ARTICLE TYPE: ORIGINAL RESEARCH ARTICLE

Improving the biomarker diagnostic capacity via functional transformations

P. Martínez-Camblor[a,b] and S. Pérez-Fernández[c] and S. Díaz-Coto[c]

[a]The Dartmouth Institute for Health Policy and Clinical Practice, Hanover (NH); USA
[b]Universidad Autonoma de Chile, Santiago, Chile [c]Universidad de Oviedo, Asturies, Spain

ARTICLE HISTORY
Compiled September 25, 2018

ABSTRACT
The use of the area under the receiver-operating characteristic, ROC, curve (AUC) as an index of diagnostic accuracy is overwhelming in fields such as biomedical science and machine learning. It seems that a larger AUC value has become synonymous with a better performance. The functional transformation of the marker values has been proposed in the specialized literature as a procedure for increasing the AUC and therefore the diagnostic accuracy. However, the classification process is based on some regions (classification subsets) which support the decision made; one subject is classified as positive if its marker is within this region and classified as negative otherwise. In this paper we study the capacity of improving the classification performance of univariate biomarkers via functional transformations and the impact of this transformation on the final classification regions based on a real-world dataset. Particularly, we consider the problem of determining the gender of a subject based on the *Mode frequency* of his/her voice. The shape of the cumulative distribution function of this characteristic in both the male and the female groups makes the resulting classification problem useful for illustrating the differences between having useful diagnostic rules and obtaining an optimal AUC value. Our point is that improving the AUC by means of a functional transformation can produce classification regions with no practical interpretability. We propose to improve the classification accuracy by making the selection of the classification subsets more flexible while preserving their interpretability. Besides, we provide different graphical approximations which allow us a better understanding of the classification problem.

KEYWORDS
Area under the ROC curve; (Bio)markers; Generalized receiver-operating characteristic (gROC) curve; Receiver-operating characteristic (ROC) curve.

1. Introduction

The term *biomarker* generally refers to a measurable indicator of some biological state or condition (`https://en.wikipedia.org/wiki/Biomarker`). They have a wide range of applications in (bio)medical sciences. Once a biomarker is identified, it can be used as the basis for diagnosing a disease [16]. Frequently, biomarkers are used as surrogate measures of the studied characteristics, but they are also used with prognostic goals such as determining the susceptibility to respond to a treatment. Besides some biologi-

CONTACT: Pablo Martínez-Camblor. The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, 7 Lebanon Street, Suite 309, Hinman Box 7251, Hanover, NH 03755, USA. Email: Pablo.Martinez.Camblor@dartmouth.edu

cal desirable properties (usually reliability, interpretability, and feasibility [14]) a good biomarker should have a strong relationship with the presence/absence of the studied characteristic, i.e., good diagnostic capacity. That is, the biomarker value should enable us to know the real status of the subject with a small (as small as possible) error. There are two potential errors associated with binary decisions: to classify a subject within the positive group when it is actually in the negative group (the number of negative subjects classified as positive out of the total number of negative subjects is the so-called false-positive rate, FPR), and to classify a subject within the negative group when it is actually positive (the number of positive subjects classified as negative out of the total number of positive subjects is the so-called false-negative rate, FNR). A classification rule is a trade-off between these two potential mistakes. The receiver-operating characteristic, ROC, curve [19] is a graphical tool routinely used to represent the biomarker diagnostic capacity for all possible classification rules in the way $[c, \infty)$ with $c \in \mathbb{R}$. That is, by assuming (without loss of generality) that larger values of the biomarker are associated with having a higher probability of being positive, the ROC curve plots the sensitivity, S_E (i.e., the capacity of the biomarker for detecting positive subjects) against the complementary of the specificity, $1 - S_P$ (the specificity is the ability of the biomarker for detecting negative subjects) classifying as positive those subjects within the subset $[c, \infty)$ for each $c \in \mathbb{R}$. That means that the decision criteria (rules) used by the ROC curve are $\{[c, \infty)$ with $c \in \mathbb{R}\}$. Hence, the decision/classification based on a particular ROC curve lies with one of those particular rules.

Considering the case in which we want to know the gender of a subject based on acoustic properties of her/his voice and speech, we use the *Gender recognition* dataset freely available at `https://www.kaggle.com/primaryobjects/voicegender`. It consists of 3,168 recorded voice samples collected from male (50%) and female (50%) speakers. The voice samples were pre-processed by acoustic analysis using the `R` packages `seewave` and `tuneR`. An original analysis of this dataset can be found at `http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning`. Particularly, we are interested in the variable *Mode frequency*. Because the mean in the female group is larger than in the male group (0.178 ± 0.07 vs. 0.152 ± 0.08, respectively), female will be considered the positive group. Figure 1-A depicts the histograms for both the female (gray) and the male (black) groups. Figures 1-D and 1-B show the subsets determining that a subject is classified as positive (that is, female, in gray) for each $1 - S_P$ value and the resulting ROC curve, respectively.

In this case, the area under the ROC curve (AUC), frequently used for summarizing the diagnostic capacity by a single number [3], is 0.581 with a 95% confidence interval of [0.561, 0.602]. Henceforth the 95% confidence interval (based on 5,000 bootstrap replications) for the AUC of the different curves will be displayed. However, the curve presents some concavities. The existence of more than one relative maximum (point around which it is possible to find an interval where that point is a maximum) in the sum of the sensitivity and the specificity (Figure 1-C stands for the curve $\mathfrak{J}(c) = S_E(c) - 1 + S_P(c)$) suggests that the decision capacity could be improved by using more flexible decision criteria. That is, allowing to the decision rules be different to $[c, \infty)$ with $c \in \mathbb{R}$ (for instance, considering as a possibility those subsets in the way $(-\infty, a] \cup [b, \infty)$ with $a < b \in \mathbb{R}$).

Two main approaches have been proposed for improving the diagnostic capacity of a given biomarker. Arguing as in McIntosh and Pepe [15], Johnson [6] suggested to transform the biomarker values by using a suitable functional transformation which

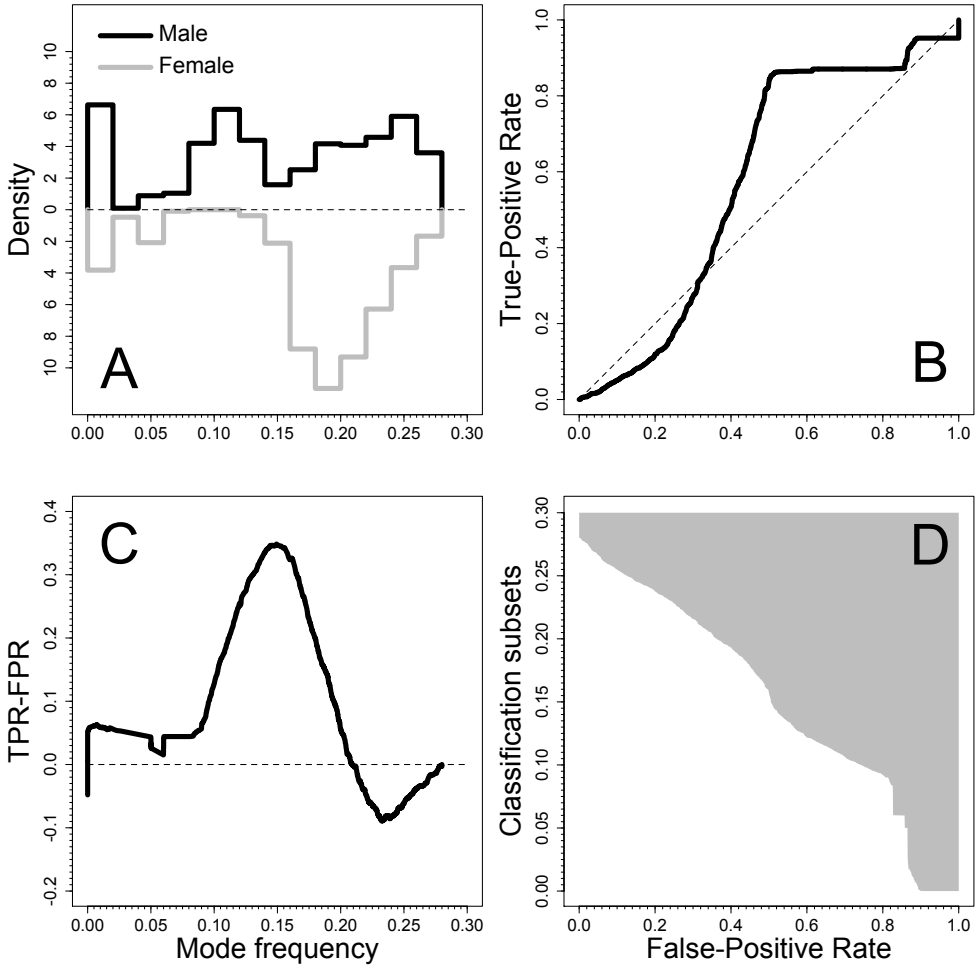

Figure 1. Gender recognition data. Top-left, histogram of the *Mode frequency* by gender. Top-right, receiver-operating characteristic, ROC, curve. Bottom-left, $\mathfrak{J}(c) = S_E(c) - 1 + S_P(c)$. Bottom-right, classification subsets (in gray, to classify as a female).

should enhance the biomarker accuracy in the AUC sense. Recently, Kauppi [7] defined the *efficient ROC curve* as the ROC curve resulting from the optimal (in the AUC sense) functional transformation of a given biomarker. In that paper, two different procedures (one semi-parametric, one non-parametric) to estimate the optimal transformation are proposed. No real-world examples are provided in those papers. On the other hand, Martínez-Camblor et al. [11] proposed to increase the performance of the biomarker by making more flexible the classification rules. They considered the problem in which both the highest and the lowest biomarker values are associated with having more probability of being positive. The parametric estimator of the resulting ROC curve (called *generalized ROC curve*, gROC) was considered by Martínez-Camblor and Pardo-Fernández [12]. For particular real problems, more customized criteria have been proposed. For instance, in the context of gene selection, Pepe et al. [17] proposed to use the partial area under the ROC curve while Li and Fine [8] considered a criterion based on a weighted area under the ROC curve.

It is worth to note that, when the biomarker is modelled by a univariate random variable, directly working on the classification regions we preserve the sense and interpretability of the classification process (white-box), while considering functional transformations provokes a loss of control and interpretability in the derived classification rules (black-box).

This paper study the capacity of improving the classification performance of univariate biomarkers via functional transformations and the impact of this transformation on the final classification regions based on a real-world dataset. Particularly, we considered a dataset regarding the identification of the individual gender based on his/her voice *Mode frequency*. Because of the specially intricate distributional shape of the characteristic considered, the problem allows us to revise the two different aforementioned approaches for improving the classification performance of a univariate biomarker. Our results suggest that, even with extremely pathological distributional shapes, in practice, the obtained improvement by a flexible functional transformation is not relevant, and indeed this procedure implies to deal with black-box (lack of practical interpretation) classification rules. Rest of the paper is organized as follows. In section 2, we consider the gROC curve and analyze the reported solution focused on the classification rules. Section 3 is devoted to find the optimal functional transformation. We revise the likelihood ratio argument and report some potential solutions. In section 4, we compare the classification rules reported by the different solutions. We conclude with a brief discussion about the topic in the section 5. As Supplementary Material we provide the `R` code used for the computations. Some technical proofs are included as Appendix.

2. gROC curve approach

Despite sometimes the ROC curve is described as a *classification tool*, it is actually just a graphical representation of the general diagnostic capacity of the considered biomarker. Underlying, there are certain *classification rules* which are the ones really determining the group in which each subject is going to be allocated. Particularly, each false-positive rate, t $(= 1 - S_P)$ value within the interval $[0, 1]$, has univocally associated one region, $r(t)$ $(\subseteq \mathbb{R})$, such that the ROC curve is

$$\mathcal{R}(t) = \mathcal{P}(\xi \in r(t)),$$

where ξ stands for the random variable representing the positive subjects. Notice that, if χ denotes the random variable representing the biomarker values for the negative subjects, $\mathcal{P}(\chi \in r(t)) = t$. In the standard ROC curve, assuming that higher values of the biomarker are associated with a greater likelihood of being positive, the regions are always in the form $r(t) = [c, \infty)$ ($c \in \mathbb{R}$). Therefore, if F_ξ and F_χ depict the cumulative distribution functions (CDF) for the positive and the negative subjects, respectively, then $c = F_\chi^{-1}(1-t)$ (with $G^{-1}(\cdot) = \inf\{y : G(y) \geq \cdot\}$) and $\mathcal{R}(t) = 1 - F_\xi(F_\chi^{-1}(1-t))$. Martínez-Camblor et al. [11] dealt with the case in which extreme values of the biomarker are associated with having a higher probability of being positive, that is, for each value of false-positive rate, t ($0 \leq t \leq 1$), they considered classification rules in the form $b(t) = (-\infty, u_L] \cup [u_U, \infty)$ with u_L, $u_U \in \mathbb{R}$ satisfying that $\mathcal{P}\{\chi \in b(t)\} \leq t$ (the probability of misclassifying a negative subject is, at most, t). This characterization is not unique. Usually, there exist infinite pairs, $\{u_L, u_U\}$ with $u_L, u_U \in \mathbb{R}$, satisfying $\mathcal{P}\{\chi \in b(t)\} \leq t$ so, the one verifying $b(t) = \arg\max_{s_t \in \mathcal{I}(t)} \mathcal{P}(\xi \in s_t)$ is taken, where $\mathcal{I}(t)$ is a class of subsets satisfying that, for each $s_t \in \mathcal{I}(t)$,

i) $s_t = (-\infty, u_L] \cup [u_U, \infty)$ with $u_L, u_U \in \mathbb{R}$,
ii) $\mathcal{P}\{\chi \in s_t\} \leq t$.

While condition *ii)* is essential for providing adequate classification rules, condition *i)* can be adapted to any particular classification problem. The shape of the selected subsets should be interpretable, but this point has to be solved for each particular situation. It is interesting that it can be applied to domains others than $\mathbb{R}$. With this definition, classifications based on, for instance, multidimensional or functional objects, among others, can be easily accommodated to the ROC curve context.

Distributions showed in Figure 1-A suggest that positive subjects (remember, female), are more likely to have medium values of *Mode frequency*. Hence, it seems adequate to base the classification on subsets in the form (u_L, u_U) with $u_L < u_U \in \mathbb{R}$. It is remarkable that the observed percentage of zeros was 4.8% in female and 10.1% in male. Therefore, we are always committing a classification error when these subjects are allocated within a group. Besides, there also exist other 14 values with ties between male and female.

Figure 4 shows the non-parametric estimation for the gROC curve (solid-gray line). Figure 6-A depicts the *classification subsets* on which the curve is based. The final gAUC (area under the gROC curve) is 0.741 [0.726, 0.758]. It is remarkable that, despite the fact that the gAUC can be seen as a measure (index) of the global diagnostic capacity for the set of rules on the basis of which the gROC curve is built, it does not share the standard AUC probabilistic interpretation ($\mathcal{P}(\chi < \xi)$). Under certain conditions, particularly that

(*C*) let $b(t_1)$ and $b(t_2)$ ($t_1 \leq t_2 \in [0,1]$) be the corresponding eligible subsets such that $\mathcal{R}(t_1) = \mathcal{P}(\xi \in b(t_1))$ and $\mathcal{R}(t_2) = \mathcal{P}(\xi \in b(t_2))$, then $b(t_1) \subseteq b(t_2)$,

Martínez-Camblor and Pardo-Fernández [12] proved that the AUC under continuous ROC curves defined as $\mathcal{R}(t) = \mathcal{P}(\xi \in b(t))$ ($0 \leq t \leq 1$) is *the probability of selecting randomly and independently two subjects, one negative and one positive, for which there exists one classification subset such that both subjects are correctly classified.* In any case, the AUC can always be interpreted as the average sensitivity (specificity) for all specificity (sensitivity) values and beyond the determined classification rules, it can always be used to compare the global classification accuracy of two different biomarkers.

3. Searching for a functional transformation

Following the celebrated Neyman-Pearson Lemma, McIntosh and Pepe [15] proved that the optimal classification rules can be obtained from binary regression procedures. Theoretically, let Y be the random variable representing the biomarker values (that is, the values of the *Mode frequency* in the population for both men and women) and D the binary random variable representing the disease status (in this case, an indicator of having the studied characteristic: to be a woman), with $\mathcal{P}(Y \leq c|D = 1) = F_\xi(c)$ and $\mathcal{P}(Y \leq c|D = 0) = F_\chi(c)$ $(c \in \mathbb{R})$, the Neyman-Pearson lemma guarantees that, for any α $(= t)$, the screening rule with the highest sensitivity based on Y, among all possible rules based on Y, is the likelihood ratio rule. That is,

$$c(\alpha) < \mathfrak{L}(Y) = \frac{\mathcal{P}(Y|D=1)}{\mathcal{P}(Y|D=0)} = \frac{f_\xi(y)}{f_\chi(y)}, \tag{1}$$

where $\alpha = \mathcal{P}(\mathfrak{L}(Y) > c(\alpha)|D = 0)$ and f_ξ and f_χ are the density functions of ξ and χ, respectively. Of course, last equality requires the existence of such density functions. Then, directly by the Bayes' rule,

$$\begin{aligned} p(Y) &= \mathcal{P}(D=1|Y) \\ &= \frac{\mathcal{P}(Y|D=1)\cdot\mathcal{P}(D=1)}{\mathcal{P}(Y|D=1)\cdot\mathcal{P}(D=1)+\mathcal{P}(Y|D=0)\cdot\mathcal{P}(D=0)} = \frac{\mathfrak{L}(Y)\cdot q}{1+\mathfrak{L}(Y)\cdot q}, \end{aligned} \tag{2}$$

where $q = \mathcal{P}(D = 1)/\mathcal{P}(D = 0)$. Therefore, the risk function, $p(Y)$, is a monotone increasing function of $\mathfrak{L}(Y)$ and then, equation (1) can be written as $p(Y) > c^*(\alpha)$, with $c^*(\alpha) = c(\alpha) \cdot q/(1 + c(\alpha) \cdot q)$. Summarizing, if $\text{logit}(p(Y)) = h(Y)$, where h is a real function, then $h(Y) > k$ $(= k(\alpha))$ with $k \in \mathbb{R}$, satisfying $\mathcal{P}(h(\chi) > k) \leq \alpha$, provides the optimal classification rules among those based on Y. Then, the problem is to estimate the function h. Besides, based on these results, it can be proved that, if both negative and positive subjects are normally distributed, the gROC curve is based on the optimal classification rules (see Appendix).

With no restrictions on the shape of h and based on a fixed sample of positive and negative subjects without ties between the positive and negative values, it is always possible to build a function which provides a perfect classification region. In the dataset under consideration, let $\{e_1, \cdots, e_{15}\}$ be the fifteen different values of the voice data which are taken for subjects in both the male and the female groups, and let $\{u_1, \cdots, u_{1457}\}$ and $\{v_1, \cdots, v_{1362}\}$ be the other 1457 and 1362 different values taken by subjects within the female and the male groups, respectively. Then, if we define the function

$$h_{of}(x) = \begin{cases} 1 & if \quad x = u_i \text{ for some } i \in \{1, \ldots, 1457\} \\ 0 & if \quad x = v_i \text{ for some } i \in \{1, \ldots, 1357\} \\ \dfrac{\textit{n. females taking the value } e_i}{\textit{n. subjects taking the value } e_i} & if \quad x = e_i \text{ for some } i \in \{1, \ldots, 15\} \end{cases} \tag{3}$$

The function $h_{of}(x)$ over those x out of the sample does not affect the empirical AUC so it can be defined, without loss of generality, as a linear interpolation of the points defined by the equation above. Classification based on this transformation is the optimal one in the AUC sense (see the Appendix section for a rigorous proof).

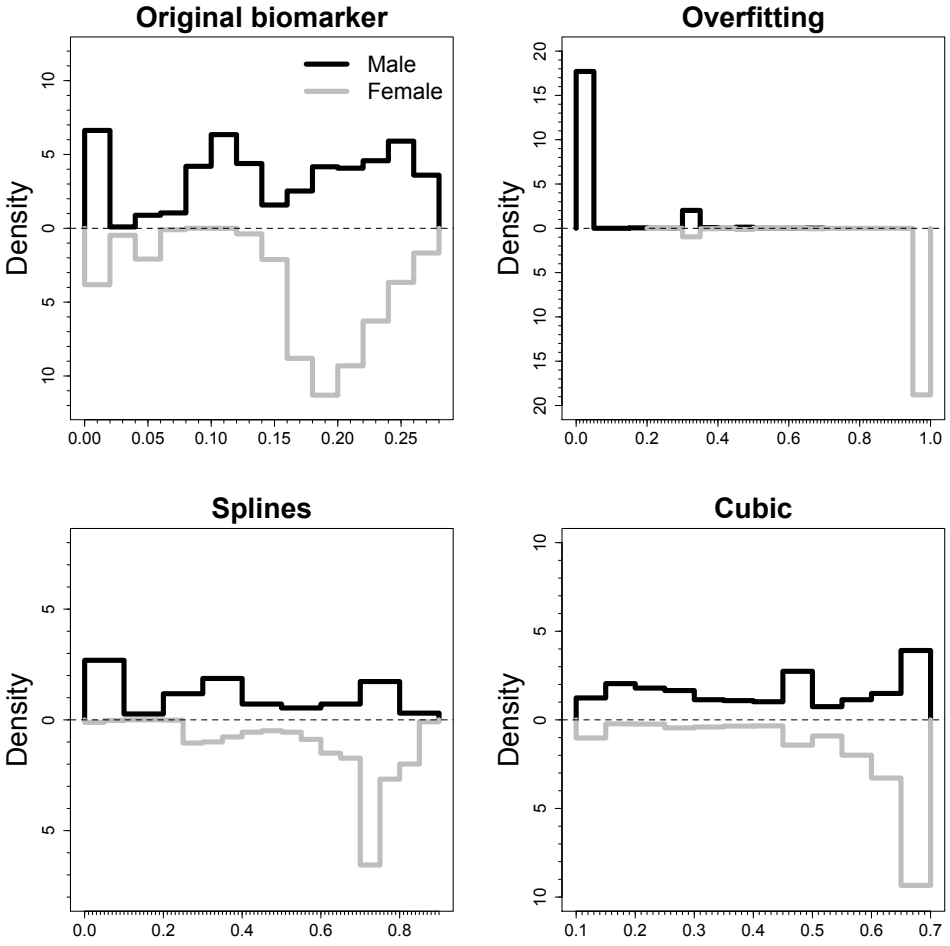

Figure 2. Histograms of different functional transformations of the *Mode frequency* by gender. Top-left, biomarker crude. Top-right, overfitted (h_{of}) curve. Bottom-left, 10 knots restricted cubic splines (h_{sp}) curve. Bottom-right, cubic (h_{cr}) function.

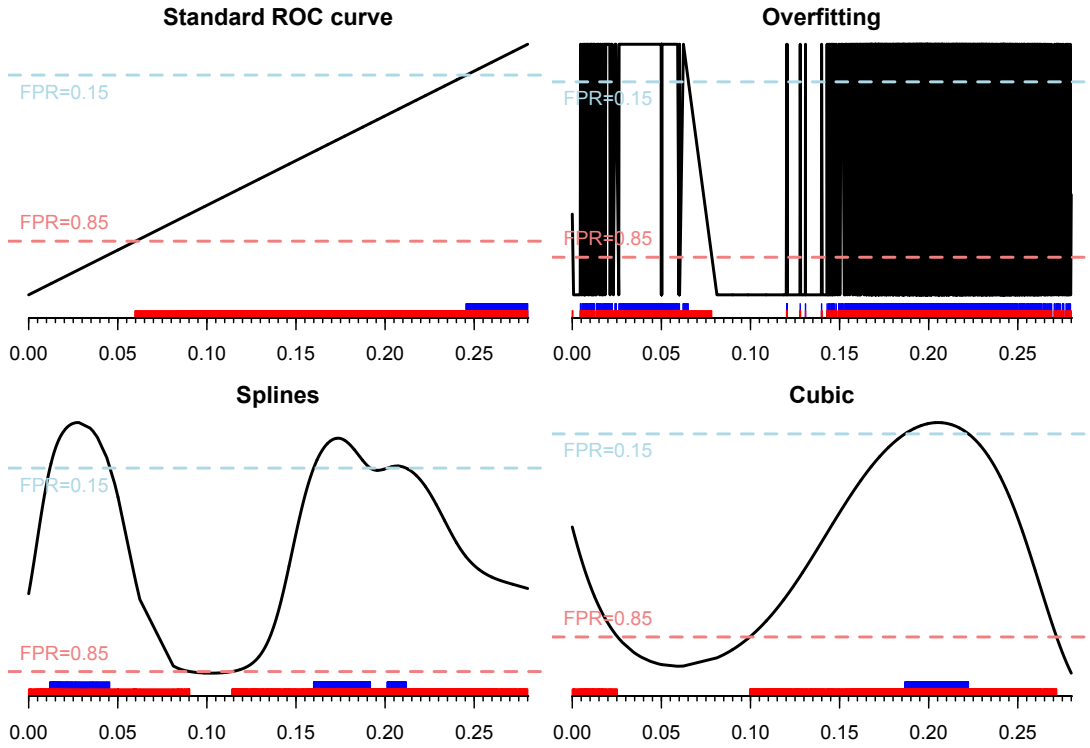

Figure 3. Different functional transformations and resulting classification subsets for false-positive rates of 0.15 and 0.85. Top-left, biomarker crude. Top-right, overfitted (h_{of}) curve. Bottom-left, 10 knots restricted cubic splines (h_{sp}) curve. Bottom-right, cubic (h_{cr}) function.

Figure 2 (top-right) depicts the histograms of the biomarker transformed by h_{of} for both the female (gray) and the male (black) groups. The AUC based on this transformation is 0.997 [0.996, 0.998]. Figure 4 (dashed-gray line) depicts the obtained ROC curve. However, regions employed to get these classifications are difficult to explain from a practical point of view and the overfitting is clear. Figure 3 (top-right) shows the function h_{of} and the regions of *Mode frequency* which provide false-positive rates (FPR) of 0.15 (blue) and 0.85 (red). Figure 6-B stands for the classification subsets on which the decisions based on $h_{of}(\cdot)$ are taken.

López-Ratón [9] proposed to estimate the function h by using general additive models (GAM). Particularly, she considered the use of *spline smoothing methods* to estimate $h(\cdot)$ from a logistic GAM regression model. Other tools such as *kernel* or *P-spline* smoothing techniques [2] are also recommended. All these procedures provide smooth functions highly adapted to the data shape. Their main handicap is – perhaps – the lack of a short (and usable) functional expression. For the voice-gender recognition data, we consider the implementation of the *logistic regression of the restricted cubic splines* [1] (natural splines) procedure included in the `R` package `rms` (Regression Modeling Strategies) developed by F.E. Harrell (version 5.1-2 uploaded to the `CRAN` on 2018-01-07). Figure 3 (bottom-left) depicts the estimated function when 10 knots

are considered,

$$h_{sp}(x) = \beta_0 + \beta_1 \cdot x + \sum_{j=2}^{9} \beta_j \cdot \left\{ \left[\frac{(x - x^*_{j-1})^3}{(x^*_{10} - x^*_1)^2} \right]_+ - \left[\frac{(x - x^*_9)^3}{(x^*_{10} - x^*_1)^2} \right]_+ \cdot \frac{x^*_{10} - x^*_{j-1}}{x^*_{10} - x^*_9} + \left[\frac{(x - x^*_{10})^3}{(x^*_{10} - x^*_1)^2} \right]_+ \cdot \frac{x^*_9 - x^*_{j-1}}{x^*_{10} - x^*_9} \right\},$$

where $[u]_+ = \max\{u, 0\}$, $\{x^*_1, \cdots, x^*_{10}\} = \{0.0007,\ 0.0153,\ 0.1038,\ 0.1475,\ 0.1752,\ 0.1919,\ 0.2054,\ 0.2225,\ 0.2424,\ 0.2708\}$ is the vector of knots and $\{\beta_0, \cdots, \beta_9\} = \{$-0.9411, 165.99, -7682.35, 9927.90, -4853.82, 456.03, 11816.26, -17863.90, 8858.09, 1809.01$\}$ is the vector of coefficients. Figure 2 (bottom-left) shows the histograms of the biomarker transformed by this function for both the female (gray) and the male (black) groups. The resulting function reflects the difference in the monotony observed for $\mathfrak{J}(\cdot)$ (see Figure 1). The ROC curve based on $h_{sp}(\cdot)$ (solid-black line in Figure 4) achieves an AUC of 0.786 [0.772, 0.803]. Figure 6-C depicts the classification subsets which lead to the ROC curve based on the $h_{sp}(\cdot)$ transformation.

Anyway, the most direct solution to estimate the function h is to consider logistic polynomial regression models. That is, to fix a degree for the polynomial, g, and to estimate the parameters of the model

$$p(Y) = \frac{\exp\{\beta_0 + \beta_1 \cdot Y + \cdots + \beta_g \cdot Y^g\}}{1 + \exp\{\beta_0 + \beta_1 \cdot Y + \cdots + \beta_g \cdot Y^g\}},$$

where the coefficients β_i $(0 \leq i \leq g)$ can be directly estimated by using the maximum likelihood procedure. The main handicap is to select the polynomial degree and the evident risk of overfitting for too large g-values. Based on the shape of $h_{sp}(\cdot)$, we consider a logistic cubic regression, $g = 3$, obtaining the model

$$p(Y) = \frac{1}{1 + \exp\{0.188 + 65.4 \cdot Y - 710.6 \cdot Y^2 + 1790.8 \cdot Y^3\}}.$$

It can be directly computed by making dummy variables $Y_2 = Y^2$ and $Y_3 = Y^3$ and including the three variables (Y, Y_2 and Y_3) in a standard multivariate logistic regression procedure (see the R code provided as supplementary material). Figure 3 (bottom-right) stands for the estimated function,

$$h_{cr}(x) = -0.188 - 65.4x + 710.6x^2 - 1790.8x^3,$$

and the corresponding classification subsets for FPR of 0.15 (blue) and 0.85 (red); while the histograms of the biomarker transformed by h_{cr} for both the female (gray) and the male (black) groups are displayed in Figure 2 (bottom-right). The ROC curve based on the transformed biomarker achieved an AUC of 0.725 [0.707, 0.743]. Figure 4 (dashed-black line) depicts this ROC curve.

While it seems clear that the $h_{of}(\cdot)$ is too close to the fixed sample and the overfitting is immediate, both smooth (restricted cubic splines, in this case) and polynomial logistic regression provide reasonable and flexible estimates for a functional transformation. Of course, it is difficult to know if there exists any other theoretical function which improves the diagnostic capacity of the biomarker. However, to consider

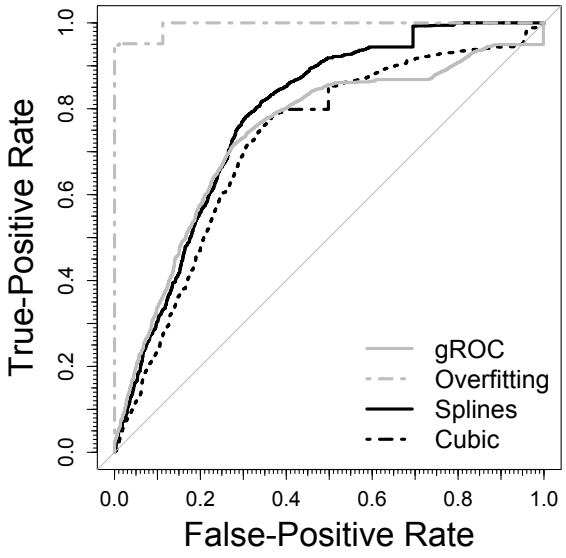

Figure 4. Solid-gray line, gROC curve. Dashed-gray line, ROC curve based on the $h_{of}(\cdot)$ transformation. Solid-black line, ROC curve based on the $h_{sp}(\cdot)$. Dashed-black line, ROC curve based on the $h_{cr}(\cdot)$ transformation.

restricted cubic splines estimator is a consistent method to get this objective. Top panels of Figure 5 depict, in gray, the functions obtained from 100 bootstrap samples when both restricted cubic splines and cubic polynomial logistic regression were used for estimating the functions. Bottom panels of Figure 5 show these functions when the considered sample size was the 25% of the original one.

4. Classification rules

As it is well-known, the ROC curve plots the sensitivity against the complementary of the specificity over the unit interval. It reports visually the global diagnostic capacity of a studied biomarker while the AUC summarizes in a single number this information. However, the process of allocating a subject within a group given a biomarker value requires to know if this value is within or outside one particular region. Once the minimum required specificity (or sensitivity) is fixed (frequently, the point leading the Youden index [18] is taken, i.e., $\arg\max_{c\in\mathbb{R}} \mathfrak{J}(c)$), in order to make the decision, we need to know explicit values for this region. Besides, in some particular problems, these regions can help to understand the problem and to make recommendations (for instance, for determining the adequate values of serum phosphorus, calcium or PTH which minimize the mortality risk in dialysis patients [4]).

Based directly on the *Mode frequency*, for a fixed specificity of 0.85, the region determining that a subject is female is the interval $[0.245, 0.28]$ (0.28 is the maximum *Mode frequency* value observed in the dataset) which reaches a sensitivity of 0.079. When we consider extreme values of *Mode frequency* as male, for the same specificity the region determining that a subject is female is $(0.156, 0.204)$, which reaches a sensitivity of 0.462. By using a three-degree polynomial logistic regression, $h_{cr}(\cdot)$, the resulting region is $(0.187, 0.222)$ with a sensitivity of 0.365. With more flexibility, by using the restrictive cubic splines transformation, $h_{sp}(\cdot)$, the region is in the form

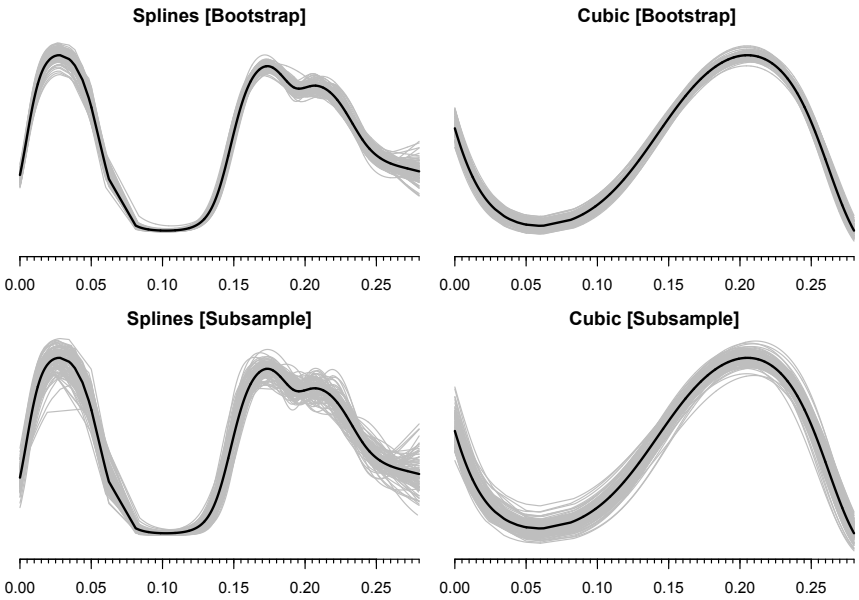

Figure 5. Top panels, in gray, functions estimated from 100 bootstrap samples by using restricted cubic splines (left) and cubic logistic regression (right). Bottom panels, identical plots but considering 25% of the values of the original sample.

$(0.012, 0.045) \cup (0.160, 0.192) \cup (0.201, 0.212)$ and reports a sensitivity of 0.417. Figure 6 depicts the classification subsets for the different considered criteria.

5. Discussion

Both the receiver-operating characteristic, ROC, curve and the area under the ROC curve, AUC, have become so popular that, frequently, the diagnostic capacity of a biomarker is directly identified with its AUC. Getting a large AUC has become an objective by itself. In this context, finding an adequate functional transformation of the biomarker values that reaches an AUC as large as possible seems to be a potential procedure to improve the biomarker capacity. However, decisions are made based on some particular regions (classification subsets) and their practical interpretation is frequently helpful for a better understanding of the problem and used for making recommendations (white–box). The optimal function can report the largest AUC (see, for instance, results reported by function $h_{of}(\cdot)$) but having associated classification regions which make difficult its interpretation (black-box). Even though the biomarker behavior should be different in each group in order to be used for diagnostic tasks, a different behavior does not guarantee its use in such problems [10].

In this paper we study the capacity of univariate biomarkers to improve their classification performance via functional transformations and the impact of this transformation on the final classification regions based on a real-world dataset. Particularly, we based our research on the problem of determining the gender of a subject given some feature of his/her voice, particularly, the *Mode frequency*. The distribution functions of this characteristic in both male and female groups are complex: in both the CDF has

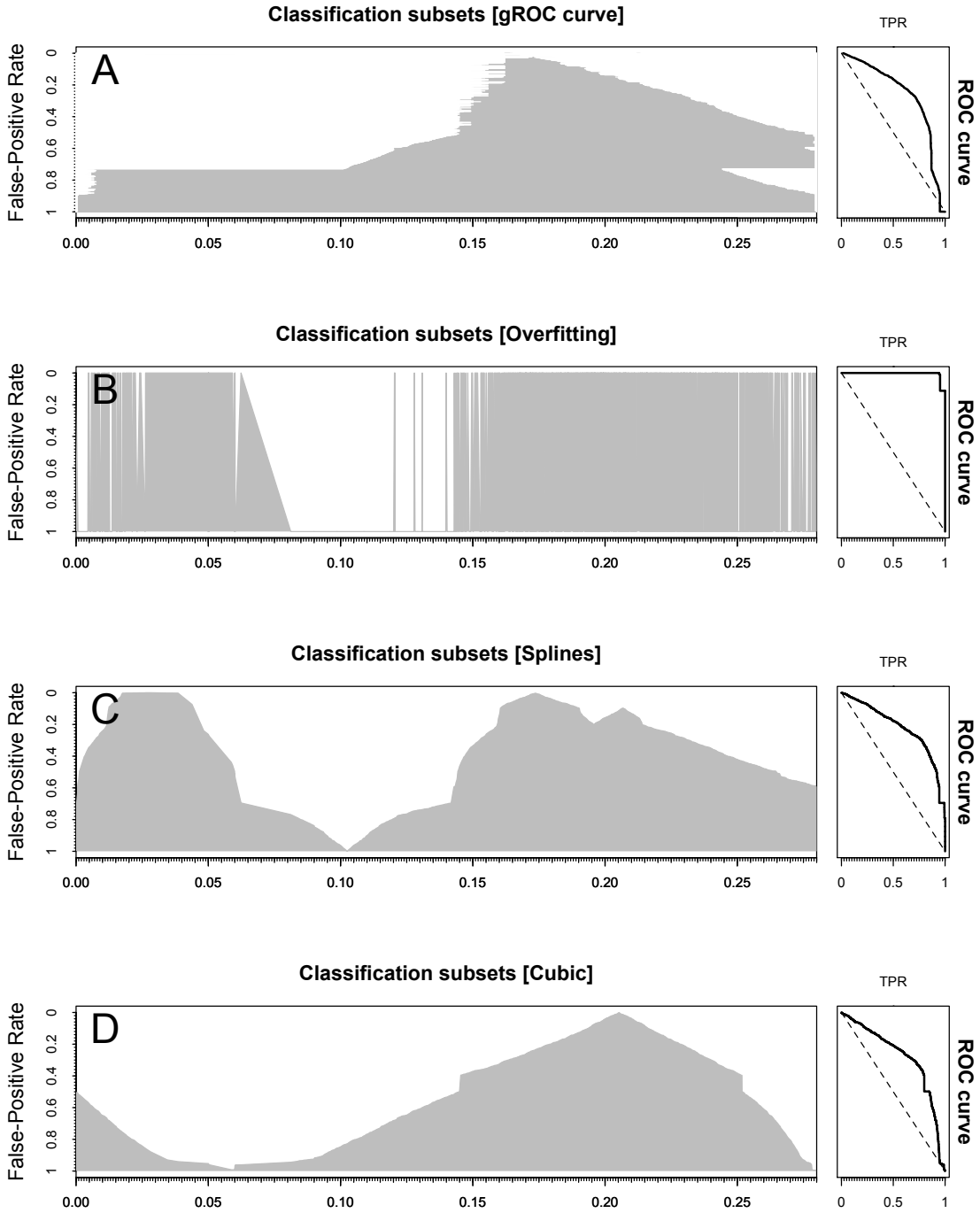

Figure 6. Classification subsets for the gROC curve (A), for the biomarker with the overfitted transformation, $h_{of}(\cdot)$ (B), the restrictive cubic splines transformation, $h_{sp}(\cdot)$ (C) and the three-degree polynomial logistic regression, $h_{cr}(\cdot)$ (D).

a non-negligible discontinuity at zero and both densities have two additional modes at different points (Figure 1-A). We can say that the behavior of the *Mode frequency* is different in male and female. However, its use with classification purpose reports a poor AUC of 0.581 [0.561, 0.602]. The reason seems to be that, although the *Mode frequency* mean is larger in female than in male group, intervals in the way $[c, \infty)$ $(c \in \mathbb{R})$ are not adequate for reflecting the CDFs complexity. A rapid overview suggests that the *Mode frequency* in male is more likely to take extreme values of the spectrum. Considering classification subsets in the form (a, b) $(a \leq b \in \mathbb{R})$ according to the so-called gROC curve approach [11, 12] improves the global diagnostic capacity and gets an AUC of 0.741 [0.726, 0.758]. McIntosh and Pepe [15] proved that the optimal classification criterion among those based on one particular biomarker can be computed from a binary regression with an adequate functional transformation of the biomarker. This argument has already been exploited [6, 7, 9]. It is worth noting that no real-world examples were reported in those papers. In addition, the estimation of this function has several drawbacks. Some restrictions must be imposed in order to avoid overfitting (Figure 6-B). Based on the Gender recognition data, we mainly consider a three-degree polynomial transformation and a cubic restricted splines (with 10 knots). Those report AUCs of 0.725 [0.707, 0.743] and 0.786 [0.772, 0.803], respectively. Besides, both methods seem to be consistent and report similar resulting functions when different resampling procedures were considered (Figure 5). Their main handicap is the interpretation of the associated classification subsets which, in this case, sometimes result in a union of two or even three intervals. Depending on the considered problem, these regions are not realistic and, in any case, each particular problem should be cautiously studied in order to provide a reasonable explanation.

Of course, one can design artificial distributions on which the functional transformation clearly improves the results obtained by the gROC curve. Besides the splines procedure has proved its capacity for finding this underlying pattern (see, for instance, Hansen and Kooperberg [5]). However, this kind of distributions are rarely found in actual practice. In this work, we show that even when the shapes of the distributions are specially intricate, functional transformations get a slight improvement in terms of the area under the curve.

In practice, a flexible binary regression can be used in order to get the optimal (or close to the optimal) AUC based on one particular biomarker. Then, additional restrictions can be added to this regression (for instance reducing the number of knots in a penalized cubic splines procedure) and weighing the simplicity of the models and the improvement in the AUC one can decide what is the adequate criteria for each model. We also encourage the reader to consider and to prioritize the gROC curve if the difference between the derived AUCs is not relevant. This curve produces direct and easy to interpret classification rules.

In short, particular polynomial binary regression and, preferably, some binary flexible functional regression (such as restricted cubic splines) provides a valuable tool to know the limits in the global classification capacity of a studied marker. When the resulting classification subsets lack practical interpretation, additional restrictions on the function can be imposed to simplify their shape. Furthermore, the gROC curve approach provides the optimal AUC when extreme values of the biomarker are associated with having a higher probability of being positive (or negative). The fact that in the considered Gender recognition dataset, with extremely atypical CDFs for the *Mode frequency* in both the male and the female groups, we found only a small difference between the gAUC (area under the gROC curve) and the AUC provided for the $h_{sp}(\cdot)$ (just an improvement of 6.1%) suggests that, in most real situations, considering

the gROC curve would be enough to get both simple classification regions and a good diagnostic capacity.

Supplementary Material

As supplementary material of this paper we provide the `R` code used for computing plots and models reported herein. `R` packages `nsROC` (developed by Sonia Pérez-Fernández) and `rms` (developed by Frank E. Harrel Jr.) are required. The used Gender-recognition dataset is freely available at `https://www.kaggle.com/primaryobjects/voicegender`.

Appendix: Technical issues

Theorem. *If the studied biomarker is normally distributed for both positive and negative groups, then the classification rules on which the gROC curve is based are the optimal ones among those based on this biomarker.*

Proof. Notice that if ξ and χ (random variables representing the values of the biomarker in the positive and the negative group, respectively) follow $\mathscr{N}(\mu_\xi, \sigma_\xi)$ and $\mathscr{N}(\mu_\chi, \sigma_\chi)$ distributions, respectively ($\mathscr{N}(\mu, \sigma)$ stands for a normal distribution with mean μ and standard deviation σ) the ROC curve derived from ξ and χ is equivalent to the one derived from $\xi^* = (\xi - \mu_\chi)/\sigma_\chi$ and $\chi^* = (\chi - \mu_\chi)/\sigma_\chi$. Thus ξ^* follows a $\mathscr{N}(a, b)$ with $a = (\mu_\xi - \mu_\chi)/\sigma_\chi$ and $b = \sigma_\xi/\sigma_\chi$ and χ^* follows a $\mathscr{N}(0, 1)$.

Martínez-Camblor and Pardo-Fernández [12] proved that, assuming that extreme values of the biomarker are associated with having more probability of being positive (that implies $b \geq 1$), for each $t \in [0, 1]$, the gROC curve is based on the subsets

$$b^*(t) = \left(-\infty, \Phi^{-1}(\gamma_t \cdot t)\right] \cup \left[\Phi^{-1}(1 - t + \gamma_t \cdot t), \infty\right), \tag{4}$$

where Φ stands for the CDF of a standard normal distribution and γ_t is the solution to the equation

$$\gamma_t = \frac{1}{t} \cdot \Phi\left(\frac{2 \cdot a}{1 - b^2} - \Phi^{-1}(1 - t + \gamma_t \cdot t)\right), \tag{5}$$

Notice that if $\sigma_\xi = \sigma_\chi$ ($b = 1$), then the equation (5) does not have a real solution and γ_t falls on the extremes, $\gamma_t = 0$ for $a > 0$ ($b^*(t) = [\Phi^{-1}(1 - t), \infty)$) and $\gamma_t = 1$ for $a < 0$ ($b^*(t) = (-\infty, \Phi^{-1}(t)]$).

Equation (1) implies that, for the considered context, assuming $\sigma_\xi > \sigma_\chi$, the optimal classification rules are given by

$$\begin{aligned} \log(c(\alpha)) < & \log(\mathfrak{L}(Y)) \\ & = -\log(b) + \frac{b^2 - 1}{2 \cdot b^2} \cdot \left\{\left(x - \frac{a}{1 - b^2}\right)^2 - \frac{a^2 \cdot b^2}{(1 - b^2)^2}\right\}. \end{aligned} \tag{6}$$

Equivalently, the classification rules based on the transformation $h(x) = (x - a/(1 - b^2))^2$ are optimal among those based on the considered biomarker. Notice that if

$\sigma_\xi = \sigma_\chi$, the quadratic factor is cancelled, then: if $a > 0$, no transformation is needed; if $a < 0$, just a change of sign is required. In these cases, the classification subsets are the same as the ones employed in the gROC curve. For each $t \in [0, 1]$ the classification subset is

$$a^*(t) = (-\infty, -\sqrt{c} + a/(1-b^2)] \cup [\sqrt{c} + a/(1-b^2), \infty), \tag{7}$$

with $\mathcal{P}(h(\chi) \leq c) = 1 - t$, that is

$$1 - t = \Phi\left(\sqrt{c} + \frac{a}{1-b^2}\right) - \Phi\left(-\sqrt{c} + \frac{a}{1-b^2}\right).$$

Making

$$\Phi\left(-\sqrt{c} + \frac{a}{1-b^2}\right) = \gamma \cdot t,$$

then

$$\begin{aligned} -\sqrt{c} =& \Phi^{-1}(\gamma \cdot t) - a/(1-b^2) \text{ and,} \\ \sqrt{c} =& \Phi^{-1}(1 - t + \gamma \cdot t) - a/(1-b^2). \end{aligned}$$

Directly, $a^*(t) = \left(-\infty, \Phi^{-1}(\gamma \cdot t)\right] \cup \left[\Phi^{-1}(1 - t + \gamma \cdot t), \infty\right)$ where γ is the solution to the equation,

$$\gamma = \frac{1}{t} \cdot \Phi\left(\frac{2 \cdot a}{1-b^2} - \Phi^{-1}(1 - t + \gamma \cdot t)\right). \tag{8}$$

Therefore, $a^*(t) = b^*(t)$ for each $t \in [0, 1]$ and both gROC and optimal ROC curves based on the biomarker defined by ξ^* and χ^* are the same. Directly, these curves are the same as the ones based on ξ and χ, so the theorem is proved. □

It is remarkable that the result does not need the normality of ξ and χ but just that there exists a monotone transformation, H, such that both $H(\xi)$ and $H(\chi)$ are normally distributed.

Next result proves the transformation which gets the optimal AUC in two finite populations of positive and negative subjects is the one described in the main manuscript. This results can be applied on empirical samples.

Proposition. *Let $\{x_i, \cdots x_n\}$ and $\{y_1, \cdots, y_m\}$ be populations of positive and negative subjects, respectively. Let $\{u_1, \cdots, u_{d_n}\}$ and $\{v_1, \cdots, v_{d_m}\}$ be the values taken just by the positive or the negative subjects, respectively, and let $\{e_1, \cdots, e_k\}$ be those values taken by both the positive and the negative subjects with frequencies $n_1, \cdots, n_k$ and*

$m_1, \cdots, m_k$, *respectively. Then, the transformation*

$$h_{op}(x) = \begin{cases} 1 \text{ if } x = u_i \text{ for some } 1 \leq i \leq d_n \\ 0 \text{ if } x = v_j \text{ for some } 1 \leq j \leq d_m \\ \dfrac{n_l}{n_l + m_l} \text{ if } x = e_l \text{ for some } 1 \leq l \leq k \end{cases}$$

gets the largest AUC among all possible transformations.

Proof. Let I_0^n and I_0^m be the set of indices whose original values are included in those sets. That is $i \in I_0^n$ if $x_i = u_j$ for some $1 \leq j \leq d_n$ and $j \in I_0^m$ if $y_j = v_i$ for some $1 \leq i \leq d_m$. Notice that if n_0 and m_0 are the cardinality of I_0^n and I_0^m, respectively, then $n = n_0 + n_1 + \cdots + n_k$ and $m = m_0 + m_1 + \cdots + m_k$.

Given any transformation, h, we have that the AUC is defined [13] by

$$\begin{aligned} AUC(h) =& \frac{1}{nm} \sum_{i \in I_0^n} \sum_{j \in I_0^m} W(h(y_j), h(x_i)) + \frac{1}{nm} \sum_{i \in I_0^n} \sum_{l=1}^{k} m_l W(h(e_l), h(x_i)) \\ &+ \frac{1}{nm} \sum_{s=1}^{k} n_s \sum_{j \in I_0^m} W(h(y_j), h(e_s)) + \frac{1}{nm} \sum_{s=1}^{k} n_s \sum_{l=1}^{k} m_l W(h(e_l), h(e_s)) \\ =& \frac{1}{nm} \left(S_1 + S_2 + S_3 + S_4 \right), \end{aligned} \tag{9}$$

where if $I(A)$ takes the value 1 if A is true and 0 otherwise, for $a, b \in \mathbb{R}$, $W(a, b) = I(a < b) + 1/2 \cdot I(a = b)$.

S_1, S_2 and S_3 are simultaneously maximized if h satisfies

(C_1) $h(v_j) < h(e_l) < h(u_i)$ for $1 \leq j \leq d_m$, $1 \leq l \leq k$ and $1 \leq i \leq d_n$.

In addition, S_4 is not affected by this condition, only by the order of $h(e_1), \ldots, h(e_k)$, then we can define

$$h_{op}(v_j) = 0 \text{ for all } 1 \leq j \leq d_m \quad \text{and} \quad h_{op}(u_i) = 1 \text{ for all } 1 \leq i \leq d_n \tag{10}$$

without loss of generality provided that $h(e_l) \in (0, 1) \ \forall 1 \leq l \leq k$.

In order to maximize S_4 what we should find is the order of $h(e_1), \ldots, h(e_k)$ which optimizes the following sum depending on the frequencies n_l and m_l:

$$S_4 = \frac{1}{nm} \sum_{s=1}^{k} n_s \sum_{l=1}^{k} m_l I(h(e_l) < h(e_s)) + \frac{1}{2nm} \sum_{s=1}^{k} n_s \sum_{l=1}^{k} m_l I(h(e_l) = h(e_s)). \tag{11}$$

McIntosh and Pepe [15] proved that the optimal h is the likelihood ratio rule (or a monotone increasing function of it), which in the empirical case over $\{e_1, \cdots, e_k\}$ is

$$h_{op}(e_l) = \frac{\mathcal{P}(e_l | D = 1)}{\mathcal{P}(e_l | D = 0)} = \frac{n_l / n}{m_l / m} = \frac{m}{n} \cdot \frac{n_l}{m_l}.$$

Then, the transformation maximizing the term S_4 satisfies

(C_2) $h(e_l) = g\left(\frac{m}{n} \cdot \frac{n_l}{m_l}\right)$ with g any monotone strictly increasing function.

Since m/n is only a positive constant, the order induced by n_l/m_l is the one leading to optimize equation (11). In order to satisfy (C_1) and definition in (10) we can define

$$h_{op}(e_l) = \frac{n_l}{n_l + m_l} \text{ for } 1 \leq l \leq k \tag{12}$$

because the order of induced by $n_l/(n_l + m_l)$ is that induced by n_l/m_l but the values are in the unit interval.

Notice that, since (C_2) does not affect (C_1) those functions which satisfy both simultaneously maximize S_1, S_2, S_3 and S_4. Since h_{op} defined by (10) and (12) satisfies them it gets the optimal AUC. □

It is remarkable that the result assigns the same h to those e_l with the same rate n_l/m_l. Even if regarding definition of $W(a, b)$ it seems better to make them different but consecutive since the second term is multiplied by $1/2$, it can be seen that the gain in AUC for being different is lost through the loss in number of ties $(h(e_l), h(e_s))$. Below we can see a proof of this fact:

Proof. If $n_i/m_i = c$ for some $i \in I^* := \{i_1^*, \ldots, i_{k^*}^*\} \subseteq \{1, \ldots, k\}$, the loss in AUC

- if $h(e_{i_1^*}) = \cdots = h(e_{i_{k^*}^*})$ is

$$\frac{1}{2nm} \sum_{i \in I^*} n_i \cdot \sum_{i \in I^*} m_i = \frac{1}{2nm} \sum_{i \in I^*} n_i m_i + \frac{1}{2nm} \sum_{i \in I^*} \left(n_i \cdot \sum_{\substack{j \in I^* \\ j \neq i}} m_j \right) \tag{13}$$

- if $h(e_{i_1^*}) < \cdots < h(e_{i_{k^*}^*})$ but consecutive is

$$\frac{1}{2nm} \sum_{i \in I^*} n_i m_i + \frac{1}{nm} \sum_{i \in I^*} \left(n_i \cdot \sum_{\substack{j \in I^* \\ j < i}} m_j \right) \tag{14}$$

Then it has to be proven that (13) = (14):

$$(13) = (14) \iff \frac{1}{2nm} \sum_{i \in I^*} \left(n_i \cdot \sum_{\substack{j \in I^* \\ j \neq i}} m_j \right) = \frac{1}{nm} \sum_{i \in I^*} \left(n_i \cdot \sum_{\substack{j \in I^* \\ j < i}} m_j \right)$$

$$\iff \sum_{i \in I^*} \left(n_i \cdot \sum_{\substack{j \in I^* \\ j > i}} m_j \right) = \sum_{i \in I^*} \left(n_i \cdot \sum_{\substack{j \in I^* \\ j < i}} m_j \right)$$

Since $n_i/m_i = c$ for all $i \in I^*$, then substituting n_i by cm_i:

$$(13) = (14) \iff \sum_{i\in I^*}\left(m_i \cdot \sum_{\substack{j\in I^*\\ j>i}} m_j\right) = \sum_{i\in I^*}\left(m_i \cdot \sum_{\substack{j\in I^*\\ j<i}} m_j\right)$$

But

$$\sum_{i\in I^*}\left(m_i \cdot \sum_{\substack{j\in I^*\\ j>i}} m_j\right) - \sum_{i\in I^*}\left(m_i \cdot \sum_{\substack{j\in I^*\\ j<i}} m_j\right) = \sum_{i\in I^*}\left(m_i \cdot \left(\sum_{\substack{j\in I^*\\ j>i}} m_j - \sum_{\substack{j\in I^*\\ j<i}} m_j\right)\right)$$

$$= \sum_{i\in I^*}\left(m_i \cdot \sum_{\substack{j\in I^*\\ j>i}} m_j\right) - \sum_{i\in I^*}\left(m_i \cdot \sum_{\substack{j\in I^*\\ j>i}} m_j\right) = 0.$$

□

Acknowledgement

This work is supported by the Grants MTM2015-63971-P, MTM2014-55966-P and MTM2017-89422-P (ERDF support included) from the Spanish Ministerio of Economía y Competitividad, FC-15-GRUPIN14-101 from the Asturies Government and Severo Ochoa Grant BP16118 (this one for S. Pérez-Fernández).

References

[1] T. Devlin and B. Weeks, *Spline functions for logistic regression modeling* (1986), pp. 646–651.

[2] P. Du and L. Tang, *Transformation-invariant and nonparametric monotone smooth estimation of ROC curves*, Statistics in Medicine 28 (2009), pp. 349–359.

[3] D. Faraggi and B. Reiser, *Estimation of the area under the ROC curve*, Statistics in Medicine 21 (2002), pp. 3093–3106.

[4] J. Fernández-Martín, P. Martínez-Camblor, M. Dionisi, J. Floege, M. Ketteler, G. London, F. Locatelli, J. Gorriz, B. Rutkowski, A. Ferreira, W.J. Bos, A. Covic, M. Rodríguez-García, J. Sánchez, D. Rodríguez-Puyol, and J. Cannata-Andia, *Improvement of mineral and bone metabolism markers is associated with better survival in haemodialysis patients: the COSMOS study*, Nephrology Dialysis Transplantation 30 (2015), pp. 1542–1551.

[5] M. Hansen and C. Kooperberg, *Spline adaptation in extended linear models (with comments and a rejoinder by the authors*, Statistical Science 17 (2002), pp. 2–51.

[6] N. Johnson, *Advantages to transforming the receiver operating characteristic (ROC) curve into likelihood ratio co-ordinates*, Statistics in Medicine 23 (2004), pp. 2257–2266.

[7] H. Kauppi, *The generalized receiver operating characteristic curve*, Aboa Centre for Economics, Discussion paper 114 (2016), pp. 1–50.

[8] J. Li and J. Fine, *Weighted area under the receiver operating characteristic curve and its application to gene selection*, Journal of the Royal Statistical Society: Series C (Applied Statistics) 59 (2010), pp. 673–692.

[9] M. López-Ratón, *Optimal cutoff points for classification in diagnostic studies: new contributions and software developments*, PhD Dissertation, Universidade de Santiago de Compostela, Galicia, Spain, 2016.

[10] P. Martínez-Camblor, *On the paper: 'Notes on the overlap measure as an alternative to the Youden index: How are they related?'*, Statistics in Medicine 37 (2018), pp. 1222–1224.

[11] P. Martínez-Camblor, N. Corral, C. Rey, J. Pascual, and E. Cernuda-Morollón, *Receiver operating characteristic curve generalization for non-monotone relationships*, Statistical Methods in Medical Research 26 (2017), pp. 113–123.

[12] P. Martínez-Camblor and J. Pardo-Fernández, *Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships*, Statistical Methods in Medical Research Accepted (2017), pp. 1–20.

[13] P. Martínez-Camblor, *Area under the ROC curve comparison in the presence of missing data*, Journal of the Korean Statistical Society 42 (2013), pp. 431 – 442.

[14] R. Mayeux, *Biomarkers: Potential uses and limitations*, NeuroRX 1 (2004), pp. 182–188.

[15] M. McIntosh and M. Pepe, *Combining several screening tests: optimality of the risk score*, Biometrics 58 (2002), pp. 657–664. Available at `http://dx.doi.org/10.1111/j.0006-341X.2002.00657.x`.

[16] M. Pepe, R. Etzioni, Z. Feng, J.D. Potter, M. Thompson, M. Thornquist, M. Winget, and Y. Yasui, *Phases of biomarker development for early detection of cancer*, JNCI: Journal of the National Cancer Institute 93 (2001), pp. 1054–1061.

[17] M. Pepe, G. Longton, G. Anderson, and M. Schummer, *Selecting differentially expressed genes from microarray experiments*, Biometrics 59 (2003), pp. 133–142.

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

[19] X. Zhou, N. Obuchowski, and D. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley Blackwell, New York, 2002.