HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2018 October 30.

Published in final edited form as:

Stat Med. 2017 October 30; 36(24): 3830–3843. doi:10.1002/sim.7394.

Estimation of Smooth ROC Curves for Biomarkers With Limits of Detection

Leonidas E. Bantis[*,1], **Qingxiang Yan**[1], **John V. Tsimikas**[2], and **Ziding Feng**[1]

[1]Dept of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, U.S.A

[2]Dept of Mathematics, Division of Statistics and Data Analysis, University of the Aegean, Samos, 83200, Greece

Abstract

Protein biomarkers found in plasma are commonly used for cancer screening and early detection. Measurements obtained by such markers are often based on different assays that may not support detection of accurate measurements due to a limit of detection (LOD). The *ROC* curve is the most popular statistical tool for the evaluation of a continuous biomarker. However, in situations where LODs exist, the empirical *ROC* curve fails to provide a valid estimate for the whole spectrum of the false positive rate (FPR). Hence, crucial information regarding the performance of the marker in high sensitivity and/or high specificity values is not revealed. In this paper, we address this problem and propose methods for constructing *ROC* curve estimates for all possible *FPR* values. We explore flexible parametric methods, transformations to normality, and robust kernel-based and spline-based approaches. We evaluate our methods though simulations and illustrate them in colorectal and pancreatic cancer data.

Keywords

Biomarker; Box-Cox; Cancer; Censoring; Classification; Early detection; Generalized Gamma; Kernels; Limit of detection; ROC; Spline

1. Introduction

At the preclinical level, performing a research-grade biomarker assay might involve in vitro experiments and mass spectrometry techniques, among other procedures, with the objective of identifying genes or proteins that are over- or under-expressed in tumor tissue compared to control tissue (see [1]). Promising biomarkers obtained during that first phase are further

[*]Correspondence to: Leonidas E. Bantis, 1400 Pressler St. Pickens Tower, Dept. of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, U.S.A. lebantis@mdanderson.org.

[†]Please ensure that you use the most up to date class file, available from the SIM Home Page at www.interscience.wiley.com/jpages/0277-6715

Supplementary Material: Regarding the implementation of our approaches a user friendly software is available online by the first author's website with detailed instructions and a reproducible example at: www.leobantis.net. All simulations contained in the Web Appendix are available by the journal's website. All simulations presented in the supplementary material are also provided in a unified excel file for convenience.

explored using a clinical assay, which is usually developed by an industrial partner at great expense. It is common to have limited resources and technology at the preclinical level, which might impose detection limits (LODs), depending on the platform. Having to base conclusions on naive techniques that cannot accommodate such LODs might cause researchers to reject a promising biomarker at the early stage of biomarker discovery. Hence, it is crucial to properly evaluate biomarkers at the early stage of research. Furthermore, this will in turn result in decisions related to spending resources to improve technology for those promising biomarkers and eliminate these technical issues. Examples of biomarkers with such issues are not rare and might be related to colorectal cancer as well as pancreatic cancer among others.

Colorectal cancer is the third most common cancer diagnosed in both men and women in the United States. Early detection of colorectal cancer is key to successful treatment since it might be curable if detected at an early stage (see [2]). Carcinoembryonic antigen (CEA), normally produced in gastrointestinal tissue, is known to increase in individuals with colorectal carcinoma and to exhibit increased levels in individuals with pancreatic, lung and breast carcinoma as well. CEA is produced during the fetal period by endothelial cells, and postnatal by epithelial cells. Adenocarcinoma cells of the colon as well as fetal intestinal cells produce CEA, which is used as a biomarker for treatment monitoring purposes and to identify disease recurrence after surgical intervention. However, laboratory limitations play a role in measuring CEA, making its value undetectable below some LOD that may vary among different assay techniques. Carbohydate antigen 19-9 (CA19-9) is another well-known biomarker that is mainly used for diagnosis or monitoring response to therapy in pancreatic cancer. CA19-9 might also be subject to a lower LOD, the value of which depends on the assay. Although the most common situation is the existence of a lower LOD, there might be more rare cases that an upper limit of detection is present. An example is CA19-9. Its levels may be elevated in many types of cancer such as colorectal cancer, esophageal cancer, or hepatocellular carcinoma, as well as in pancreatitis and cirrhosis (see [3], [4], among others). Hence, it might not be safe to conclude that non-detected CA19-9 measurements above the upper LOD correspond to pancreatic cancer. More specifically, as pancreatic cancer is a rare disease and there is no suitable high-risk population for screening, any biomarker that identifies an average-risk population for screening will require very high specificity, which in turn will require unbiased estimators of the associated sensitivity. A very high specificity level of CA19-9 or the target biomarker, in cases where an upper LOD is present, could compromise the ability to derive an accurate estimate of the associated sensitivity. Likewise, when an accurate estimate of the specificity at a high sensitivity level (e.g., 95%) is required in a clinical context, the lower LOD will limit the ability to accurately estimate the desired specificity at the cutoff that corresponds to that sensitivity. It is common that a potential upper limit of detection might be resolved within the laboratory by dilution. However, such a strategy costs in time, resources and available samples that might be limited. This is the case presented in [5] that focuses on Hepatitis B and the Cobas Amplicor HBV Monitor test for which there is a limited range of detection due to an upper limit of detection. In that study even the dilution took place prior to the assay to avoid issues with the upper limit of detection, some sera had to be retested because they were still found to be above the upper limit of detection. However, such a retesting might cause a batch effect.

Another example with an upper limit of detection is discussed in [6]. In that paper, the authors investigate kidney biomarkers for patients with cirrhosis and acute kidney injury. The measurements that correspond to the upper limit of detection are simply replaced by the exact value of the limit of detection and analysis is performed as if these measurements were exactly observed. Such an analysis might cause (downward) biased results. Therefore, the lower or upper LODs limit the ability to select candidate biomarkers for further development of clinical assays. We propose new approaches to address these issues.

The typical statistical tool to evaluate either continuous or ordinal biomarkers is the receiver operating characteristic (*ROC*) curve. If a lower LOD is present, clinicians cannot obtain exact measurements for patients whose measurements are below that lower LOD, d_L. It is also not possible to obtain exact measurements when they exceed the upper LOD, d_U. Hence, data that are left, right, or doubly censored arise in cases of a lower LOD, upper LOD, or both. We denote the biomarker measurements of healthy individuals as Y_0, and those of individuals with disease as Y_1. Assuming a continuous setting and that higher marker measurements are more indicative of the disease, we define sensitivity as $P(Y_1 > c)$ and specificity as $P(Y_0 < c)$, where $-\infty < c < \infty$ is a decision cutoff, above which we consider all measurements to be "positive." For each cutoff value c, we obtain a pair of sensitivity and specificity values. The *ROC* curve is the plot of all possible pairs of sensitivity and specificity values obtained by scrutinizing all possible values of the cutoff c. Namely, $ROC(c) = \{FPR(c), TPR(c), c \in (\infty, \infty)\}$. Let $F_0(x)$ and $F_1(x)$ be the distribution functions of the healthy and the diseased group respectively. The definition of the *ROC* curve as a function of the underlying distribution functions is

$$\mathrm{ROC}(t) = S_1(S_0^{-1}(t)), t \in (0, 1). \quad (1)$$

where $S_0(x) = 1 - F_0(x)$ and $S_1(x) = 1 - F_1(x)$ are the survival functions of the healthy and the diseased group respectively. For a detailed overview of the *ROC* curve, see [7] and [8]. When a lower and/or upper LOD is present, the construction of the *ROC* curve must account for the censored nature of the available data. Some practices in the setting of a lower LOD of a positively defined continuous biomarker involve ignoring the censored nature of the data and proceeding with the usual empirical *ROC* curve analysis (see [11] for an overview). This strategy is essentially equivalent to simply treating all left-censored measurements as if they were observed exactly at d_L, which is expected to result in bias. Other similar approaches might involve the replacement of these left-censored values with $d_L/2$ (see [12]) or $d_L/\sqrt{2}$ (see [13]). The first approach is based on the assumption that the distribution of the data is uniform in the interval $[0, d_L]$. The rationale for using $d_L/\sqrt{2}$ is a strategy that completes the tail of the underlying density with a straight line toward the (0,0) point, forming an orthogonal triangle with corners at the points $(0, 0)$, $(d_L, 0)$, $(d_L, f(d_L))$, where f is the underlying density function. It is not a surprise that the resulting empirical *ROC* curve remains the same regardless of which of these nave approaches one uses. This is due to the ranking of the data that remains unchanged no matter the replacement value used. The disadvantage of using such a replacement approach is threefold: (1) It only applies to positive supported marker measurements, which might not be the case even though it is

common; (2) such an approach cannot be applied for the case of an upper LOD, as theoretically the interval of the true measurement in that case is $[d_U, +\infty)$; and more importantly, (3) under the natural assumption of concavity of the ROC curve, the resulting ROC curve will most often be biased, which will result in downward bias of the targeted sensitivity and specificity values.

Some researchers have proposed parametric approaches. The use of the binormal model and the bigamma model was investigated in [14]. Mumford and others (2006) (see [15]) explored the binormal model as well, under a framework of taking random samples and pooling biospecimens. In [16] maximum likelihood ratio tests to compare biomarkers subject to LODs were considered. The multivariate normality assumption is employed in [17] for multiple biomarkers subject to an LOD. Maximum likelihood-based approaches under the normality assumption for making inference for optimal linear combinations of two biomarkers subject to LODs are discussed in [18]. They considered multiple biomarkers in the presence of lower LODs as well. However, such parametric approaches might be too restrictive. A spline-based approach was proposed in [19]. Their approach relaxes the need for common parametric assumptions; however, it cannot accommodate both upper and lower LODs.

The paper is organized as follows: In Section 2, we show that the naive empirical approach will provide a biased area under the ROC curve (*AUC*) estimate. We further show that, after the use of any replacement value, the empirical-based *ROC* curve will result in a biased estimate of the $ROC(t)$. In Sections 3 and 4, we explore power transformations as well as the use of a flexible parametric model, the extended generalized gamma (GG) distribution (see [20] and [21]). In Sections 5 and 6, we investigate an imputation of kernel- and spline-based approaches to achieve robustness. We evaluate our approaches through simulations in Section 7, and apply them to colon cancer and pancreatic cancer data in Section 8. We close with a discussion.

2. Empirical ROC

A naive approach to the problem under study is to simply use the empirical *ROC* curve and ignore the censored nature of the data. The empirical-based *ROC* estimate for a continuous marker is obtained by all possible pairs of $\hat{TPR}(c)=\sum_{i=1}^{n_1}\frac{I(Y_1>c)}{n_1}$ and $\hat{FPR}(c)=\sum_{i=1}^{n_0}\frac{I(Y_0>c)}{n_0}$, $-\infty < c < \infty$. Some techniques (see Hughes, 2000 for a discussion) involve simply using a fixed replacement value for the censored data that are piling up on the lower LOD, d_L. Some replacement values discussed in the literature are $a_L = d_L/2$ or $a_L = d_L/\sqrt{2}$, and are used to attempt to reduce bias in terms of the estimated *AUC*, which in the case of a continuous marker corresponds to the probability $P(Y_0 < Y_1)$. However, for any replacement value $a_L < d_L$ and/or $a_U > d_U$, the empirical estimate of the *ROC* curve remains the same when it is obtained by an analysis that ignores the censored nature of the data. We show that any estimate based on a replacement value strategy induces bias. We generalize the result of [14] and show that the theoretical *ROC* curve in the

presence of LODs d_L and d_U yields a biased estimate for any replacement value, both for the $ROC(t)$ curve and for the AUC.

Proposition 1 Let the continuous marker that yields measurements for healthy individuals and those with disease be denoted by Y_0 and Y_1 from distributions F_0, F_1, respectively. Assume that the marker is subject to a lower LOD, d_L, and an upper, LOD d_U. Let $a_L \leq d_L$ and $a_U \geq d_U$ be the replacement values for the censored scores, which define the imputed scores M_0 and M_1 for the healthy and diseased, respectively. Denote with $ROC_Y(t)$ the true unknown $ROC(t)$ based on the unobserved scores Y_0 and Y_1, and denote with $ROC_M(t)$ the ROC curve that is based on M_0 and M_1. It then holds that:

i. $ROC_M(t)$ remains the same for any replacement values a_U and a_L. $ROC_M(0) = 0$, $ROC_M(1) = 1$, and $ROC_M(t) = ROC_Y(t)$ for $t \in [S_0(d_U), S_0(d_L)]$. It is not defined elsewhere.

ii. If we define $\mathrm{ROC}^*_M(t)$ by linearly completing $ROC_M(t)$, then we have $\mathrm{ROC}^*_M(t) = \mathrm{ROC}_Y(t)$ for $t \in (0, 1)$ if and only if $\frac{S_1(x)}{S_0(x)} = \frac{S_1(d_U)}{S_0(d_U)} x > d_U$, $\frac{F_1(x)}{F_0(x)} = \frac{F_1(d_L)}{F_0(d_L)} x < d_L$.

iii. $\mathrm{ROC}^*_M(t) = \mathrm{ROC}_Y(t)$ for $t \in (0, 1)$ if and only if the following holds for the positive and negative predictive values of the (fully observed) marker: $PPV_Y(c) = PPV_Y(d_U)$ $c > d_U$, $NPV_Y(c) = NPV_Y(d_L)$ $c < d_L$.

iv. If Y_0 is smaller than Y_1 in the hazard rate order ($Y_0 \leq_{hr} Y_1$), then $\mathrm{ROC}^*_M(t) < \mathrm{ROC}_Y(t)$ for $t \in (0, S_0(d_U))$. If Y_0 is smaller than Y_1 in the reverse hazard rate order ($Y_0 \leq_{rh} Y_1$), then $\mathrm{ROC}^*_M(t) < \mathrm{ROC}_Y(t)$ for $t \in (S_0(d_L), 1)$. Equivalently $\mathrm{ROC}^*_M(t) < \mathrm{ROC}_Y(t)$ for $t \in (0, S_0(d_U))$ if $PPV_Y(c)$ is increasing in c and $\mathrm{ROC}^*_M(t) < \mathrm{ROC}_Y(t)$ for $t \in (S_0(d_L), 1)$ if $NPV_Y(c)$ is decreasing in c. In particular, if $ROC_Y(t)$ is concave (equivalent to Y_0 being smaller than Y_1 in the likelihood ratio order, $Y_0 \leq_{lr} Y_1$), then $\mathrm{ROC}^*_M(t) < \mathrm{ROC}_Y(t)$ for $t \notin [S_0(d_U), S_0(d_L)]$.

v. For the area under $\mathrm{ROC}^*_M(t)$, $\mathrm{AUC}^*_M = \frac{S_0(d_U)S_1(d_U)}{2} + \int_{S_0(d_U)}^{S_0(d_L)} \mathrm{ROC}(t)\mathrm{d}t + \frac{(1 - S_0(d_L))(1 - S_1(d_L))}{2}$, we have that $\mathrm{AUC}^*_M < \mathrm{AUC}_Y$, when the marker is subject to an upper LOD and $PPV_Y(c)$ is increasing in c, or when the marker is subject to a lower LOD and $NPV_Y(c)$ is decreasing in c, or when the marker is subject to both upper and lower $LODs$ and with $PPV_Y(c)$ increasing and $NPV_Y(c)$ decreasing in c.

The proofs of the statements in the proposition are given in the Appendix. The first result in the proposition states that the ROC curve that corresponds to any replacement value strategy for the LODs is identical to the ROC curve obtained by setting all values outside the LODs equal to the LODs, and is a mixed-type ROC curve. In practice, one must be cautious with inference based on the empirical ROC since observed FPR and TPR values may not be

attainable in the true *ROC* curve for the marker when specific LODs are set. Statements (ii) and (iii) of the proposition concern the common practice of replacing values outside the LODs and then proceeding to inference for accuracy measures while ignoring the LOD issue. In general, this practice results in biases. Strictly speaking, this practice is valid when one is willing to assume that, in terms of patient classification, knowledge of the actual marker measurement is useless when classifying a patient as healthy based on a value smaller than the lower LOD or when classifying a patient as having the disease based on a value larger than the upper LOD. We note here that there are some scenarios in which $\mathrm{ROC}^{*}_{M}(t)$ may be close to $ROC_Y(t)$. For example, consider the marker values for the diseased group being a mixture of a subpopulation identical to the values for the healthy group and another subpopulation of marker values that are well separated (in the *ROC* sense) from those for the healthy group. In the case of a reasonable lower LOD, d_L, based on the lower tail for the healthy group, we will have an approximately constant cdf ratio, resulting in $\mathrm{ROC}^{*}_{M}(t) \approx \mathrm{ROC}_{Y}(t)$. Statements (iv) and (v) of the proposition imply that the use of the empirical *ROC* in the presence of LODs results in the underestimation of the true ROC for markers that have desirable properties. Thus, LODs may lead to 'good' markers being 'shortchanged' on the basis of estimated accuracy measures when using the empirical *ROC*.

3. Box-Cox based ROC

The available data under the framework studied in this paper are of the form $\{Y_i, \Delta_i\}$, $i = 1, 2, 3, \ldots, n$ where $Y = [Y_{0j}, Y_{1k}]$ with $j = 1, 2, \ldots, n_0$ with $k = 1, 2, \ldots, n_1$, ($n_0 + n_1 = n$) and Δ_i is the censoring indicator that takes values -1, 0, and 1 for a left-censored, an exactly observed and a right-censored measurement, respectively. When common parametric models cannot be justified by the data at hand, power transformations are often employed. Such a transformation is the so-called Box-Cox transformation (see [22]). The Box-Cox transformation has been used under the ROC curve framework (see [8], [9], [10]). Its form is given by $Y^{(\lambda)} = \frac{Y^{\lambda} - 1}{\lambda}$ if $\lambda \neq 0$, and by $Y^{(\lambda)} = \log(Y)$ if $\lambda = 0$. The likelihood under this context is based on the power normal distribution implied for the original untransformed data:

$$f(y_i) = \frac{y_i^{\lambda - 1}}{\Phi_Y \sigma^{(\lambda)}} \phi\left(\frac{y_i^{(\lambda)} - \mu^{(\lambda)}}{\sigma^{(\lambda)}}\right),$$

where $\Phi_Y = \Phi(\mathrm{sgn}(\lambda)|\frac{\lambda \mu_1^{(\lambda)} + 1}{\lambda \sigma_1^{(\lambda)}}|)$ if $\lambda \neq 0$, and 1 otherwise. This is a truncation term often neglected by many authors, as discussed in [23]. However, as the support of the data must be positive in order for the Box-Cox transformation to be feasible, this truncation term must be taken into account when building the corresponding likelihood, which is given by

$$
\begin{aligned}
L(p) \quad &= \prod_{i=1}^{n_0} \left(\frac{1}{\Phi_0} \left[\Phi\left(\frac{y_{0i}^{\lambda}}{\lambda\sigma_0^{(\lambda)}} - \frac{1+\lambda\mu_0^{(\lambda)}}{\lambda\sigma_0^{(\lambda)}} \right) - \Phi\left(\frac{1+\lambda\mu_0^{(\lambda)}}{\lambda\sigma_0^{(\lambda)}} \right) \right] \right)^{I(\delta_i=-1)} \left(\frac{y_{0i}^{\lambda-1}}{\Phi_0\sigma_0^{(\lambda)}} \phi\left(\frac{y_{0i}^{(\lambda)}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}} \right) \right)^{I(\delta_i=0)} \\
&\times \left(1 - \left(\frac{1}{\Phi_0} \left[\Phi\left(\frac{y_{0i}^{\lambda}}{\lambda\sigma_0^{(\lambda)}} - \frac{1+\lambda\mu_0^{(\lambda)}}{\lambda\sigma_0^{(\lambda)}} \right) - \Phi\left(-\frac{1+\lambda\mu_0^{(\lambda)}}{\lambda\sigma_0^{(\lambda)}} \right) \right] \right) \right)^{I(\delta_i=1)} \\
&\times \prod_{i=1}^{n_1} \left(\frac{1}{\Phi_1} \left[\Phi\left(\frac{y_{1i}^{\lambda}}{\lambda\sigma_1^{(\lambda)}} - \frac{1+\lambda\mu_1^{(\lambda)}}{\lambda\sigma_1^{(\lambda)}} \right) - \Phi\left(\frac{1+\lambda\mu_1^{(\lambda)}}{\lambda\sigma_1^{(\lambda)}} \right) \right] \right)^{I(\delta_i=-1)} \left(\frac{y_{1i}^{\lambda-1}}{\Phi_1\sigma_1^{(\lambda)}} \phi\left(\frac{y_{1i}^{(\lambda)}-\mu_1^{(\lambda)}}{\sigma_1^{(\lambda)}} \right) \right)^{I(\delta_i=0)} \\
&\times \left(1 - \left(\frac{1}{\Phi_1} \left[\Phi\left(\frac{y_{1i}^{\lambda}}{\lambda\sigma_1^{(\lambda)}} - \frac{1+\lambda\mu_1^{(\lambda)}}{\lambda\sigma_1^{(\lambda)}} \right) - \Phi\left(-\frac{1+\lambda\mu_1^{(\lambda)}}{\lambda\sigma_1^{(\lambda)}} \right) \right] \right) \right)^{I(\delta_i=1)}
\end{aligned}
\tag{2}
$$

where $p = \left(\mu_0^{(\lambda)}, \sigma_0^{(\lambda)}, \mu_1^{(\lambda)}, \sigma_1^{(\lambda)}, \lambda\right)$, $\Phi_k = \Phi(\mathrm{sgn}(\lambda)|\frac{\lambda\mu_k^{(\lambda)}+1}{\lambda\sigma_k^{(\lambda)}}|)$ if $\lambda \neq 0$, and 1 otherwise with $k = 0, 1$. After maximizing likelihood (2), we obtain an estimate of all the underlying parameters $\hat{p} = \left(\hat{\mu}_0^{(\lambda)}, \hat{\sigma}_0^{(\lambda)}, \hat{\mu}_1^{(\lambda)}, \hat{\sigma}_1^{(\lambda)}, \hat{\lambda}\right)$. Note that $I(A) = 1$ if A is true, and 0 otherwise. Based on the estimate of the transformation parameter λ, $\hat{\lambda}$, we then obtain the proposed estimate of the *ROC* curve by assuming that the transformed scores follow a truncated normal distribution, truncated at $-1/\lambda$. Namely, the cdf of the healthy is assumed to be

$$
F_0(y) = \frac{\Phi\left(\frac{y-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}} \right) - \Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}} \right)}{1 - \Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}} \right)} \times I\left(x > -\frac{1}{\lambda} \right), \tag{3}
$$

and this is similarly determined for the diseased. The inverse cdf function of the healthy is given by

$$
F_0^{-1}(t) = \Phi^{-1}\left(\Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}} \right) + t\left(1 - \Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}} \right) \right) \right) \sigma_0^{(\lambda)} + \mu_0^{(\lambda)} t \in (0,1). \tag{4}
$$

Hence, the corresponding proposed *ROC* curve can be written in closed form based on the corresponding survival functions $S_0 = 1 - F_0$, $S_1 = 1 - F_1$, and is given by

$$\mathrm{ROC}(t)=1-\frac{\Phi\left(\frac{\Phi^{-1}\left(\Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}}\right)+(1-t)\left(1-\Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}}\right)\right)\right)\sigma_0^{(\lambda)}}{\sigma_0^{(\lambda)}}\right)-\Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}}\right)}{1-\Phi\left(\frac{-\frac{1}{\lambda}-\mu_0^{(\lambda)}}{\sigma_0^{(\lambda)}}\right)} \quad (5)$$

where $t \in (0, 1)$. An estimate of this *ROC* curve can then be obtained by simply plugging in the estimated parameters $\hat{p}=\left(\hat{\mu}_0^{(\lambda)},\hat{\sigma}_0^{(\lambda)},\hat{\mu}_1^{(\lambda)},\hat{\sigma}_1^{(\lambda)},\hat{\lambda}\right)$. The tails of the underlying density of the transformed scores are expected to conform with the tails of an estimated truncated normal distribution, and hence are nicely captured below the lower LOD and beyond the upper LOD in contrast to a naive approach that considers a simple replacement value.

4. Extended Generalized Gamma ROC

In case there are biological insights for speculating the shape of the tail of the underlying density of each group, and to allow for more flexibility in the tails, we explore a flexible parametric model known as generalized gamma (GG), which is based on the initial work of [20]. Extensions and computationally robust parameterizations for this model have proposed in [24]. For a thorough and elegant investigation of the extended GG model, we refer to [21]. Its density (see [24]) can be written as

$$f(x;\alpha,\lambda,\gamma)=\frac{|\alpha|}{\Gamma(\gamma)}\gamma^{\gamma}\lambda^{\alpha\gamma}x^{\alpha\gamma-1}\exp\{-\gamma(\lambda x)^{\alpha}\},$$

where $\alpha \neq 0$ and $\gamma > 0$ are the shape parameters, and $\lambda > 0$ is the scale parameter. When $\alpha = 0$, the limiting case of the lognormal distribution is obtained. When $\gamma = 1$ or $\alpha = 1$, we obtain the Weibull or the gamma distribution, respectively. Here, we denote the GG distribution by GG(α, λ, γ). The survival function is

$$S(x;\alpha,\lambda,\gamma)=\begin{cases} I\{\gamma(\lambda x)^{\alpha},\gamma\}, & \text{if } \alpha<0 \\ 1-I\{\gamma(\lambda x)^{\alpha},\gamma\}, & \text{if } \alpha>0. \end{cases}$$

where $I\{\cdot,\cdot\}$ is the incomplete gamma function. It is defined as $I(x,\theta)=\frac{1}{\Gamma(\theta)}\int_0^x t^{\theta-1}e^{-t}\mathrm{dt}$ and is also known as the regularized gamma function. This distribution provides a parametric platform under a survival analysis concept. Its hazard rate can take the form of the four most common types, i.e., increasing, decreasing, arc-shaped and bathtub-shaped. Even though the hazard rate under the *ROC* curve framework has no appealing interpretation, it loosely expresses the risk of having a measurement $q + \Delta q$ given that the measurement is $> q$, and directly relates to the tails of the underlying densities, the form of which are of crucial interest in our setting. For example, a decreasing hazard rate refers to a

heavy tail of the underlying density. Similarly, when the hazard rate is increasing, the tail of the density is light. The tails are of primary interest in our setting since it is the tails that are not 'revealed' due to the upper or lower LOD. We denote $I\overline{(x,\alpha)}=\frac{1}{\Gamma(\alpha)}\int_x^\infty t^{a-1}e^{-t}\mathrm{dt}=1-\frac{1}{\Gamma(\alpha)}\int_0^\infty t^{a-1}e^{-t}\mathrm{dt}=1-I(x,\alpha)$ and let $I^{-1}(x, \alpha)$ and $I^{-1}\overline{(x,\alpha)}$ be the inverse functions of $I(x, \alpha)$ and $I\overline{(x,\alpha)}$, respectively. The proposed GG-based ROC curve can be derived after inverting the survival distribution of the healthy group. We denote the parameter vector of the GG that refers to the healthy group as (α_0, γ_0, λ_0), and the corresponding parameters for the group with disease as (α_0, γ_0, λ_0). The corresponding likelihood is given by

$$L(p)=\prod_{i=1}^{n_0}F_0(y_{0i};\alpha_0,\gamma_0,\lambda_0)^{I(\delta_i=-1)}f_0(y_{0i};\alpha_0,\gamma_0,\lambda_0)^{I(\delta_i=0)}S_0(y_{0i};\alpha_0,\gamma_0,\lambda_0)^{I(\delta_i=1)}\\ \times\prod_{j=1}^{n_1}F_1(y_{1j};\alpha_1,\gamma_1,\lambda_1)^{I(\delta_j=-1)}f_1(y_{1j};\alpha_1,\gamma_1,\lambda_1)^{I(\delta_j=0)}S_1(y_{1j};\alpha_1,\gamma_1,\lambda_1)^{I(\delta_j=1)} \quad (6)$$

where $F_0(y; \alpha_0, \gamma_0, \lambda_0) = 1 - S_0(y; \alpha_0, \gamma_0, \lambda_0)$, and $f_0(y; \alpha_0, \gamma_0, \lambda_0)$ is the GG cdf and pdf that correspond to the healthy group. We note that δ_i is -1 if the i-th measurement is left censored, 1 if it is right censored and 0 if the measurement is exactly observed. The indicator function $I(A)$ is 1 if A is true, and 0 otherwise. This is performed similarly for the group with disease. Hence, the proposed ROC curve under this parameterization is given by

$$\mathrm{ROC}(t)=\begin{cases} I\{\gamma_1(\lambda_1(I^{-1}\{\gamma_0(\lambda_0 t)^{\alpha_0},\gamma_0\}))^{\alpha_1},\gamma_1\}, & \text{if } \alpha_1<0,\alpha_0<0\\ 1-I\{\gamma_1(\lambda_1(I^{-1}\{\gamma_0(\lambda_0 t)^{\alpha_0},\gamma_0\}))^{\alpha_1},\gamma_1\}, & \text{if } \alpha_1>0,\alpha_0<0.\\ I\{\gamma_1(\lambda_1(I^{-1}\overline{\{\gamma_0(\lambda_0 t)^{\alpha_0},\gamma_0\}}))^{\alpha_1},\gamma_1\}, & \text{if } \alpha_1<0,\alpha_0>0\\ 1-I\{\gamma_1(\lambda_1(I^{-1}\overline{\{\gamma_0(\lambda_0 t)^{\alpha_0},\gamma_0\}}))^{\alpha_1},\gamma_1\}, & \text{if } \alpha_1<0,\alpha_0<0 \end{cases} \quad (7)$$

After maximizing the likelihood (6), the estimated ROC curve can be obtained by simply plugging $\hat{\alpha_0}$, $\hat{\gamma_0}$, $\hat{\lambda_0}$, $\hat{\alpha_1}$, $\hat{\gamma_1}$, $\hat{\lambda_1}$ into expression (7). The calculation of the incomplete gamma function as well as its inverse are built into software such as `MATLAB` (see `gammainc` and `gammaincinv` functions) and `Mathematica` (see `GammaRegularized` and `InverseGammaRegularized`). Furthermore, the fit of the GG model is readily available by using `SAS` (see `LIFEREG procedure`); a `MATLAB` routine named `aft` for the same purpose is available in the `File exchange` link of `Mathworks` provided by the first author.

5. Kernel-Imputation Based ROC

Here, we focus on a hybrid approach based on non parametric kernel based distribution estimates and parametric based imputation. A non parametric kernel based estimate of a density of a fully observed random variable X is given by:

$$\hat{f}_X(x)=\frac{1}{n_X h_X}\sum_{i=1}^{n_X}K\left(\frac{x-X_i}{h_X}\right), \quad (8)$$

where $K(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$, and h_X is the bandwidth for which one can choose to use a simple plug in type equal to $h_X=0.9\min(\mathrm{std}(X_i),\mathrm{IQR}(X_i))n_X^{0.2}$ (see [25]. When it comes to censored data the kernel estimate depends on the steps of the empirical survival curve that accommodates censoring. For example in the presence of right censoring the kernel based estimate depends on the steps of the Kaplan Meier estimate (see [26]). However, the Kaplan Meier survival estimate, and hence the kernel based distribution estimates, suffer from the 'last is censored' phenomenon. Namely, the Kaplan Meier step curve cannot provide a valid estimate beyond the last observation if that is right censored. The same problem persists for doubly censored data even if one decides to use Turnbull's algorithm (see [27]). That is, no valid estimate can be obtained before the first observation if that is left censored.

To overcome this problem we consider imputing the censored data by random draws of plausible parametric models in order to have some pseudodata in the tails of both distributions (of the diseased and the healthy) and then proceed with the non-parametric kernel based estimators for both distributions separately. For an overview regarding imputation see [28]. Assume that the number of left and right censored observations of the healthy are denoted with n^*_{0L} and n^*_{0U} respectively (notation is similar for the diseased related censored data). The proposed algorithm is:

- Step 1: Draw n^*_{0L} data from a plausible parametric model f_0, Y^*_{0Li}, $i=1,\ldots,n^*_{0L}$ for which $Y^*_{0i}<d_L\,\forall_i$.
- Step 2: Draw n^*_{0U} data from a plausible parametric model f_0, Y^*_{0Ui}, $i=1,\ldots,n^*_{0U}$ for which $Y^*_{0i}>d_U\,\forall_i$. The pseudodata set for the healthy is denoted with $Y^*_0=[Y^*_{0L},(Y_0{:}\delta_i=0),Y^*_{0U}]$. Namely Y^*_0 consists of all imputed data in place of the previously censored ones, and the exactly observed from the original Y_0.
- Step 3: Repeat the analogous Step 1 and 2 for the diseased group and obtain Y^*_1.
- Step 4: Obtain the kernel based distribution estimates based on Y^*_0 and Y^*_1 separately and hence obtain an estimate of the kernel based *ROC*, $\hat{ROC}(t)$, $t\in(0,1)$.
- Step 5: Repeat Steps 1-4, 30 times, and obtain 30 smooth kernel based *ROC* curves.
- Step 6: The proposed kernel based *ROC* curve is the average of the *ROC* curves obtained in Step 5, namely $\hat{ROC}(t)=\frac{\sum_{j=1}^{30}\hat{ROC}^j(t)}{30}$.

The proposed estimate of Step 6 of the previous algorithm can be considered as a robust alternative to simply assuming a parametric model. It only uses the parametric assumption to fill in the data of the censored tails of each distribution while the main body of the distributions of both groups is estimated non-parametrically. This hybrid estimate has also the property of providing a smooth ROC curve which is a natural assumption for the true ROC curve. Here, based on the two previous sections, we consider two versions of this estimate: The kernel based *BC* ROC estimate (Kernel(BC)), and the kernel based *GG* ROC estimate where imputation is based on the *GG* model (Kernel(GG)). For the former, we first transform the data based on the maximum likelihood parameter estimates, and then imputation is done based on the obtained truncated normal distributions.

6. Spline Based ROC

Here we explore a spline based technique that is based on estimating the underlying survival functions of the healthy and the diseased, namely $S_0(t)$ and $S_1(t)$. We extend a monotone spline initially presented in [29]. Consider the K knots placed at $\tau_1 < \ldots < \tau_K$ and let the natural spline for the cumulative hazard which can be written as

$$H(x)=\theta_1(x-\tau_1)_+^3+\ldots+\theta_{K-2}(x-\tau_{K-2})_+^3+\theta_{K-1}(x-\tau_{K-1})_+^3+\theta_K(x-\tau_K)_+^3 \quad (9)$$

$$\theta_{K-1}=\frac{\theta_1(\tau_1-\tau_K)+\theta_2(\tau_2-\tau_K)+\ldots+\theta_{K-2}(\tau_{K-2}-\tau_K)}{\tau_K-\tau_{K-1}}$$
$$\theta_K=\frac{\theta_1(\tau_1-\tau_{K-1})+\theta_2(\tau_2-\tau_{K-1})+\ldots+\theta_{K-2}(\tau_{K-2}-\tau_{K-1})}{\tau_{K-1}-\tau_K}. \quad (10)$$

The above model has the following appealing properties: (i) It is linearly extrapolated beyond the last knot τ_K, (ii) It equals to zero before the first knot τ_1, (iii) It is a smooth function since its first two derivatives are continuous. The number of parameters of interest $\theta_1, \theta_2,\ldots, \theta_k$, is $K-2$ since the last two parameters are functions of all previous.

We attempt the fit of the above model to the cumulative hazard nonparametric step estimate. For that reason, monotonicity conditions must be imposed since the cumulative hazard is a non decreasing function. There is an implied region of monotonicity which is non-linear (see [29]). We employ an optimal polygon strategy to approximate linearly the nonlinear monotonocity region so that the minimization problem reduces to a least squares one with linear restrictions on the parameters (see [29]). We propose the use of six knots and regarding the knot placement we consider the following strategy: We derive the following 10 percentiles derived by the fully observed data: 0, 2.5*th*, 5*th*, 10*th*, 20*th*, 40*th*, 50*th*, 60*th*, 80*th*, and 100*th*. There are 10!/(6!4!) = 210 possible combinations, and thus 210 possible knot schemes. In a given application all 210 knot schemes are explored by fitting model (9) to the non-parametric maximum likelihood based cumulative hazard function. Finally, the knot scheme that results to the smallest distance from the corners of the step cumulative

hazard function is the one chosen. That is, the knot scheme selection is based on the objective function for minimization:

$$\Psi_0(\hat{\boldsymbol{\theta}}_0)=\sum_i(\hat{H}_0(Y_{0i}|\Delta_{0i}=0) - \hat{H}_0^{\mathrm{KM}}(Y_{0i}|\Delta_{0i}=0))^2, \quad (11)$$

where $\hat{\boldsymbol{\theta}}_0$ is the estimated parameter vector of the spline related to the control group, Y_{0i} is the score of the i-th individual of the control group and Δ_{0i} refers to the censoring status of that individual (taking the value of 0 for an exactly observed score and the value of 1 for a right censored score), $\hat{H}_0$ is the fitted model (related to the control group) defined in (9) under the appropriate constraints of monotonicity, and $\hat{H}_0^{\mathrm{KM}}$ is the Kaplan Meier based cumulative hazard estimator obtained by the biomarker data of the control group. The notation for the diseased group is similar. In the case of doubly censored data the above criterion is defined through Turnbull's non parametric cumulative hazard function estimate:

$$\Psi_0(\hat{\boldsymbol{\theta}}_0)=\sum_i(\hat{H}_0(Y_{0i}|\Delta_{0i}=0) - \hat{H}_0^{\mathrm{TB}}(Y_{0i}|\Delta_{0i}=0))^2, \quad (12)$$

where $\hat{H}_0^{\mathrm{TB}}(\cdot)$ is Turnbull's estimate for H_0. Notation is similar for the diseased group. Resampling methods are feasible for inference. The proposed spline based ROC estimate is given by $\hat{ROC}(t)=\hat{S}_1(\hat{S}_0^{-1}(t))$ where $\hat{S}_i(t) = exp(-\hat{H}_i(t))$, $i = 0, 1$ and $\hat{H}_0(t)$ and $\hat{H}_1(t)$ are the spline based cumulative hazard estimates for the healthy and the diseased whose forms are given by (9). We note here that the placement of the first knot in the case of doubly censored data can be derived with the help of the box-cox transformation and the first percentile of the underlying density estimate. Namely, for a given data set, we apply the box-cox transformation, estimate the first percentile of the implied truncated normal as discussed in section (3), and then back-transform on the original scale. This is done because the cumulative hazard is naturally assumed to be zero before the first knot. A simulated example that includes all methods discussed is given in Figure 1. In that figure we visualize how for a simulated example that is generated from two normal distributions, all methods approximate better the true underlying ROC curve compared to the Naive Empirical.

7. Simulations

To evaluate our approaches we conduct simulation studies for 105 different scenarios and use comparisons based on different ranges of the underlying integrated squared errors. The scenarios that generate the data depend on sample size, direction of censoring, level of censoring, AUC, and the true underlying model. We consider sample sizes fo $(n_0, n_1) = (100, 100)$, (200, 200), and (500, 500). We generate data based on normal, gamma, and lognormal distributions with parameters set to achieve true values of the AUC equal to 0.7 and 0.8. All underlying true parameters for all scenarios are given in Table 1 of the Web Appendix (WA).

Note that we consider scenarios that lie in the Box-Cox family as well as out of the Box-Cox (e.g. gammas). The levels of censoring we consider are 30% and 50% caused by a lower LOD, an upper LOD or both. The criterion on which we base our evaluation is the relative integrated squared error for different *FPR* ranges, namely

$$\mathrm{rISE}_{a-b}=\frac{\int_a^b (\hat{ROC}^{(\mathrm{naive})}-\mathrm{ROC}(t)^{\mathrm{true}})^2 \mathrm{dt}}{\int_a^b (\hat{ROC}^{(\mathrm{proposed})}-\mathrm{ROC}(t)^{\mathrm{true}})^2 \mathrm{dt}}.$$ We consider exploring this criterion to intervals (a – b) = (0 – 0.2), (0 – 0.4), (0.6 – 1), (0.8 – 1) and the total (0 – 1). The value of our methods will be revealed in those cases that this criterion is > 1 and we are particularly interested in those areas where the tail of the true *ROC* is not supported by fully observed data. Namely, for lower LOD we do not fully observe the upper part of the *ROC* that corresponds to a high *FPR* (or equivalently high *TPR* range). Conversely, when an upper LOD is considered, particular interest in terms of performance lies in the part of the *ROC* estimate that corresponds to a range of low *FPRs*.

For the parametric approaches (see Tables 2 and 3(WA)) we observe the following: The BC approach appears to outperform the empirical naive estimate by far in all cases of the normal scenarios. For example, for the scenario of $AUC = 0.8$ and 50% censoring due to a lower LOD it provides $rISE_{0–0.2} = 3.1032$, namely over three times smaller ISE. We observe that as we increase the sample size the gain is greater: (n_0, n_1) = (200, 200), and (500, 500) the $rISE_{0–0.2}$ equals to 5.6420 and 14.4367 respectively. The results of the lognormal related scenarios are equally impressive. The reason is that both the normal and the lognormal models are in the family of the Box-Cox. However, the BC approach demonstrates satisfactory robustness. For the gamma related scenarios we observe for $AUC = 0.7$, and 0.8 with 50% censoring due to an upper LOD $rISE_{0–0.2} = 4.1410$ and 13.7343 respectively. As we increase the sample size, we outperform the naive empirical in almost all cases. The GG model seems to not provide so impressive results especially for smaller sample sizes. It seems, however, to outperfrom the naive empirical for larger sample sizes. For example it provides $rISE_{0–0.2} = 8.9095$ for a lognormal related scenario with an upper LOD causes 50% censoring when $AUC = 0.8$. For the same scenario and a lower LOD it yields $rISE_{0.8–1} = 1.5484$. The GG is a very flexible parametric model that contains as special cases the Weibull, the Lognormal, the Gamma, the log-logistic, the logistic, and the so called ‘ammag’ among other not so common parametric models as well (for a detailed discussion on this aspect see [21]). Due to its flexibility it seems that it requires a larger sample size to outperform the BC, which is also quite robust, but does not capture all these models theoretically within its family. In finite sample sizes, the BC seems to work well even for distributions like gamma which are out of the Box-Cox family. This is also stressed in [10]. However, as the sample size increases and of course asymptotically, being within the gamma family is in favor of the Generalized Gamma model. This is also verified by a large sample simulation we ran for $n = (5000, 5000)$ where the GG outperforms the BC throughout. This result justifies the theoretical asymptotic expectation that as the sample size increases the GG will eventually outperform the BC when we lie within the GG family and outside the BC family. However, such huge sample sizes are not realistic in the early stage of biomarker discovery.

For the kernel based approach (see Table 4(WA)) we observe that when used in combination with the BC (Kernel(BC)), it performs nicely in all scenarios, outperforming the Naive empirical in nearly all cases. For larger sample sizes we achieve up to 27 times smaller $rISE_{0-0.2}$. The results of the Kernel(GG) approach are inconclusive for smaller sample sizes (see Table 5(WA)). For $(n_0, n_1) = (500, 500)$ it seems to attain descent results in many cases, outperforming the Naive in the regions of interest as well as in terms of total rISE ($rISE_{0-1}$). Conclusions for the spline based approach are similar (see Table 6(WA)). We note, however, results for $AUC = 0.8$ are better compared to those that refer to $AUC = 0.7$. The explanation for that is the curvature of the underlying ROC curve. For poor biomarkers, it is a fact that the ROC near its tails is almost linear, hence the Naive empirical provides a computationally very effective approach for such cases. We additionally explore cases for which $AUC = 0.6$ and $AUC = 0.9$ that support this argument (see Tables 7-9(WA)).

In summary, the BC and the Kernel(BC) approaches perform nicely in all explored cases outperforming the Naive empirical. They provide satisfactory level of coverage for the underlying bootstrap based confidence intervals for the $\hat{ROC}(t)$, while the Naive Empirical dramatically fails in that aspect yielding coverage close to 0 (see Table 10(WA)). The Kernel(BC) based approach is considered as a more robust alternative compared to the *BC*. The GG, Kernel(GG), and the spline approach should be used only in cases of very large sample sizes, and still they do not attain such a good performance compared to the BC and Kernel(BC). To further illustrate that we considered an additional simulation study for which the diseased group is generated from a bimodal normal mixture density. The results are given in Table 12 of the Web Appendix. In this scenario we observe that the BC underperforms as it fails to capture satisfactorily the true underlying density which is now bimodal. However, the Kernel(BC) as being more robust by construction is outperforms the Naive empirical in terms of $rISE_{(0-1)}$ in all scenarios. The spline seems to generally be unstable and does not perform well in cases where the true AUC is not close to 0.8. The efficiency of all approaches depends heavily on the underlying curvature of that portion of the ROC curve that is undetected. In cases where the biomarkers under study have poor performance then the true ROC curve tends to be linear in the tails. Hence, for such cases the Naive empirical might have an advantage but these cases are unidentifiable in practice. We present an additional simulation for $AUC = 0.6$ for the normal related scenario for that purpose (see Tables 7-9 (WA)) in which we explore settings with $AUC = 0.6$ and $AUC = 0.9$. For the $AUC = 0.9$ the Box-Cox and the Kernel(BC) still perform nicely. It is a fact that below the lower LOD and above the upper LOD one deals with a 'black box' situation and caution is needed, since this is not a simple setting or random censoring out of which the pattern of the two underlying densities can be easily revealed. We observe that the Box-Cox based approaches tackle the problem nicely by extracting all information under its robust assumption, attempting to approximate the hidden tails. The notion of imputation under the same assumption is used for the Kernel(BC) which can be considered as a hybrid approach since it only uses the BC assumption to impute the censored data that only refer to the underlying tails of the corresponding densities.

As concluding remarks we provide the following guidance: (1) if the BC approach can be justified by the data at hand then it should preferred over all other explored approaches. (2) The spline based (HCNS) approach is not to be preferred over the other proposed

approaches. It performs better than the Naive empirical when the sample size is large and the curvature allows it (around AUC=0.8). (3) When the available sample size is very large and the GG assumption can be justified by the data at hand then we propose the GG. Based on a large sample simulation (not presented here for brevity) for the gamma related scenario, where $n = (5000, 5000)$, the *GG* outperforms the BC as expected. Such large sample sizes, however, are unrealistic in the early stage of biomarker discovery. (4) For any sample size, in case the BC assumption is not justified by the data at hand we propose the Kernel(BC) which performs satisfactorily in all cases illustrating robustness. The Kernel(BC) is the most robust alternative to the fully parametric BC method and is to be preferred when certain parametric assumptions cannot be justified by the data at hand.

8. Applications

We employ our methods by using two examples that motivated our study. First, we use CEA measurements from a case-control validation study presented in Taguchi and others (2015). The purpose of the study is to determine the performance of biomarker MARPE1 in the presence of other biomarkers such as CEA in detecting early colorectal cancer and colon adenoma. Plasma samples are used from 60 patients with adenomas, 30 patients with early colorectal cancer, 30 with advanced colorectal cancer, and 60 individuals participating as controls. CEA plasma levels were measured based on ELISA. Here we focus on the performance of CEA in discriminating between subjects with either early or advanced colorectal cancer ($n_0 = 60$) from controls ($n_1 = 60$). The lower LOD as induced by the ELISA assay results in 66.67% left censoring. The empirical Naive AUC estimate is 0.6561 (see Figure 2). We illustrate the BC (AUC=0.7016), the Kernel-BC (AUC=0.6861), as well as the spline method (AUC=0.7440) since the generalized gamma did not converge for this data set. The LOD affects the Naive empirical for $FPR > 0.2$. We observe that all approaches yield a higher $ROC(t)$ for $t \in (0.2, 1)$ compared to the empirical naive. For the underlying AUC, we obtain 95% percentile bootstrap based confidence intervals see Table 1. We also focus on specific TPR values: 0.7, 0.8, and 0.9. For the sensitivity levels we calculate 95% confidence intervals for the associated FPRs (see Table 1). We observe that the confidence intervals obtained by the naive approach are narrower as expected since variability is reduced due to the linear part that is always used for estimation in that *FPR* range in all bootstrap samples. Narrow confidence intervals are indicative of the bias induced by the Naive empirical approach. The BC and Kernel(BC) yield fairly similar results. The spline based approach yields somewhat lower FPRs at the explored TPR values but with less certainty since all corresponding condidence intervals are wider. This is the price we pay for not making any parametric assumption at no stage. For this example we recommend the estimates obtained by the BC and Kernel(BC) since simulations have shown better performance of these approaches for scenarios related to *AUCs* approximately equal to 0.7. Based on the results for TPR=0.8 and 0.9 we observe that the point estimates of the Naive Empirical (0.7 and 0.85 respectively) barely make it to lie within the 95% confidence intervals as obtained by the BC and the kern(BC) method ([0.358-0.720] and [0.382, 0.738] respectively), giving evidence that conclusions based on the Naive approach might be deceiving.

As a second example we consider CA19.9 measurements from a study conducted at the MD Anderson Cancer Center ([30]). Plasma samples were obtained from the University of Michigan Comprehensive Cancer Center, the University of California San Francisco, the University of Utah School of Medicine and the University of Texas MD Anderson Cancer Center. The purpose of the study is the evaluation of candidate protein based biomarkers for the diagnosis of early stage pancreatic cancer. We focus only on the CA19-9 marker for which a lower LOD yields approximately 15% left censoring. Since in this example only a small portion of the empirical ROC estimate is affected by the lower LOD, we use it as illustration by artificially censoring more measurements and hence observe how our methods would have captured the empirical ROC based on all observed data, which is expected to be closer to the true and uknown ROC curve. We impose an upper LOD that along with the original lower LOD yields a total of 55% censoring (see Figure 2b). For this example we focus on very low FPR values (=0.01, 0.02, 0.04, 0.05, 0.10) and provide point estimates for the corresponding sensitivity as well as percentile bootstrap based 95% confidence intervals for all approaches. Results are given in Table 1. For this example, we have the luxury of a 'substitute' for the truth, and we are able to compare how all approaches perform at these low FPR values. We can observe how our approaches capture the empirical based ROC curve that uses all data as compared to the naive empirical. This is nicely visualized in Figure 2b. The fact that the Naive empirical approach can be misleading is evident in this example. For FPR=0.01, 0.02, and 0.05 we observe that the sensitivity point estimates, as are provided by the Naive empirical, fail to lie within the 95% confidence intervals of all methods, including the 'full' empirical which takes into account all available data. We observe that the point estimates of the Box-Cox and the Kernel(BC) approach are fairly close to the 'full' empirical that takes into account all available data. For example, the Kernel(BC) approach yields $\hat{TPR} = 0.554, 0.613$, and 0.633 for FPR=0.02, 0.04, and 0.05 respectively. The corresponding sensitivity estimates based on the 'full' empirical are 0.593, 0.686, and 0.721, while for the Naive empirical we get 0.182, 0.364, and 0.455 respectively. Conclusions are similar for the confidence intervals as well. There is an impressive overlap of the CIs obtained by our approaches with those obtained by the 'full' empirical, as opposed to those that are obtained by the Naive empirical. Regarding the GG method we observe that is very close to the 'full' empirical for FPR=0.05, 0.05 and 0.10 yielding $\hat{TPR} =$ 0.686, 0.721, 0.756 that are fairly similar to the TPR estimates provided by the 'full' empirical: 0.686, 0.7210, and 0.756 respectively. The similarity is also epressed in terms of overlap of the corresponding 95% CIs. However, it seems that for even lower FPR values: $FPR = 0.01$ and 0.02 the GG has some discrepancies compared to the full empirical, a disadvantage that is not expressed by the kernelBC which generally performs better as shown in our simulation studies.

9. Discussion

Cancer screening is commonly based on biomarkers related to technology that includes gene expression profiles based on microarrays and protein expression profiles based on mass spectroscopy (see [1]). During a research based assay (Phase I of biomarker development) such technological resources might be limited due to technical reasons and as a result limits of detection might be present. Improper evaluation techniques at this stage might yield

deceiving results against a promising biomarker. As a consequence, a potentially good biomarker might fail to pass the criteria of Phase I, since a clinical assay might be a very expensive step to undertake. When LODs are present, one has to take into account the censored nature of the data during the estimation of the underlying *ROC*. In the case of a lower LOD, d_L, many researchers refer to the replacement value technique that involves imputing all scores to a value less than d_L, typically $d_L/2$ or $d_L/\sqrt{2}$ (see [11] for a discussion). However, if one is based on these imputed scores to construct the empirical *ROC* curve, then the resulting empirical estimate remains unchanged regardless the actual replacement value used. We show that the naive empirical approach is biased both in terms of *AUC* and in $ROC(t)$ for any pair of replacement values chosen. Even if one aims to build a parametric *ROC* curve based on some assumptions justified by the data, the rational of such an imputation technique is vague, and in that case building the underlying likelihood by taking into account the censored nature of the data is more appropriate. Furthermore, such an approach is inapplicable when the data lie on the real line, which often can be the case. This is due to transformations such as the *log*, that are typically employed in such settings, and project intrinsically positive scores on the real line. In addition, even if the scores are actually positive, an upper LOD is not bounded by an upper bound so as to justify any kind of replacement value.

We investigate flexible parametric, power transformations, as well as spline based techniques. More specifically, we explore the generalized gamma distribution which is an umbrella for many known distributions typically used in the *ROC* framework. The nature of censoring, however, is not random. Random censoring would allow at least some fully observed measurements to be available in the tails of the distributions of the healthy and the diseased. Under the investigated setting, all censored data pile up at the point of the LOD, resulting in a black box as far as the corresponding tails of the underlying densities concerns. An approach that deals nicely with this problem, is the proposed Box-Cox approach. The rational behind it is that it manages to transform the scores of the healthy and the diseased to exhibit an approximate known distribution. Hence, we attain some information regarding the tails of the underlying transformed densities, and thus construct the Box-Cox based ROC. One may argue that the Box-Cox assumption cannot be justified by the data at hand. For that purpose we also investigate a hybrid approach, that involves both the Box-Cox and kernel smoothers to non-parametrically estimate the fully observed measurements, and to only rely on the Box-Cox transformation to impute the censored data that lie on the tails of the underlying densities. Both approaches yield nice results and outperform the naive empirical approach throughout. An additional spline based approach is also explored. The latter does not make any distributional assumptions for the data and attempts to extrapolate the cumulative hazard of the measurements of the healthy and the diseased. The price to pay is that it is less efficient that the Box-Cox and the Kernel(BC) approach. Our simulation studies indicate that the Box-Cox based approaches are to be preferred over the others. We note also that the naive empirical might offer a nice robust and computationally simple strategy to deal with settings where very poor or very accurate biomarkers are under study. The reason is that the tails of the underlying true ROC are almost linear in those cases.

Once the ROC estimate is obtained any subsequent measures such as the AUC, the maximized Youden index or its associated cutoff point can also be straightforwardly derived. We stress that our approaches are mainly focused on the early stage of biomarker discovery in which it is usual that resources are an issue and hence technological limitations might impose limits of detection. At this early stage we attempt to explore the potential usefulness of a given biomarker so that further exploration can be suggested which would imply spending more resources in specific technologies in order to alleviate any similar limitations at subsequent stages of a biomarker study and thus be able to derive the ROC at the full spectrum of the FPR values. Estimation and inference with respect to other measures such as an optimal cut-off based on some criterion, is not generally appropriate at this early stage. Under a limit of detection framework such an optimal cutoff might lie at the undetected region of FPRs. Therefore, estimation of such a crucial index that might be used for clinical decision making should not be based on extrapolation of any kind and should generally be avoided under the studied framework.

Our proposed approaches, as shown in the simulation studies, clearly outperform the Naive empirical method which under some assumptions that are stated in our proposition is negatively biased. Thus, our approaches can provide the methodology framework to reveal potential markers that otherwise might have been ‘killed’ in the early discovery stage due to the Naive empirical estimate. As a referee pointed out, the followed methodology is appropriate only in the search and evaluation of biomarker potential. Use of our approaches is not recommended when interest lies in the actual discriminatory ability of the markers, or when interest lies in comparing the actual discriminatory capacity of two biomarkers. Our approaches are extremely useful in providing the biomarker potential, based on which subsequent decision making regarding the improvement of technology can follow. Once the actual measurements are technically possible, then the aforementioned issues of the actual accuracy of the biomarkers can be appropriately addressed. Without the proposed methods, useful biomarkers might have been totally missed and discarded as uninformative due to the Naive empirical method that most practitioners follow.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Drs. Michela Capello, Ayumu Taguchi and Samir Hanash for their biological insight and data access. The authors would also like to thank two anonymous referees and the associate editor for their comments that significantly improved this paper. The research presented in this paper is supported by NIH grants U24086368, U01DK108328, U01CA194733.

References

1. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of Biomarker Development for Early Detection of Cancer. Journal of National Cancer Institute. 2001; 93:1054–1061.
2. Taguchi A, Rho Jh, Yan Q, Zhang Y, Zhao Y, Xu H, Tripathi SC, Wang H, Brenner DE, Kucherlapati M, Kucherlapati R, Boutin AT, Wang AY, DePinho RA, Feng Z, Lampe PD, Hanash

SM. MAPRE1 as a Plasma Biomarker for Early-Stage Colorectal Cancer and Adenomas. Cancer Prevention Research. 2015; doi: 10.1158/1940-6207.CAPR-15-0077

3. McLaughlin R, O'Hanlon D, Kerin M, Kenny P, Grimes H, Given HF. Are elevated Levels of the Tumour Marker CA19-9 of any clinical significance?-An evaluation. Clin Chem Lab Med. 1999; 168(2):124–126.
4. Kim HR, Lee CH, Han SK, Shim YS, Yim JJ. Increased CA 19-9 level in patients without malignant disease. Clin Chem Lab Med. 2009; 47:750–754. [PubMed: 19402792]
5. Kessler HH, Stelzl E, Daghofer E, Santner BI, Marth E, Lackner H, Stauber RE. Semiautomated Quantification of Hepatitis B Virus DNA in a Routine Dignaostic Laboratory. Clinical and Diagnostic Laboratory Immunology. 2000; 7(5):853–855. [PubMed: 10973470]
6. Belcher JM, Sanyal AJ, Peixoto AJ, Perazella MA, Lim J, Thiessen-Philbrook H, Ansari N, Coca SG, Garcia-Tsao G, Parikh CR. Kidney Biomarkers and Differential Diagnosis of Patients With Cirrhosis and Acute Kidney Injury. Hepatology. 2014; 60(2):622–632. [PubMed: 24375576]
7. Pepe, MS. The Statistical Evaluation of Medical Diagnostic Tests for Classification and Prediction. Oxford: Oxford University Press; 2003.
8. Zhou, KH., Obuchowski, NA., McClish, DK. Statistical Methods in Diagnostic Medicine Wiley Series in Probability and Statistics Hoboken. John Wiley & Sons; 2002.
9. Faraggi D, Reiser B. Estimation of the area under the ROC curve. Statistics in Medicine. 2002; 21:3093–3106. [PubMed: 12369084]
10. Bantis LE, Nakas CT, Reiser B. Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. Biometrics. 2014; 70:212–223. [PubMed: 24261514]
11. Hughes MD. Analysis and design issues for studies using censored biomarker measurements with an example of viral load measurements in HIV clinical trials. Statistics in Medicine. 2000; 19:3171–191. [PubMed: 11113952]
12. Nehls GJ, Akland GG. Procedures for handling aerometric data. Journal for air pollution control association. 2006; 7:585–598.
13. Hornung, Reed. Estimation of average concentration in the presence of nondetectable values. Applied Occupational and Environmental Hygiene. 2007; 5:46–51.
14. Perkins NJ, Schisterman EF, Vexler A. Receiver Operating Characteristic Curve Inference from a Sample with a Limit of Detection. American Journal of Epidemiology. 2007; 165:325–333. [PubMed: 17110640]
15. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. Biostatistics. 2006; 7:585–598. [PubMed: 16531470]
16. Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to a limit of detection. Biometrics. 2008; 64:895–903. [PubMed: 18047527]
17. Perkins NJ, Schisterman EF, Vexler A. Multivariate Normally Distributed Biomarkers Subject to Limits of Detection and Receiver Operating Characteristic Curve Inference. Academic radiology. 2013; 20(7):838–846. [PubMed: 23747152]
18. Perkins NJ, Schisterman EF, Vexler A. ROC curve inference for best linear combination of two biomarkers subject to limits of detection. Biometrical Journal. 2011; 53:464–476. [PubMed: 22223252]
19. Bantis LE, Tsimikas JV, Georgiou SD. Smooth ROC curves and surfaces for markers subject to a limit of detection using monotone natural cubic splines. Biometrical Journal. 2013; 55(5):719–740. [PubMed: 23553499]
20. Stacy EW. A Generalization of the Gamma Distribution. The Annals of Mathematical Statistics. 1962; 33:1187–1192.
21. Cox C, Chu H, Schneider MF, Muoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine. 2007; 23:4352–74.
22. Box GEP, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society, Series B. 1964; 26:211–252.

23. Hamasaki T, Seo Young K. Box and Cox power-transformation to confined and censored nonnormal responses in regression. Computational Statistics and Data Analysis. 2007; 51:3788–3799.

24. Lee T, Wang J. Increased CA 19-9 level in patients without malignant disease. Statistical Methods for Survival Analysis Wiley Series in Probability and Statistics. 2003

25. Silverman, BW. Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC; 1998.

26. Wand, MP., Jones, MC. Kernel Smoothing. Boca Raton: Chapman & Hall/CRC; 1995.

27. Turnbull BW. Nonparametric estimation of a survivorship function with doubly censored data. Journal of the American Statistical Association. 1974; 38:290–295.

28. Rubin, D. Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons Inc; Hoboken: 2004.

29. Bantis LE, Tsimikas JV, Georgiou SD. Survival estimation through the cumulative hazard function with monotone natural cubic splines. Lifetime data analysis. 2012; 18(3):364–396. [PubMed: 22399231]

30. Capello M, Bantis LE, Scelo G, Zhao Y, Li P, Dhillon DS, Patel NJ, Kundnani DL, Wang H, Abbruzzese JL, Maitra A, Tempero MA, Brand R, Firpo MA, Mulvihill SJ, Katz MH, Brennan P, Feng Z, Taguchi A, Hanash SM. Sequential Validation of Blood-Based Protein Biomarker Candidates for Early-Stage Pancreatic Cancer. Journal of the National Cancer Institute. 2017; 109(4) djw266.

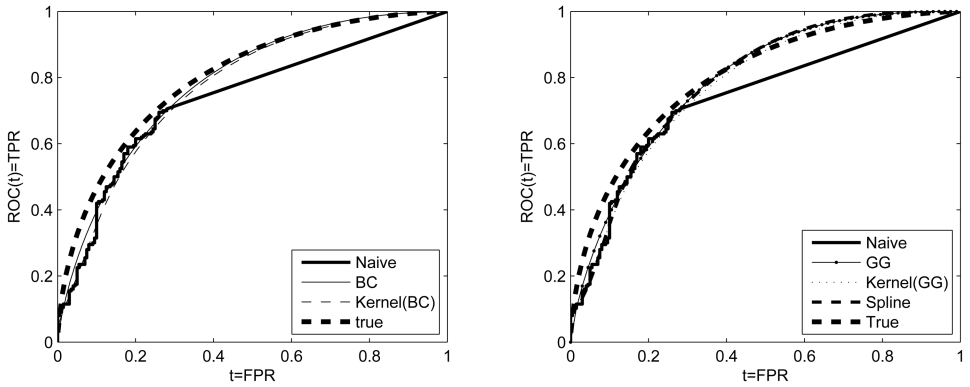

Figure 1.

Simulated data set of two normal distributions generated to attain true AUC=0.8 and expected level of censoring 50% due to a lower LOD. Left panel: Box-Cox (BC) and kernel(BC) based *ROCs* along with the true and the naive empirical. Right panel: Generalized gamma (GG), kernel(GG), and spline based *ROCs* along with the true and the naive empirical.

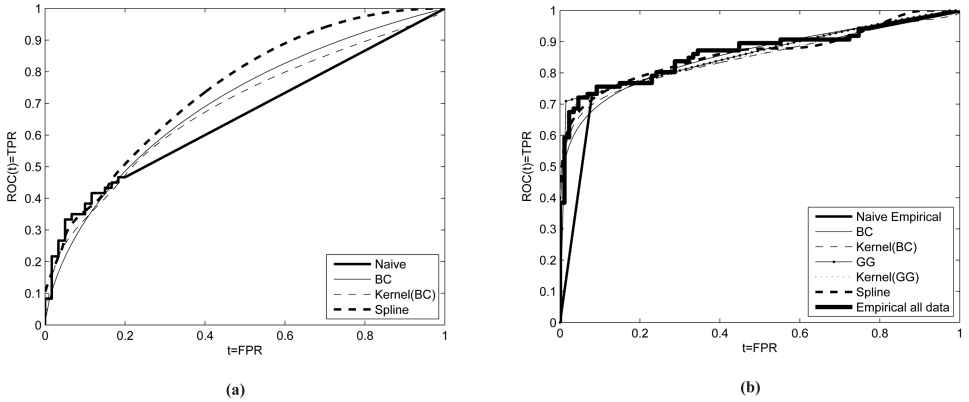

Figure 2.

Figure (2a): ROC curves for CEA that refers to colon cancer. AUCs for the Box-Cox, Kernel(BC), HCNS (spline) and naive empirical are 0.699, 0.686, 0.744 and 0.6566 respectively. Figure (2b): ROC curves for CA19-9 that refers to pancreatic cancer. AUCs for the Box-Cox, Kernel(BC), GG, kernel(GG), spline and naive empirical are 0.8577, 0.8474, 0.8657, 0.8569, 0.8627, 0.8428. The corresponding AUC for the empirical when using all available data is 0.8651. We note that even though it seems so, the GG based ROC estimate does not exhibit a step on the right panel. It simply has a very steep curvature close to $TPR = 0.7$.

Table 1

Results for the CEA and CA19-9 biomarkers.

CA19-9 related results							
	Naive	**Box-Cox**	**Kernel(BC)**	**GG**	**Kernel(GG)**	**Spline**	**Empirical using all data for CA19-9**
FPR	$\hat{TPR}$**(95%*CI*)**	$\hat{TPR}$**(95%*CI*)**	$\hat{TPR}$**(95%*CI*)**	$\hat{TPR}$**(95%*CI*)**	$\hat{TPR}$**(95%*CI*)**	$\hat{TPR}$**(95%*CI*)**	$\hat{TPR}$**(95%*CI*)**
0.01	0.091(0.052,0.219)	0.501(0.253,0.709)	0.515(0.255,0.686)	0.707(0.6406,0.8289)	0.710(0.614,0.781)	0.573(0.625,0.803)	0.384(0.291,0.674)
0.02	0.182(0.105,0.438)	0.554(0.339,0.741)	0.574(0.345,0.714)	0.711(0.6571,0.8327)	0.737(0.634,0.784)	0.617(0.641,0.806)	0.593(0.314,0.744)
0.04	0.364(0.209,0.756)	0.613(0.428,0.774)	0.637(0.456,0.746)	0.718(0.6776,0.8379)	0.767(0.656,0.789)	0.664(0.658,0.810)	0.686(0.500,0.802)
0.05	0.455(0.262,0.779)	0.633(0.463,0.785)	0.657(0.491,0.756)	0.722(0.6818,0.8401)	0.778(0.656,0.791)	0.680(0.664,0.811)	0.721(0.581,0.802)
0.10	0.756(0.524,0.837)	0.699(0.601,0.821)	0.712(0.594,0.790)	0.739(0.6981,0.8555)	0.813(0.676,0.796)	0.732(0.664,0.811)	0.756(0.651,0.826)
AUC:	0.795(0.773,0.898)	0.858(0.773,0.898)	0.846(0.773,0.898)	0.866(0.799,0.935)	0.906(0.729,0.832)	0.846(0.773,0.898)	0.865(0.816,0.917)
CEA related results							
TPR	$\hat{FPR}$(95%*CI*)	$\hat{FPR}$(95%*CI*)	$\hat{FPR}$(95%*CI*)	$\hat{FPR}$(95%*CI*)	$\hat{FPR}$(95%*CI*)	$\hat{FPR}$(95%*CI*)	
0.7	0.550(0.394,0.650)	0.417(0.235,0.591)	0.455(0.261,0.611)	-	-	0.364(0.193,0.993)	-
0.8	0.700(0.596,0.776)	0.561(0.358,0.720)	0.612(0.382,0.738)	-	-	0.473(0.236,0.999)	-
0.9	0.850(0.798,0.883)	0.748(0.535,0.857)	0.807(0.589,0.881)	-	-	0.618(0.306,1.000)	-
0.95	0.925(0.899,0.941)	0.863(0.686,0.928)	0.917(0.752,0.951)	-	-	0.723(0.356,1.000)	-
AUC:	0.656(0.577,0.735)	0.699(0.594,0.803)	0.686(0.584,0.786)	-	-	0.744(0.447,0.858)	-