0.00067	0.9509	1	-0.00092	0.9670	1	-0.02141	0.9467	1
	1,600	-0.00013	0.9504	1	-0.00005	0.9664	1	-0.02103	0.9388	1
	2,000	0.00109	0.9504	1	0.00075	0.9662	1	-0.02081	0.9363	1
0.2	200	0.02271	0.9533	0.6048	0.01992	0.9490	0.6102	0.03718	0.9561	0.6184
	400	0.00763	0.9496	0.8700	0.00634	0.9488	0.8727	0.00863	0.9439	0.8836
	1,000	-0.00082	0.9515	0.9985	-0.00166	0.9545	0.9986	-0.00515	0.9511	0.9986
	1,600	-0.00207	0.9519	1	-0.00289	0.9555	1	-0.00747	0.9485	1
	2,000	0.00300	0.9528	1	0.00267	0.9556	1	-0.00129	0.9485	1
0.1	200	0.03636	0.9554	0.2005	0.03423	0.9489	0.2084	0.17649	0.9671	0.2013
	400	0.02256	0.9503	0.3398	0.02148	0.9474	0.3467	0.05647	0.9700	0.3444
	1,000	-0.00036	0.9561	0.6605	-0.00088	0.9549	0.6627	0.00176	0.9573	0.6798
	1,600	-0.00885	0.9501	0.8513	-0.00905	0.9514	0.8504	-0.00991	0.9496	0.8634
	2,000	-0.00157	0.9503	0.9180	-0.00252	0.9500	0.9187	0.00144	0.9474	0.9283

$\Delta_\theta$ , area between ROC curves;  $N$ , sample size; MRB, mean relative bias; CP, coverage probability;  $\beta_3$ , interaction coefficient.

**TABLE 6** Performance of the treatment selection indicators (Scenario 3)

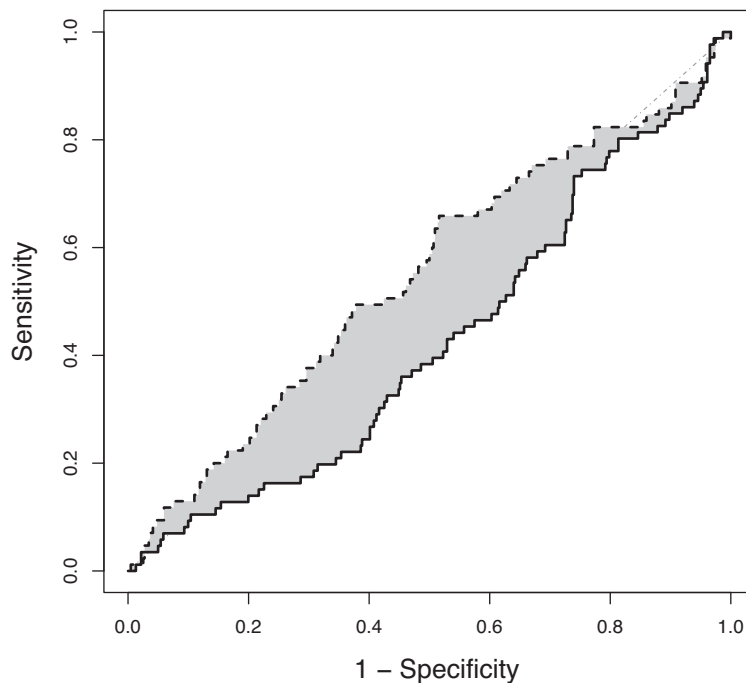
$\Delta_\theta$	$N$	$\hat{\Delta}_\theta$			$\hat{\gamma}$			$\widehat{TG}$		$\hat{\beta}_3$
		MRB	CP	Power	MRB	CP	Power	MRB	CP	Power
0.4	200	0.00580	0.9458	0.9406	-0.00215	0.9848	0.8764	0.00434	0.9277	0.9516
	400	0.00296	0.9477	0.9989	0.00061	0.9854	0.9951	-0.00481	0.9390	0.9997
	1,000	-0.00243	0.9495	1	-0.00542	0.9859	1	-0.01636	0.9482	1
	1,600	0.00058	0.9518	1	0.00101	0.9879	1	-0.01227	0.9474	1
	2,000	0.00005	0.9529	1	-0.00093	0.9868	1	-0.01496	0.9449	1
0.2	200	0.01893	0.9449	0.3624	0.01393	0.9580	0.3244	0.05878	0.9491	0.3471
	400	0.00556	0.9459	0.5880	0.00331	0.9606	0.5584	0.01065	0.9501	0.6045
	1,000	0.00157	0.9493	0.9339	0.00054	0.9626	0.9248	-0.00024	0.9476	0.9476
	1,600	-0.00050	0.9536	0.9939	-0.00068	0.9657	0.9926	-0.00271	0.9488	0.9961
	2,000	0.00190	0.9501	0.9981	0.00171	0.9627	0.9980	-0.00072	0.9484	0.9992
0.1	200	0.02403	0.9446	0.1327	0.02026	0.9473	0.1172	0.37949	0.9547	0.1158
	400	0.01717	0.9478	0.1942	0.01630	0.9493	0.1841	0.15611	0.9623	0.1921
	1,000	0.00072	0.9504	0.3800	0.00013	0.9530	0.3713	0.02104	0.9723	0.3910
	1,600	0.00049	0.9547	0.5482	0.00097	0.9573	0.5416	0.00438	0.9675	0.5648
	2,000	0.00228	0.9508	0.6514	0.00199	0.9536	0.6452	0.00613	0.9546	0.6704

$\Delta_\theta$ , area between ROC curves;  $N$ , sample size; MRB, mean relative bias; CP, coverage probability;  $\beta_3$ , interaction coefficient.

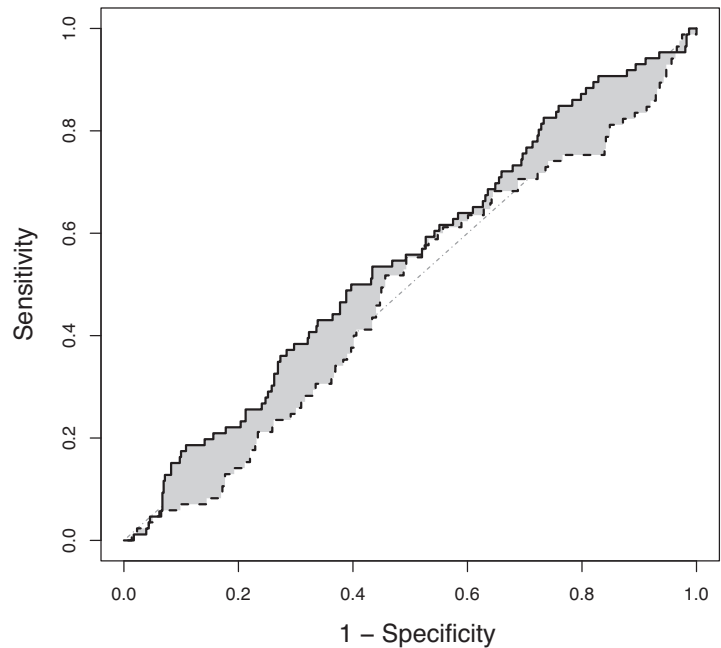
**TABLE 7**  $\alpha$ -Risk under  $H_0$  (third simulation study)

$\Delta_\theta$	$N$	$\hat{\Delta}_\theta$	$\hat{\beta}_3$
		$\alpha$	$\alpha$
0	40	0.0603	0.0465
	100	0.0503	0.0459
	200	0.0500	0.0506
	400	0.0527	0.0512
	1,000	0.0519	0.0487

$\Delta_\theta$ , area between ROC curves;  $N$ , sample size.

**FIGURE 2** ROC curves associated with the *DDR2* gene (dotted line: FOLFOX4 + cetuximab treatment; solid line: FOLFOX4 treatment)

**FIGURE 3** ROC curves associated with the *FBXW7* gene (dotted line: FOLFOX4 + cetuximab treatment; solid line: FOLFOX4 treatment)



The  $\widehat{TG}$  and the  $\widehat{\gamma}$  concordance measure were higher for the *DDR2* gene than for the *FBXW7* gene, in agreement with the  $\widehat{\Delta}_\theta$  results. This means that the capacity of the *DDR2* gene for treatment selection is higher than that of the *FBXW7* gene. Moreover, the  $\widehat{\Delta}_\theta$  of the *DDR2* gene (but not that of the *FBXW7* gene) was significantly different from 0, in agreement with the results from the  $\widehat{\gamma}$  concordance index and the interaction coefficient. The *DDR2* gene is thus a treatment selection marker whereas the *FBXW7* gene cannot be considered as a treatment selection marker.

## 5 | DISCUSSION

We present in this article the ABC that may be measured by the  $\Delta_\theta$  metric, a very simple indicator to quantify and test the overall capacity of a quantitative marker for treatment selection when the overall risk of event in each treatment arm is the same. The simulation results showed that the proposed estimation method has good performances. This is reflected by the low mean relative bias in  $\widehat{\Delta}_\theta$  ( $< 3 \times 10^{-2}$ ) in all scenarios and all sample sizes, which was comparable to that of the  $\widehat{\gamma}$  concordance measure. Furthermore, the mean relative bias of  $\widehat{\Delta}_\theta$  was lower than that of the  $\widehat{TG}$  in almost all scenarios and settings, which is explained by the interaction not being linear on the logit scale (except in Scenario 1).

As for  $\widehat{\beta}_3$  (the estimator of the interaction coefficient) and the  $\widehat{\gamma}$  concordance measure, the power of  $\widehat{\Delta}_\theta$  decreased along with the decrease of the sample size and the decrease of the risk of event occurrence. In Scenario 2 and a theoretical  $\Delta_\theta = 0.2$ , 400 patients were sufficient to reach a power  $> 80\%$ , whereas in Scenarios 1 and 2 and a theoretical  $\Delta_\theta = 0.1$ , 1,600 patients or more were needed to reach sufficient power. Even higher  $N$  values were needed in Scenario 3 because the risk of event occurrence was equal to 0.1 in each arm. Overall, the number of patients needed to achieve sufficient power is in line with the classical need for rather high sample sizes in clinical trials when a test for interaction is performed to detect a treatment selection marker (Janes, Brown, & Pepe, 2015a). In all settings, the power of  $\widehat{\Delta}_\theta$  was comparable to the one of  $\widehat{\gamma}$  and to the interaction coefficient estimated by a logistic regression model. These results are interesting because they show that using a non-parametric method—that does not require verifying the complex assumptions required by the parametric methods—would not impact the power of detecting a treatment selection marker, and are in favor of  $\widehat{\Delta}_\theta$ .

The coverage probability of the asymmetric confidence interval of  $\Delta_\theta$  was always close to 95%, compared to the one of  $\gamma$  that sometimes exceeded 95% (whereas the working model used to calculate the variance of  $\widehat{\gamma}$  led to the most optimal variance estimate in the simulations) and to the one of  $\widehat{TG}$  that exceeded 95% in multiple settings. Thus, the use of the asymmetric confidence interval of  $\Delta_\theta$  is recommended. Again, these results are in favor of  $\widehat{\Delta}_\theta$ .

Three assumptions are necessary to use the proposed  $\Delta_\theta$  metric. The first is that the risk of event in both arms either increases or decreases monotonically over the marker values. This is in fact an assumption similar to but a little more stringent than the assumption of monotonicity of the risk difference between the two treatment arms over the marker values required for other

methods proposed to assess a treatment selection marker. The second assumption is that the two ROC curves do not intersect, so that the ABC can be calculated by a simple difference between AUCs. It can be shown that when the ROC curves intersect (and that  $\rho_{(-1)} = \rho_{(1)}$ ) the marker has a treatment selection ability (see Supporting Information), so the ABC could still be an indicator of treatment selection ability in that situation, the ABC being calculated in this case using partial AUCs (McClish, 1989). However this approach would need further investigation. The third assumption is that the overall risk of event is the same in the two treatment arms. This assumption can be tested, but small departure from this hypothesis may not be identified by the Chi-squared test. Even in the case of small deviations from this assumption, a simulation study showed that the mean  $\hat{\Delta}_\theta$  was always  $< 3 \times 10^{-3}$  for markers without treatment selection capacity. The  $\alpha$ -risk was still close to 0.05, except for a small overall risk of event ( $\approx 0.1$ ).

Contrary to the interaction coefficient in a risk model, the  $\Delta_\theta$  metric does not depend on the range of marker values. The  $\gamma$  and the TG indicators have also this property; however, it can be easily demonstrated that their maximum depends on the overall risk of event occurrence in each arm, which is not the case for  $\Delta_\theta$  which is always bounded between  $-1$  and  $1$ . Hence,  $\Delta_\theta$  is an indicator that facilitates comparisons of treatment selection capacities between markers (and between studies). The  $\Delta_\theta$  indicator is thus useful to detect markers that may be used for treatment decision-making, and is complementary with all aforementioned parametric methods.

A binary outcome is necessary to build ROC curves or marker-by-treatment predictiveness curves. However, a frequent problem with binary outcomes is the presence of censored data in long-term trials. In the example of application herein, as there were few censored data at the time of follow-up chosen (40 in a sample of 1,068 patients) imputation was not considered to affect significantly the conclusions, but with a longer follow-up the potential impact of imputation may become problematic. To avoid the use of imputation, methods to calculate AUC in the presence of censored data may be used (such as the time-dependent AUC) (Blanche, Dartigues, & Jacqmin-Gadda, 2013; Heagerty, Lumley, & Pepe, 2000; Li, Greene, & Hu, 2018).

The capacities of two markers for treatment selection may be compared by testing the difference in  $\Delta_\theta$ s against zero; the standard error of this difference requires the calculation of the covariance between the AUCs of the two markers in each arm. This may use the method developed by DeLong et al. (1988) for correlated ROC curves. But one has to keep in mind that a large  $\hat{\Delta}_\theta$  estimate is not sufficient to detect clinically useful markers and compare them. For example, when the benefit from a treatment increases together with the marker value and the marker-by-treatment predictiveness curves do not intersect, the marker can still be considered as a treatment selection marker because the difference in risks of event is not constant over all marker values. Nevertheless, in this example, one of two treatments is always associated with a lower risk of event; this means that this treatment should be preferred whatever the marker value (and the marker is not useful for treatment choice). Yet, taking into account the mean risks in the target population and the consequences of each treatment strategy (e.g., adverse events), the treatment associated with a higher risk of event occurrence may nevertheless be preferred at marker values where the difference in risks is low if it is associated with fewer adverse events than the other. An extension of the present work would be to include the clinical utility of the treatments in the assessment of treatment selection markers; this requires the estimation of the optimal threshold for treatment allocation that takes into account the clinical utility and the mean risk of event occurrence in each treatment arm (Blangero et al., 2019; Jund, Rabilloud, Wallon, & Ecochard, 2005; Subtil & Rabilloud, 2015).

In conclusion, the area between ROC curves was able to quantify and test the capacity of a quantitative marker for treatment selection. The method is easy-to-use and complements previous parametric methods when the parametric assumptions cannot be verified.


## ACKNOWLEDGMENT

The authors thank Dr. Philip Robinson (Hospices Civils de Lyon) whose suggestions improved significantly the present manuscript. This work was supported by the Fondation pour la Recherche Médicale (FRM grant number ECO 20160736050 to YB).

## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Yoann Blangero  <https://orcid.org/0000-0003-0309-3399>

## REFERENCES

- Ballman, K. V. (2015). Biomarker: Predictive or prognostic? *Journal of Clinical Oncology*, *33*, 3968–3971.
- Blanche, P., Dartigues, J. F., & Jacqmin-Gadda, H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, *55*, 687–704.
- Blangero, Y., Rabilloud, M., Ecochard, R., & Subtil, F. (2019). A Bayesian method to estimate the optimal threshold of a marker used to select patients' treatment. *Statistical Methods in Medical Research*, *29*, 29–43. <https://doi.org/10.1177/0962280218821394>
- Böhning, D., Hempfling, A., Schelp, F. P., & Schlattmann, P. (1992). The area between curves (ABC): Measure in nutritional anthropometry. *Statistics in Medicine*, *11*, 1289–1304.
- Byar, D. P. (1985). Assessing apparent treatment-covariate interactions in randomized clinical trials. *Statistics in Medicine*, *4*, 255–263.
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.
- De Roock, W., Piessevaux, H., De Schutter, J., Janssens, M., De Hertogh, G., Personeni, N., & Tejpar, S. (2008). KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. *Annals of Oncology*, *19*, 508–515.
- Di Fiore, F., Blanchard, F., Charbonnier, F., Le Pessot, F., Lamy, A., Galais, M. P., & Frebourg, T. (2007). Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by cetuximab plus chemotherapy. *British Journal of Cancer*, *96*, 1166–1169.
- Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., & Offen, W. (2005). *Analysis of clinical trials using SAS: A practical guide*. Cary, NC: SAS Institute.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: CRC Press.
- Gioacchetti, K. E. D., & Heyse, J. (2015). Using proportion of similar response to evaluate correlates of protection for vaccine efficacy. *Statistical Methods in Medical Research*, *24*, 273–286.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.
- Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, *56*, 337–344.
- Hoëfding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, *19*, 293–325.
- Huang, Y., Gilbert, P. B., & Janes, H. (2012). Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*, *68*, 687–696.
- Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions! *Journal of Clinical Oncology*, *29*, 4718.
- Janes, H., Brown, M. D., Huang, Y., & Pepe, M. S. (2014a). An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics*, *10*, 99–121.
- Janes, H., Brown, M. D., & Pepe, M. S. (2015a). Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Statistics in Medicine*, *34*, 3503–3515.
- Janes, H., Pepe, M. S., Bossuyt, P. M., & Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine* *154*, 253–259.
- Janes, H., Pepe, M. S., & Huang, Y. (2014b). A framework for evaluating markers used to select patient treatment. *Medical Decision Making*, *34*, 159–167.
- Janes, H., Pepe, M. S., McShane, L. M., Sargent, D. J., & Heagerty, P. J. (2015b). The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *Journal of the National Cancer Institute*, *107*, djv157. <https://doi.org/10.1093/jnci/djv157>
- Jund, J., Rabilloud, M., Wallon, M., & Ecochard, R. (2005). Methods to estimate the optimal threshold for normally or log-normally distributed biological tests. *Medical Decision Making*, *25*, 406–415.
- Li, L., Greene, T., & Hu, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical Methods in Medical Research*, *27*, 2264–2278.
- Lièvre, A., Bachet, J. B., Boige, V., Cayre, A., Le Corre, D., Buc, E., & Laurent-Puig, P. (2008). KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. *Journal of Clinical Oncology*, *26*, 374–379.
- McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*, 190–195.
- Metz, C. E., & Pan, X. (1999). "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology*, *43*, 1–33.
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.
- Skougaard, K., Nielsen, D., Jensen, B. V., Pfeiffer, P., & Hendel, H. W. (2016). Early (18)F-FDG-PET/CT as a predictive marker for treatment response and survival in patients with metastatic colorectal cancer treated with irinotecan and cetuximab. *Acta Oncology*, *55*, 1175–1182.
- Song, X., & Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics*, *60*, 874–883.
- Subtil, F., & Rabilloud, M. (2015). An enhancement of ROC curves made them clinically relevant for diagnostic-test comparison and optimal-threshold determination. *Journal of Clinical Epidemiology*, *68*, 752–759.
- Taieb, J., Taberero, J., Mini, E., Subtil, F., Folprecht, G., Van Laethem, J. L., & Lepage, C. (2014). Oxaliplatin, fluorouracil, and leucovorin with or without cetuximab in patients with resected stage III colon cancer (PETACC-8): An open-label, randomised phase 3 trial. *The Lancet Oncology*, *15*, 862–873.

- Viallon, V., & Latouche, A. (2011). Discrimination measures for survival outcomes: Connection between the AUC and the predictiveness curve. *Biometrical Journal*, 53, 217–236.
- Vickers, A. J., Kattan, M. W., & Sargent, D. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*, 8, 14.
- Weidhaas, J. B., Harris, J., Schaeue, D., Chen, A. M., Chin, R., Axelrod, R., & Chung, C. H. (2016). The KRAS-variant and cetuximab response in head and neck squamous cell cancer: A secondary analysis of a randomized clinical trial. *JAMA Oncology*, 3, 483–491.
- Zhang, Z., Ma, S., Nie, L., & Soon, G. (2017). A quantitative concordance measure for comparing and combining treatment selection markers. *The International Journal of Biostatistics*, 13. <https://doi.org/10.1515/ijb-2016-0064>
- Zhang, Z., Nie, L., Soon, G., & Liu, A. (2014). The use of covariates and random effects in evaluating predictive biomarkers under a potential outcome framework. *Annals of Applied Statistics*, 8, 2336–2355.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Blangero Y, Rabilloud M, Laurent-Puig P, Le Malicot K, Lepage C, Ecochard R, Taieb J, Subtil F. The area between ROC curves, a non-parametric method to evaluate a biomarker for patient treatment selection. *Biometrical Journal*. 2020;1–18. <https://doi.org/10.1002/bimj.201900171>

## APPENDIX

### A.1 Example of a perfect treatment selection marker

In this scenario, two different  $\Delta_\theta$  are computed. One is equal to  $-1$  (Scenario A) and the other to  $-0.5$  (Scenario B). As stated in the manuscript,  $\Delta_\theta = -1$  has the same interpretation as  $\Delta_\theta = 1$ . The marker-by-treatment predictiveness curves that correspond to these scenarios are presented in Figure A.1.

In Scenario A ( $\Delta_\theta = -1$ ), the marker-by-treatment predictiveness curves show that below the threshold of 0.5, the risk of event occurrence is equal to 1 in the innovative arm and 0 in the reference one, and that above the threshold of 0.5, the risk of event occurrence is equal to 0 in the innovative arm and 1 in the reference one. Theoretically, no patient is likely to experience the event if the best treatment is given according to the marker values.

In Scenario B ( $\Delta_\theta = -0.5$ ), the ability of the marker to distinguish patients who would and would not experience the event is not maximal. Thus,  $\Delta_\theta$  is not equal to  $-1$  (or 1). The marker remains interesting because the decision to treat changes according to the biomarker value with sometimes high risk differences between the two treatments. In this case, the patients would not be optimally treated because the treatment allocation process would not prevent the occurrence of the event in all patients. The higher the risk difference between the arms, the farther are the two ROC curves from each other, the higher is  $\Delta_\theta$ , and the stronger is the treatment selection ability of the marker.

### A.2 Validity of the $\Delta_\theta$ indicator

First, let us demonstrate that when a marker has no capacity for treatment selection then the  $\Delta_\theta$  indicator is equal to 0. When a marker has no capacity for treatment selection:

$$\delta(v) = P(E = 1|T = -1, V = v) - P(E = 1|T = 1, V = v) = \rho_{(-1)} - \rho_{(1)} = 0 \quad \forall v$$

With the assumption (1) that the overall risk in each treatment arm is equal,  $\rho_{(-1)} - \rho_{(1)} = 0$ .

The expression of the  $\Delta_\theta$  indicator is

$$\Delta_\theta = \frac{\int_{-\infty}^{+\infty} F(v) \times \delta(v) dF(v)}{\rho(1 - \rho)}$$

It is easy to see that when a marker has no capacity for treatment selection then the numerator is equal to 0. So when a marker has no capacity for treatment selection  $\Delta_\theta = 0$ .

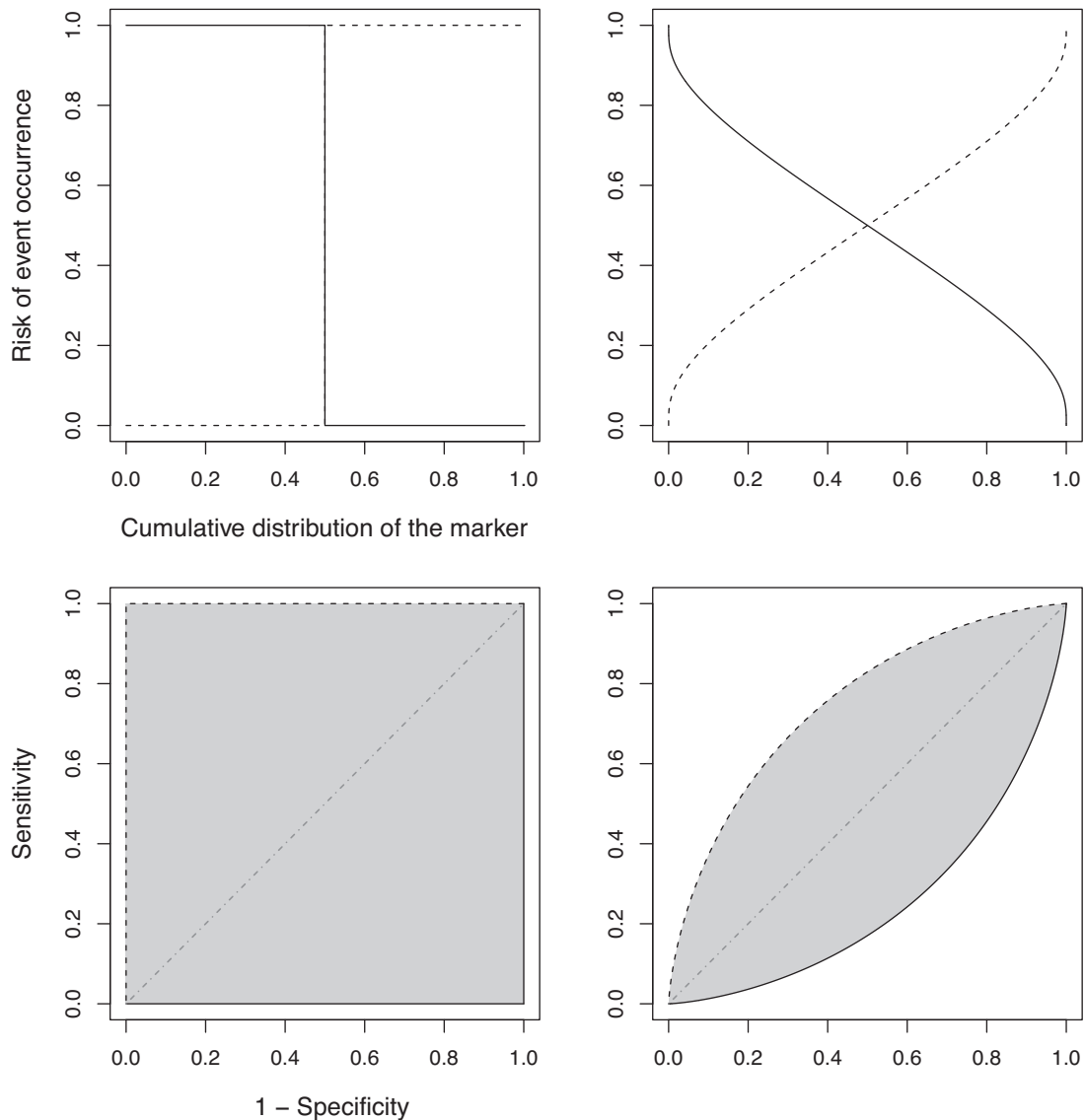


Now, let us demonstrate that when  $\Delta_\theta = 0$ , then a marker has no capacity for treatment selection. With the assumption that the ROC curves do not intersect,  $\Delta_\theta = 0$  means that the ROC curves for the two treatment arms overlap. This means that for the infinity of couples  $(c_{(-1)}, c_{(1)})$  that correspond to the set  $\Omega$  of marker values for which the ROC curves overlap, the following systems must hold:

$$\begin{cases} P(V > c_{(-1)} | E = 1, T = -1) = P(V > c_{(1)} | E = 1, T = 1) \\ P(V > c_{(-1)} | E = 0, T = -1) = P(V > c_{(1)} | E = 0, T = 1) \end{cases} \quad \forall (c_{(-1)}, c_{(1)}) \in \Omega$$

$$\begin{cases} P(V \leq c_{(-1)} | E = 1, T = -1) = P(V \leq c_{(1)} | E = 1, T = 1) \\ P(V \leq c_{(-1)} | E = 0, T = -1) = P(V \leq c_{(1)} | E = 0, T = 1) \end{cases} \quad \forall (c_{(-1)}, c_{(1)}) \in \Omega \quad (4)$$

For a given couple  $(c_{(-1)}, c_{(1)})$ , the sensitivity and specificity in one treatment arm must be equal to those of the other treatment arm.



**FIGURE A.1** Marker-by-treatment predictiveness curves of two markers, and their corresponding ROC curves (dotted line: innovative treatment; solid line: reference treatment; shaded area: area between ROC curves)

It is also important to respect the randomization constraint that is assumed in a clinical trial context (Equation (3)). This can be expressed as:

$$P(V \leq c|T = -1) = P(V \leq c|T = 1) \quad \forall c$$

This expression can be rewritten according to assumption (1) :

$$\begin{aligned} & \rho[P(V \leq c|E = 1, T = -1) - P(V \leq c|E = 1, T = 1)] \\ & - (1 - \rho)[P(V \leq c|E = 0, T = 1) - P(V \leq c|E = 0, T = -1)] = 0 \end{aligned} \quad \forall c$$

This equation must be true at the point  $c_{(-1)}$ :

$$\begin{aligned} & \rho[P(V \leq c_{(-1)}|E = 1, T = -1) - P(V \leq c_{(-1)}|E = 1, T = 1)] \\ & - (1 - \rho)[P(V \leq c_{(-1)}|E = 0, T = 1) - P(V \leq c_{(-1)}|E = 0, T = -1)] = 0 \end{aligned}$$

Replacing by the information provided in system (4), it is possible to write:

$$\begin{aligned} & \rho[P(V \leq c_{(1)}|E = 1, T = 1) - P(V \leq c_{(-1)}|E = 1, T = 1)] \\ & - (1 - \rho)[P(V \leq c_{(-1)}|E = 0, T = 1) - P(V \leq c_{(1)}|E = 0, T = 1)] = 0 \end{aligned}$$

If  $c_{(-1)} > c_{(1)}$ , the sensitivity in the innovative treatment for  $c_{(-1)}$  is lower than for  $c_{(1)}$ , and the specificity of the innovative treatment for  $c_{(-1)}$  is greater than for  $c_{(1)}$ ; thus, the left part of the equation is strictly negative and the equation cannot be true.

If  $c_{(-1)} < c_{(1)}$ , the sensitivity in the innovative treatment for  $c_{(-1)}$  is greater than for  $c_{(1)}$ , and the specificity of the innovative treatment for  $c_{(-1)}$  is lower than for  $c_{(1)}$ ; thus, the left part of the equation is strictly positive and the equation again cannot be true.

So the equation is only true when  $c_{(-1)} = c_{(1)} = c$ . It is then possible to write again the system (4) including this constraint:

$$\begin{cases} P(V \leq c|E = 1, T = -1) = P(V \leq c|E = 1, T = 1) \\ P(V \leq c|E = 0, T = -1) = P(V \leq c|E = 0, T = 1) \end{cases} \quad \forall c$$

Using the Bayes theorem, this implies that

$$P(E = 1|T = -1, V \leq c) = P(E = 1|T = 1, V \leq c) \quad \forall c$$

where the result comes from assumption (1) and the randomization constraint. It follows that

$$P(E = 1|T = -1, V = c) = P(E = 1|T = 1, V = c) = \rho_{(-1)} - \rho_{(1)} = 0 \quad \forall c$$

Which is the definition of a marker without capacity for treatment selection given in Equation (2).