ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT

Statistical Journal

Special issue on
"Statistical Models for Diagnosis and ROC Analysis "

Guest Editors:
Vanda Inácio de Carvalho
Miguel de Carvalho
Wenceslao González Manteiga

Volume 12, No.1
March 2014

REVSTAT

Statistical Journal

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística =
ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

CREDITS

- **PUBLISHER**
  - *Instituto Nacional de Estatística, I.P. (INE, I.P.)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: + 351 21 842 61 00
    Fax: + 351 21 845 40 84
    Web site: http://www.ine.pt
    Customer Support Service
    (National network) : 808 201 808
    Other networks: + 351 218 440 695
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística, I.P.*
- **EDITION**
  - *150 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*
- **PRICE [VAT included]**
  - *€ 11,00*

FOREWORD

Modern biomarker data analysis entails challenging modeling issues of the utmost importance for public health. Medical tests often use biomarker data as input, and the statistical evaluation of these tests—before their widespread application in clinical practice—requires the collaborative effort from experts with a wealth of backgrounds.

This special issue of *Revstat—Statistical Journal* gives an account of recent advances in the evaluation of medical tests, with a special emphasis on methodological, graphical, and inferential methods related to the well-known ROC curve. The key themes being surveyed include estimation, inference, and statistical modeling of ROC curves, ROC surfaces, and ROC regression, as well as modeling issues on diagnostic testing data when a verification bias exists or when no gold standard is available.

Statistical Models for Diagnosis and ROC Analysis offers a fresh look into recent advances, with an eye on future developments and on trending topics for the upcoming years.

We hope these papers encourage debate between all the experts which take part in the statistical evaluation of medical tests, and that they can provide newcomers to the field some directions on latest progresses.

On the behalf of the Editorial Board we would like to thank the authors for contributing to this special issue. Lastly, we would like to take this opportunity of putting on record our indebtedness to Professor M. Ivette Gomes, Editor-in-Chief of *Revstat—Statistical Journal*, for supporting our initiative and encouraging us throughout this editorial challenge.

V. Inácio de Carvalho and M. de Carvalho
Pontificia Univ. Católica de Chile
Santiago
Chile

W. González-Manteiga
Univ. de Santiago de Compostela
Santiago de Compostela
Spain

INDEX

Abstracted/indexed in: *Current Index to Statistics, DOAJ, Google Scholar, Journal Citation Reports/Science Edition, Mathematical Reviews, Science Citation Index Expanded®, SCOPUS* and *Zentralblatt für Mathematic.*

REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 1–20

ROC CURVE ESTIMATION: AN OVERVIEW

Authors: Luzia Gonçalves
– Unidade de Saúde Pública Internacional e Bioestatística,
Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa,
CEAUL
luziag@ihmt.unl.pt

Ana Subtil
– Departamento de Estatística e Investigação Operacional,
Faculdade de Ciências da Universidade de Lisboa,
CEAUL
asubtil@ihmt.unl.pt

M. Rosário Oliveira
– Departamento de Matemática, Instituto Superior Técnico,
Universidade de Lisboa, Portugal,
CEMAT
rsilva@math.tecnico.ulisboa.pt

Patricia de Zea Bermudez
– Departamento de Estatística e Investigação Operacional,
Faculdade de Ciências da Universidade de Lisboa,
CEAUL
pcbermudez@fc.ul.pt

Abstract:

- This work overviews some developments on the estimation of the Receiver Operating Characteristic (ROC) curve. Estimation methods in this area are constantly being developed, adjusted and extended, and it is thus impossible to cover all topics and areas of application in a single paper. Here, we focus on some frequentist and Bayesian methods which have been mostly employed in the medical setting. Although we emphasize the medical domain, we also describe links with other fields where related developments have been made, and where some modeling concepts are often known under other designations.

Key-Words:

- *Bayesian analysis; bi-normal; kernel; receiver operating characteristic curve; robustness.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

The Receiver Operating Characteristic (ROC) curve was developed by engineers during World War II for detecting enemy objects in battlefields (Collison, 1998). Its expansion to other fields was prompt and, for instance, in psychology it was used to study the perceptual detection of stimuli (Swets, 1996). Over the years, it has been widely applied in many fields including atmospheric sciences, biosciences, experimental psychology, finance, geosciences, and sociology (Marzaban, 2004; Krzanowski and Hand, 2009, and the references therein). ROC analysis has also been increasingly used in machine learning and data mining, and other relevant applications have also emerged in economics (Lasko *et al.*, 2005). Yet in another setting, Morrison *et al.* (2003) described the ROC curve as a simple and effective method to compare the accuracies of reference variables of bacterial beach water quality. Since several fields have contributed independently to the development of ROC analysis, many concepts and techniques are often known under different names in different communities.

This paper provides an overview on some inference methods used in ROC analysis—which have been mostly employed in the medical setting—, and points out the usefulness of transferring knowledge from one field to another. The estimation target of interest is the so-called ROC curve which is a graphical representation of the relationship between false positive and true positive rates or, using an epidemiological language, it is a graphical representation of Se as a function of $1 - \text{Sp}$, where Se is the sensitivity and Sp is the specificity of a diagnostic test. Se is the probability that a truly diseased individual has a positive test result, and Sp is the probability that a truly non-diseased individual has a negative test result. Using the true/false positive/negative rates or the specificity and sensitivity, we deal with conditional probabilities of belonging to a particular predicted class given the true classification (Krzanowski and Hand, 2009), in a two-class classification (*e.g.*, diseased and nondiseased subjects, email messages are spam or not, credit card transactions are fraudulent or not).

In medicine, one of the earliest applications of ROC analysis was published in the 1960s (Lusted, 1960), although the ROC curve only gained its popularity in the 1970s (Martinez *et al.*, 2003; Zhou *et al.*, 2011). Nowadays, medical technologies offer a vast array of ways to diagnose a disease, or to predict the disease progression, and new diagnostic tests and biomarkers are continuously being studied. ROC analysis is widely used for evaluating the discriminatory performance of a continuous variable representing a diagnostic test, a marker, or a classifier.

According to different aims, the ROC analysis is useful to: (i) evaluate the discriminatory ability of a continuous marker to correctly assign into a two-group

classification; (ii) find an optimal cut-off point to least misclassify the two-group subjects; (iii) compare the efficacy of two (or more) diagnostic tests or markers; and (iv) study the inter-observer variability when two or more observers measure the same continuous variable.

Many parametric, semiparametric, and nonparametric estimation methods have been proposed for estimating the ROC curve and its associated summary measures. Here, we focus on some frequentist and Bayesian methods which have been mostly employed in the medical setting. In Section 2 we introduce notation and the basic modeling concepts. Frequentist and Bayesian approaches are reviewed in Section 3 and Section 4, respectively. The paper ends with a short discussion in Section 5.

2. DEFINITIONS AND MODELING FRAMEWORK

Let X and Y be two independent random variables, respectively denoting the diagnostic test measure for a healthy population ($D = 0$) and for a diseased population ($D = 1$), defined using a gold standard. Without loss of generality, and for an appropriate cut-off point c, the test result is positive if it is greater than c and negative otherwise.

Let F and G be the distribution functions of the random variables X and Y, respectively. The sensitivity of the test is given by $\mathrm{Se}(c) = 1 - G(c)$, and the specificity is defined as $\mathrm{Sp}(c) = F(c)$. An example is presented in Figure 1.

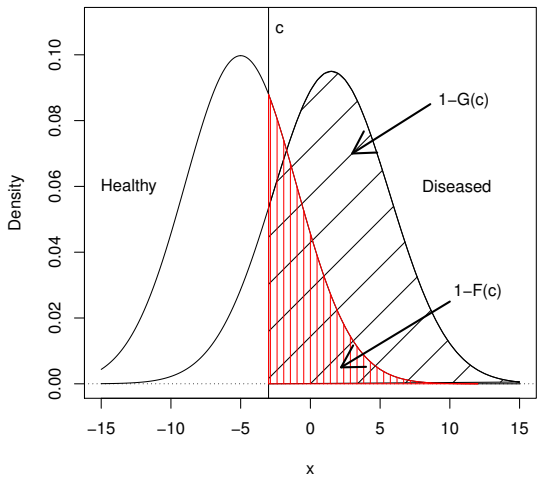


Figure 1: Distribution of the diagnostic test measures for the healthy and the diseased populations.

The ROC curve is defined as a plot of $\mathrm{Se}(c)$ versus $1-\mathrm{Sp}(c)$ for $-\infty \leq c \leq \infty$, or equivalently as a plot of

$$\mathrm{ROC}(t) = 1 - G\big(F^{-1}(1-t)\big)\,, \tag{2.1}$$

over $t \in [0,1]$, where $F^{-1}(1-t) = \inf\big\{x \in \mathbb{R}\colon F(x) \geq 1-t\big\}$.

The ROC curve is increasing and invariant under any monotone increasing transformation of the variables X and Y. Several ROC curve summary measures have been proposed in the literature, such as the area under the curve (AUC) or the Youden index ($\max_c\{\mathrm{Se}(c) + \mathrm{Sp}(c) - 1\}$). They are considered as summaries of the discriminatory accuracy of a test. The AUC is given by

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(u)\,\mathrm{d}u\ . \tag{2.2}$$

Different approaches to estimate the ROC curve lead to different estimates of the AUC. The AUC can be interpreted as the probability that, in a randomly selected pair of nondiseased and diseased individuals, the diagnostic test value is higher for the diseased subject, *i.e.*, $\mathrm{AUC} = P(Y > X)$. Values of AUC close to 1 suggest a high diagnostic accuracy of the test or marker. Bamber (1975) established an important link with the popular nonparametric test of Mann–Whitney. The area of the empirical ROC curve is equal to the Mann–Whitney U statistic that provides an unbiased nonparametric estimator for the AUC (Faraggi and Reiser, 2002). Since the seminal work of Bamber (1975), several authors have proposed refining the nonparametric approach to obtain smoothed ROC curves, for example, by using the kernel method to be described below. Parametric estimation of the ROC curve is also an active area of research and several proposals for F and G are considered. The most widely used parametric ROC model is the bi-normal, which is described in the next section.

3. FREQUENTIST METHODS

3.1. Parametric approaches

3.1.1. The bi-normal estimator

Parametric methods are used when F and G in nondiseased and diseased populations are known. The bi-normal model is commonly considered, and it is applicable when both diseased and nondiseased test outcomes follow normal distributions (Faraggi and Reiser, 2002). If data are actually bi-normal, or a Box–Cox transformation, such as the logarithm or the square root, makes the data

bi-normal, then the relevant parameters can be easily estimated by the means and variances of test values in diseased and nondiseased populations.

Let X and Y be independent normal variables with mean values μ_0, μ_1 and variances σ_0^2, σ_1^2. Then, the ROC curve can be summarized in the following way:

$$\text{ROC}(t) = \Phi\big\{a + b\,\Phi^{-1}(t)\big\}\ , \qquad 0 \le t \le 1\ , \tag{3.1}$$

where, Φ is the standard normal distribution function and a and b are the separation and the symmetry coefficients, respectively, given by $a = (\mu_1 - \mu_0)/\sigma_1$ and $b = \sigma_0/\sigma_1$. In this case, the AUC has a closed form given by

$$\text{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \tag{3.2}$$

Returning to the example presented in Figure 1, the graphical representation of the ROC curve is illustrated in Figure 2.

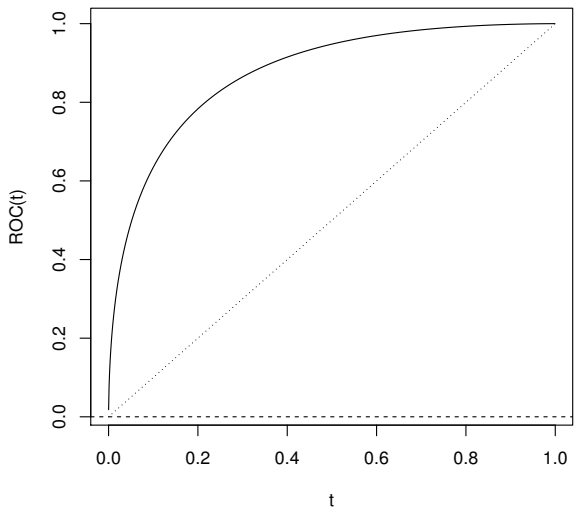


Figure 2: Example of an ROC curve for a bi-normal model, constructed using Equation (3.1).

The bi-normal model leads to convenient maximum likelihood estimates (and corresponding asymptotic variances) of the ROC curve parameters.

In this example, the normal distributions for healthy and diseased populations have the same variance and, hence, the curve is concave. Concavity is a characteristic of *proper* ROC curves (Dorfman *et al.*, 1996). This is a desirable property because it guarantees that the ROC curve will never cross the main

diagonal. Moreover, it is a property of the optimal ROC curve to establish decision rules (Huang and Pepe, 2009). However, a problem with using the bi-normal ROC model is that it is not concave in (0, 1) unless $b = 1$, as noted by Huang and Pepe (2009). Hughes and Bhattacharya (2013) characterize the symmetry properties of bi-normal and bi-gamma ROC curves in terms of the Kullback–Leibler divergences. Considering the negative diagonal of the plot, a ROC curve may be symmetric or skewed towards the left-hand axis or the upper axis of the plot. ROC curves with different symmetry properties may have the same AUC value. Not all continuous parametric ROC curves are proper. It is well known that the bi-normal ROC curve is not proper in general, while the bi-gamma ROC curve is proper (Dorfman *et al.*, 1996; Hughes and Bhattacharya, 2013). Several alternative models have been explored and compared in simulation studies, considering bi-gamma, bi-beta, bi-logistic, bi-exponential (a particular case of bi-gamma), bi-lognormal, bi-Rayleigh and even other proposals, such as the triangular distribution with constrained or unconstrained support (Dorfman *et al.*, 1996; Zou *et al.*, 1997; Marzaban, 2004; Tang *et al.*, 2010; Pundir and Amala, 2012; Tang and Balakrishnan, 2011; Hussain, 2012; Hughes and Bhattacharya, 2013).

3.1.2. Robustness of the bi-normal estimator

The choice of the bi-normal estimator to fit a ROC curve is usually justified by theoretical considerations, mathematical tractability, familiarity with the normal model or just by convenience. Hanley (1988) presents a table summarizing the most common arguments in favor of the use of this estimator. But some authors also argue that the bi-normal estimator is robust. The word robust can have many different meanings. Here it is used in the sense of robust statistics, *i.e.* meaning that in the presence of a certain amount of observations coming from a non-normal distribution the bi-normal estimator will yield reliable results. Lately, the impact of model misspecification in the parametric or semiparametric models used in health sciences is gaining importance, since practitioners are aware that theoretical models are only approximations of reality, and statistical procedures that give reliable results under model departures are essential for solving real problems. This concern is addressed by Heritier *et al.* (2009) and Farcomeni and Ventura (2010).

In the case of the bi-normal estimator of the ROC curve, authors like Swets (1986) argue that "*Empirical ROC's drawn from experimental psychology and several practical fields, (...) are fitted well on a binormal graph...*". This statement is reinforced by Hanley (1988), who claims that "*...the binormal-based fits are certainly good enough for all practical purposes.*". Hajian-Tilaki *et al.* (1997) state that, "*The results suggested that the AUC is robust to departures from binomality if one uses the binormal model as implemented in* `LARROC` *program.*". Neverthe-

less, these authors were more cautious adding that a possible explanation relies in the use of ranks instead of the original data, in both estimation procedures.

Walsh (1997) clarifies these arguments. Robustness, in Swets (1986) and Hanley (1988), is understood as the ability of the bi-normal estimator to fit a ROC curve that 'looks right' in comparison either with the theoretical ROC curve or with the observed rating method. But this author goes further, discussing the ability of the bi-normal estimator to produce valid inferences in circumstances in which the data does not satisfy the normality assumption. A simulation study to analyze the impact of data coming from a bi-logistic model combined with bi-normal estimator was developed to study: (i) the AUC estimator, (ii) the performance of the statistical test to compare AUC from two ROC curves, and (iii) the impact on size and power of this statistical test. The choice of the bi-logistic distributions to model departures from bi-normal assumption relies on the difficulty to distinguish these models, since the logistic model was considered one of the possible hardest scenarios to detect departures from the normality assumption. In his simulation study, Walsh also considers the effect of different sets of decision thresholds, and concludes that the bi-normal estimator is sensitive to model misspecification and to the location of the decision thresholds.

The problem of robustness has deserved the attention of other authors. Greco and Ventura (2011) develop an M-estimator for the $P(Y > X)$ in the context of a stress-strength model, that has direct application in AUC estimation. Recently, Devlin *et al.* (2013) discuss the impact of model misspecification in three estimators resulting from modeling the parametric form of the ROC curve directly.

3.2. Nonparametric estimation of the ROC curve

3.2.1. Empirical estimator and variants

The simplest nonparametric method is the empirical estimator, which is based plugging in empirical estimates into (2.1). Specifically, the empirical estimate of the ROC curve is given by

$$\widetilde{\mathrm{ROC}}(t) = 1 - \widetilde{G}\big(\widetilde{F}^{-1}(1-t)\big)\,, \tag{3.3}$$

where $\widetilde{F}^{-1}$ and $\widetilde{G}$ respectively denote the empirical quantile function and the empirical distribution function associated to healthy and diseased populations; roughly speaking, the empirical distribution function is defined, for any given value t, as the percentage of sample points smaller or equal to t.

The empirical ROC curve preserves many properties of the empirical distribution function and it is uniformly convergent to the theoretical curve (Hsieh and Turnbull, 1996). Nevertheless, the estimator has some drawbacks, and it may suffer from large variability, particularly for small sample sizes (Lloyd, 1998; Lloyd and Yong, 1999; Jokiel-Rokita and Pulit, 2013). While this is not a major problem in machine learning, data mining, and finance—where large samples are common—in medicine this may be inadequate, as small samples are commonplace in clinical practice. In addition to all this, the estimated ROC curve is not continuous, and thus its interpretation becomes more complex (Jokiel-Rokita and Pulit, 2013).

Other methods have been explored to obtain smooth ROC curve estimates, either through kernel smoothing (Lloyd, 1998; Lloyd and Yong, 1999) or through smooth versions of the empirical distribution function (Jokiel-Rokita and Pulit, 2013).

3.2.2. Kernel estimator

To overcome the lack of smoothness of the empirical estimator, Zou *et al.* (1997) used kernel methods to estimate the ROC curve, which were later improved by Lloyd (1998). Kernel density estimators are known to be simple, versatile, with good theoretical and practical properties (Silverman, 1986; Tenreiro, 2010), merits that the corresponding ROC curve estimator inherit.

Let $(x_1, ..., x_n)$ and $(y_1, ..., y_m)$ be two independent samples from X and Y, respectively. The kernel density estimators of f and g, the probability density functions associated with F and G, are:

$$\widehat{f}(x) = \frac{1}{nh_0} \sum_{i=1}^{n} K_0\left(\frac{x - x_i}{h_0}\right), \qquad \widehat{g}(y) = \frac{1}{mh_1} \sum_{i=1}^{m} K_1\left(\frac{y - y_i}{h_1}\right).$$

Here the $h_i > 0$ are bandwidths, which are used to control the amount of smoothness, and the K_i are kernel functions, that obey (i) $\int_{\mathbb{R}} K_i(x)\,\mathrm{d}x = 1$, (ii) $\int_{\mathbb{R}} x\,K_i(x)\,\mathrm{d}x = 0$, and (iii) $\int_{\mathbb{R}} x^2 K_i(x)\,\mathrm{d}x > 0$, for $i = 0, 1$. Using these estimators, the cumulative distribution functions can be estimated as

(3.4)

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{x} \frac{1}{h_0} K_0\left(\frac{u - x_i}{h_0}\right) \mathrm{d}u\,, \quad \widehat{G}(y) = \frac{1}{m} \sum_{i=1}^{m} \int_{-\infty}^{y} \frac{1}{h_1} K_1\left(\frac{v - y_i}{h_1}\right) \mathrm{d}v\,.$$

These integrals can be evaluated numerically. The choice of the kernels K_0 and K_1 among the available proposals is not problematic, since they all give comparable results, as was pointed out by Krzanowski and Hand (2009) and Jokiel-Rokita and Pulit (2013). This justifies the pragmatic option of using equal kernels, and

a popular option is the Gaussian kernel (Sheather, 2004; Hong *et al.*, 2007; Zhou *et al.*, 2011; Fabsic, 2012), and in this case Equation (3.4) can be written as

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x-x_i}{h_0}\right), \qquad \widehat{G}(x) = \frac{1}{m}\sum_{i=1}^{m}\Phi\left(\frac{y-y_i}{h_1}\right). \tag{3.5}$$

Plugging-in (3.4) into (2.1) leads to the kernel-based ROC curve estimator:

$$\widehat{\mathrm{ROC}}(t) = 1 - \widehat{G}\big(\widehat{F}^{-1}(1-t)\big). \tag{3.6}$$

The most sensitive aspect of the kernel-based ROC curve estimator in (3.6) is the choice of the 'optimal' bandwidth (Zhou and Harezlak, 2002; Hall and Hyndmann, 2003; Zhou *et al.*, 2011; Jokiel-Rokita and Pulit, 2013). This, combined with the selection of K determines the properties of the estimator. Zou *et al.* (1997) used bandwidths that are asymptotically optimal for estimating f and g. Lloyd (1998) improved the previous proposal by choosing bandwidths that are asymptotically optimal for estimating F and G, since the ROC curve depends directly on these cumulative distribution functions. Lloyd and Yong (1999) showed how kernel density estimators overcome the empirical ones. Qiu and Le (2001) proposed a ROC curve estimator based on a kernel distribution function estimator to G and a local smoothing quantile function estimator to F^{-1}. Peng and Zhou (2004) introduced another kernel estimator involving only one bandwidth, estimated in an optimal asymptotical way, that has better performance near the boundary of the support of X and Y. Koláček and Karunamuni (2009) proposed a related kernel-based estimator for the ROC curve that removes the boundary effects. Contrasting with these approaches, Jokiel-Rokita and Pulit (2013) proposed a strongly consistent estimator based on a smoothed version of the empirical ROC curve that, according to a simulation study, outperformed the empirical and a kernel estimator for small sample sizes.

Kernel-based estimators can also be used for estimating the AUC. For example, using the estimators proposed by Lloyd (1998) and a Gaussian kernel, yields the following estimator

$$\widehat{\mathrm{AUC}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\Phi\left(\frac{y_j - x_i}{\sqrt{h_0^2 + h_1^2}}\right). \tag{3.7}$$

See Fabsic (2012), for a simulation study comparing several parametric and non-parametric methods.

4. BAYESIAN METHODS

4.1. Introduction

Bayesian methods are introduced in ROC curve estimation as an alternative to maximum likelihood methods. Bayesian approaches enable the introduction of prior information into the estimation process, which reduces the uncertainty of the inferences. This point is specially important when a gold standard test, which correctly classifies all subjects as healthy or diseased, is unavailable, either because there is no gold standard for the disease or because the procedure is costly, technically demanding, harmful or even life-threatening. In this framework, the true state of the individuals is unknown and the modeling process may benefit from including existing information about the problem under study through the use of prior distributions.

The Bayesian framework enables obtaining credibility intervals for the ROC curve and for other summary measures, such as the AUC. As it is known, one of the benefits of the Bayesian methodology is the capability of producing regions in terms of the posterior distributions of the parameters. These regions, contrarily to confidence intervals resulting from frequentist analysis, allow for probabilistic interpretations of the inferences. Additionally, predictive probabilities of the health status of future individuals can be obtained through the predictive distribution. Furthermore, the Bayesian perspective is specially suited to model complex designs, namely through the use of hierarchical structures (Ishwaran and Gatsonis, 2000; O'Malley and Zou, 2006; Johnson and Johnson, 2006).

It is well known that the ability of a diagnostic test to discriminate between diseased and healthy populations, may be influenced by various factors (Pepe, 2003). Moreover, assessing the covariate impact may provide useful information regarding the test adequacy towards different populations and conditions (de Carvalho *et al.*, 2013). On the contrary, neglecting covariate effects may lead to biased inferences about the test performance. Covariate effects on the ROC curves are addressed in several works (*e.g.* Peng and Hall, 1996; Branscum *et al.*, 2008; de Carvalho *et al.*, 2013).

Traditionally, in a Bayesian framework, ROC curve estimation has been explored in a parametric manner. More recently, semiparametric and nonparametric methodologies have also been developed. In the next subsections some of these approaches will be described.

4.2. Parametric approaches

Some of the first accounts of using a Bayesian methodology in ROC curve estimation are based on regression models (Peng and Hall, 1996; Hellmich *et al.*, 1998). Probit-linked generalized linear regression models are applied to ordinal test results, leading to Bayesian inferences for ROC curves and functionals such as the AUC. In particular, the approach adopted by Peng and Hall (1996) admits latent bi-normal distributions for diseased and nondiseased populations, even though other parametric distributions could be considered. The authors use data augmentation techniques to impute unobserved continuous data from the latent distribution, thus allowing to overcome the difficulties due to the ordinal nature of the observations. Noninformative priors are applied. This ordinal regression model can explain modifications observed in the ROC curves caused by changing the value of a single covariate.

As mentioned earlier, some regression approaches to ROC curve analysis consider hierarchical structures (O'Malley and Zou, 2006; Johnson and Johnson, 2006). A Bayesian multivariate hierarchical transformation model is developed by O'Malley and Zou (2006) based on clustered continuous diagnostic test data with covariates. This approach is useful in the context of multilevel data with clustered responses, like, for example, radiologic data collected from patients (individual level) nested in different hospitals (clusters). The authors aim to model the diagnostic test accuracy and define a composite diagnostic test. The authors remark that a cluster-specific transformation of the outcomes is applied to handle the heterogeneity between the clusters and that multiple correlated outcomes may be used. The methodology is applied to prostate cancer biopsy data gathered from a multi-center clinical trial.

Johnson and Johnson (2006) address a situation frequently observed in radiology, in which several radiologists rate, in an ordinal scale, multiple exams collected from the same individual. A Bayesian hierarchical latent variable model for analyzing multirater correlated ordinal data is proposed. The three sources of variation (differences in patients characteristics, in diagnostic exams and in raters) are explicitly modeled, each one corresponding to a different level of the model hierarchy. Simulation studies show that this model is more efficient than the most widely used model for multirater correlated data analysis (Dorfman *et al.*, 1992).

4.3. Semiparametric and nonparametric approaches

Bayesian semiparametric and nonparametric approaches have been used for ROC curve estimation in the last few years (Erkanli *et al.*, 2006; Wang *et al.*, 2007; Gu *et al.*, 2008; Branscum *et al.*, 2013). These methodologies are still being developed and constitute a very active line of research.

Nonparametric Bayesian methods are meant to overcome the restrictions imposed by considering a fixed parametric model and the consequent difficulties in capturing nonstandard data features, such as multimodality and skewness. Contrarily to the traditional parametric framework, the nonparametric framework enables a more flexible modeling of the data, in the sense that no specific parametric family of distributions is considered.

The nonparametric approach entails a modeling framework that requires specifying a prior distribution over the space of all probability measures. As pointed out by Inácio (2012), this does not mean an absence of parameters in the model, on the contrary it involves an (possibly) infinite number of parameters. In this framework, Dirichlet processes, mixtures of Dirichlet processes, Polya trees, and mixtures of Polya trees are frequently used priors; for further details on this see Inácio (2012), and references therein.

A Bayesian semiparametric approach for ROC curve estimation method, based on mixtures of Dirichlet processes, was developed by Erkanli *et al.* (2006). A Gibbs sampling framework is used to obtain posterior distributions of the mixtures of Dirichlet processes model, thus providing posterior predictive estimates of sensitivity, specificity, ROC curves and AUC. The authors show that, even when a gold standard diagnostic test is not available, the results still stand. Moreover, it closely parallels the kernel density estimation approach, previously referred to in this paper.

A nonparametric Bayesian method reported by Hanson *et al.* (2008) uses Dirichlet process mixtures and mixtures of Polya trees for analyzing continuous serologic data. A novelty of this approach is the inclusion of a stochastic ordering constraint for the serologic values distributions of the infected and noninfected populations. This is a biologically reasonable assumption, since the serologic scores tend to be higher for the infected individuals than for the noninfected ones. According to the authors, the approach has the benefit of guaranteeing that the AUC is always larger than 0.5, meaning that the ROC curve never goes below the main diagonal. The two models are applied to Johne's disease data observed in dairy cattle. Qualitatively similar inferences are obtained and the same conclusions, regarding the accuracy of the serologic tests, can be drawn from both applications.

In the Bayesian nonparametric context, few works study the effect of covariates in ROC curve estimation. This issue is explored by de Carvalho *et al.* (2013). The model is based on dependent Dirichlet processes and allows the entire distribution in each group to smoothly change as a function of the covariates. This approach can accommodate multiple continuous and categorical predictors. An approximated version of the general model, based on B-splines, was compared with the semiparametric approach of Pepe (1998), with an extension of the previous approach that uses a B-splines trend and with the nonparametric kernel estimator of Rodríguez-Álvarez *et al.* (2011). The proposed model outperforms its competitors for nonlinear scenarios and small sample sizes. An application of the model to diabetes diagnosis is presented.

As explained by Inácio (2012), ROC surfaces have been proposed for the evaluation of the diagnostic accuracy in ordered three-class problems as a direct generalization of the ROC curve. A flexible Bayesian nonparametric approach based on mixtures of finite Polya trees priors is described by Inácio (2012).

The bootstrap has been used to ROC curve estimation by Gu *et al.* (2008). The authors also present estimation credible intervals of the ROC curve and apply the approach for testing the validity of the bi-normal assumption.

4.4. Absence of a gold standard

Imperfect diagnostic tests are widely used in medicine and, as we pointed out earlier, the Bayesian methodology is particularly suited for problems of this nature (Krzanowski and Hand, 2009).

Returning to the previously mentioned work of Erkanli *et al.* (2006), an extension of the nonparametric model to the case of imperfect reference test is given, in which a binary latent variable is introduced to express the true but unknown disease status. Extensive literature exists on the use of latent class models to evaluate the performance of binary diagnostic tests in the absence of a gold standard, either using maximum likelihood or Bayesian estimation methods (see Gonçalves *et al.*, 2012, and references therein).

Again, in the context of no gold standard data analysis, Choi *et al.* (2006) develop a parametric Bayesian methodology that admits two diagnostic tests applied to the same individuals. The data are modeled under the bi-normal assumption; this assumption may require a suitable transformation, which can be difficult to find in the absence of a gold standard test. Training data or previous studies with a gold standard could suggest an adequate transformation. The method is initially formulated for the gold standard case and slightly modified to address the gold standard absence. A latent variable indicating the true disease

state is introduced, resembling Erkanli *et al.* (2006). The method has difficulty in assigning the correct disease status when the overlap of diseased and nondiseased groups is too large.

Wang *et al.* (2007) explore the problem of estimating the ROC curve of a new ordinal or continuous scale diagnostic test by comparison with an imperfect binary reference test, assuming conditional independence between the two tests. Identifiability problems require data from at least two populations with different prevalences. The method is based on a multinomial model and no assumptions are needed concerning the shape of the distributions corresponding to the test values. Care is taken in guaranteeing the monotonicity of the ROC curve.

Both Choi *et al.* (2006) and Wang *et al.* (2007) illustrate their methods using different datasets from Johne's disease in cattle.

A group of Bayesian latent class models for mixed continuous and discrete diagnostic test data is explored by Weichenthal *et al.* (2010). These models are used to determine the probability of asbestos exposure from lung fiber count data. The model admits correlations between repeated measurements of the same test within individuals.

Branscum *et al.* (2008) propose Bayesian nonparametric and semiparametric approaches to ROC analysis and disease diagnosis in the absence of a gold standard. A nonparametric model using mixtures of Polya trees is proposed to estimate probabilities of disease risk and the ROC curve. Semiparametric extensions of this model are also proposed. These semiparametric models incorporate additional information regarding the disease status. Two types of information are used: standard covariate information and information from additional binary diagnostic tests. Such additional information improves the discriminatory ability to correctly classify subjects as healthy or diseased, leading to a modeling process in between the gold standard case and the nonparametric modeling in the absence of a gold standard. This is a very flexible approach that allows combining in a single framework available information on risk factors and additional diagnostic tests outcomes to enhance diagnostic predictive accuracy.

Nonparametric Bayesian analysis involving Polya tree priors is also dealt with in Branscum *et al.* (2013). The usefulness of the discussed flexible models over a standard parametric method is shown in an application to a lung cancer biomarker.

5. FINAL REMARKS

Statistical modeling of ROC curves is a vast topic and offers several future research lines. The use of flexible models that accommodate covariates and prior information is an active field of research. If proper ROC curves are desired in many applications, in Bioinformatics not proper ROC curves have been increasingly used as new tools for the analysis of differentially expressed genes in microarray experiments (*e.g.* Parodi *et al.*, 2008; Silva-Fortes *et al.*, 2012). A particularly relevant issue in this setting is robustness, but further research is definitely required on this.

ACKNOWLEDGMENTS

We would like to express our gratitude to the Editors for the invitation to present this paper. This work has been supported by *Fundação para a Ciência e Tecnologia* (FCT), Portugal, through PTDC/MAT/118335/2010, PEst-OE/MAT/UI0006/2014 and PEst-OE/MAT/UI0822/2014. A. Subtil acknowledges financial support from the FCT grant SFRH/BD/69793/2010.

REFERENCES

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

Branscum, A. J.; Johnson, W. O. and Baron, A.T. (2013). Robust medical test evaluation using flexible bayesian semiparametric regression models, *Epidemiology Research International*, ID 131232.

Branscum, A. J.; Johnson, W. O.; Hanson, T. E. and Gardner, I. A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis, *Statistics in Medicine*, **27**, 2474–2496.

Choi, Y.-K.; Johnson, W. O.; Collins, M. T. and Gardner, I. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard, *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 210–229.

Collison, P. (1998). Of bombers, radiologists, and cardiologists: time to ROC, *Heart*, **80**, 215–217.

de Carvalho, V. I.; Jara, A.; Hanson, T. E. and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling, *Bayesian Analysis*, **8**, 623–646.

Devlin, S. A.; Thomas, E. G. and Emerson, S. S. (2013). Robustness of approaches to ROC curve modeling under misspecification of the underlying probability model, *Communications in Statistics—Theory and Methods*, **42**, 3655–3664.

Dorfman, D. D.; Berbaum, K. S. and Metz, C. E. (1992). Receive operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method, *Investigative Radiology*, **27**, 723–731.

Dorfman, D. D.; Berbaum, K. S.; Metz, C. E.; Lenth, R. V.; Hanley, J. A. and Dagga, H. A. (1996). Proper receiver operating characteristic analysis: the bigamma model, *Academic Radiology*, **4**, 138–149.

Erkanli, A.; Sung, M.; Costello, J. E. and Angold, A. (2006). Bayesian semi-parametric ROC analysis, *Statistics in Medicine*, **25**, 3905–3928.

Fabsic, P. (2012). *Comparing the Accuracy of ROC Curve Estimation Methods*, Master Thesis, Swiss Federal Institute of Technology, Zurich.

Farcomeni, A. and Ventura, L. (2010). An overview of robust methods in medical research, *Statistical Methods in Medical Research*, **21**, 111–133.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve, *Statistics in Medicine*, **21**, 3093–3096.

Gonçalves, L.; Subtil, A.; Oliveira, M. R.; Rosário, V.; Lee, P. and Shaio, M.-F. (2012). Bayesian latent class models in malaria diagnosis, *PLoS ONE*, **7**, e40630.

Greco, L. and Ventura, L. (2011). Robust inference for the stress-strength reliability, *Statistical Papers*, **52**, 773–788.

Gu, J.; Ghosal, S. and Roy, A. (2008). Bayesian bootstrap estimation of ROC curve, *Statistics in Medicine*, **27**, 5407–5420.

Hajian-Tilaki, K. O.; Hanley, J. A.; Joseph, L. and Collet, J. P. (1997). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests, *Medical Decision Making*, **17**, 94–102.

Hall, P. G. and Hyndmann, R. J. (2003). Improved methods for bandwidth selection when estimating ROC curves, *Statistics and Probability Letters*, **64**, 181–189.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine Learning*, **77**, 103–123.

Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves, *Medical Decision Making*, **8**, 197–203.

Hanson, T. E.; Kottas, A. and Branscum, A. J. (2008). Modelling stochastic order in the analysis of Receiver Operating Characteristic data: Bayesian nonparametric approaches, *Journal of the Royal Statistical Society. Ser.* C, **57**, 207–225.

Hellmich, M.; Abrams, K. R., Jones, D. R. and Lambert, P. C. (1998). A Bayesian approach to a general regression model for ROC curves, *Medical Decision Making*, **18**, 436–443.

Heritier, S.; Cantoni, E.; Copt, S. and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*, John Wiley & Sons, Chichester.

Hong, X.; Chen, S. and Harris, C. J. (2007). A kernel-based two-class classifier for imbalanced data sets, *IEEE Transactions on Neural Networks*, **18**, 28–41.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *The Annals of Statistics*, **24**, 25–40.

Huang, Y. and Pepe, M. S. (2009). A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies, *Biometrics*, **65**, 1133–1144.

Hughes, G. and Bhattacharya, B. (2013). Symmetry properties of binormal and bi-gamma receiver operating characteristic curves are described by Kullback–Leibler divergences, *Entropy*, **15**, 1342–1356.

Hussain, E. (2012). The bi-gamma ROC curve in a straightforward manner, *Journal of Basic & Applied Sciences*, **8**, 309-314.

Inácio, V. (2012). *Semiparametric and Nonparametric Modeling of Diagnostic Data*, PhD Thesis, University of Lisbon, Portugal.

Johnson, T. D. and Johnson, V. E. (2006). A Bayesian hierarchical approach to multirater correlated ROC analysis, *Statistics in Medicine*, **25**, 1858–1871.

Jokiel-Rokita, A. and Pulit, M. (2013). Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions, *Statistics and Computing*, **23**, 703–712.

Ishwaran, H. and Gatsonis, A. C. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis, *The Canadian Journal of Statistics*, **28**, 731–750.

Koláček, J. and Karunamuni, R. J. (2009). On boundary correction in kernel estimation of ROC curves, *Austrian Journal of Statistics*, **38**, 17–32.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*, Chapman & Hall/CRC, Boca Raton.

Lasko, T. A.; Bhagwat, J. G.; Zou, K. H. and Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics, *Journal of Biomedical Informatics*, **38**, 404–415.

Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems, *Journal of the American Statistical Association*, **93**, 1356–1364.

Lloyd, C. J. and Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical, *Statistics and Probability Letters*, **44**, 221–228.

Lusted, L. B. (1960). Logical analysis in roentgen diagnosis, *Radiology*, **74**, 178–193.

Martinez, E. Z.; Louzada-Neto, F. and Pereira, B. B. (2003). A curva ROC para testes diagnósticos (Document in Portuguese), *Cadernos Saúde Coletiva*, **11**, 7–31.

Marzaban, C. (2004). The ROC curve and the area under it as performance measures, *Weather and Forecasting*, **19**, 1106–1114.

Morrison, A. M.; Coughlin, K.; Shine, J. P.; Coull, B. A. and Rex A. C. (2003). Receiver operating characteristic curve analysis of beach water quality indicator variables, *Applied and Enviromental Microbiology*, **69**, 6405–6411.

O'Malley, A. J. and Zou, K. H. (2006). Bayesian multivariate hierarchical transformation models for ROC analysis, *Statistics in Medicine*, **25**, 459–479.

Parodi, S.; Pistoia, V. and Muselli, M. (2008). Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments, *BMC Bioinformatics*, **8**, 410.

Peng, F. and Hall, W. J. (1996). Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods, *Medical Decision Making*, **16**, 404–411.

Peng, L. and Zhou, X.-H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves, *Journal of Statistical Planning and Inference*, **118**, 129–143.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results, *Biometrics*, **54**, 124–135.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.

Pundir, S. and Amala, R. (2012). A study on the bi-Rayleigh ROC curve model, *Bonfring International Journal of Data Mining*, **2**, 42–47.

Qiu, P. and Le, C. (2001). ROC curve estimation based on local smoothing, *Journal of Statistical Computation and Simulation*, **70**, 55–69.

Rodríguez-Álvarez, M. X.; Roca-Pardiñas, J. and Cadarso-Suárez, C. (2011). ROC curve and covariates: extending induced methodology to the non-parametric framework, *Statistics and Computing*, **21**, 483–499.

Sheather, S. J. (2004). Density estimation, *Statistical Science*, **19**, 588–597.

Silva-Fortes, C.; Amaral Turkman, M. A. and Sousa L. (2012). Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups, *BMC Bioinformatics*, **13**, 147.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance, *Psycological Bulletin*, **99**, 181–198.

Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum Associates, New Jersey.

Tang, L.; Du, P. and Wu, C. (2010). Compare diagnostic tests using transformation-invariant smoothed ROC curves, *Journal of Statistical Planning and Inference*, **140**, 3540–3551.

Tang, L. L. and Balakrishnan, N. (2011). A random-sum Wilcoxon statistic and its application to analysis of ROC and LROC data, *Journal of Statistical Planning and Inference*, **141**, 335–344.

Tenreiro, C. (2010). Uma introdução à estimação não-paramétrica da densidade (Document in Portuguese), *Edições SPE*, Lisboa.

Walsh, S. J. (1997). Limitations to the robustness of binormal ROC curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power, *Statistics in Medicine*, **16**, 669–679.

Wang, C.; Turnbull, B. W.; Gröhn, Y. T. and Nielsen, S. S. (2007). Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown, *Journal of Agricultural, Biological, and Environmental Statistics*, **12**, 128–146.

Weichenthal, S.; Joseph, L.; B é lisle, P. and Dufresne, A. (2010). Bayesian estimation of the probability of asbestos exposure from lung fiber counts, *Biometrics*, **66**, 603–612.

Zhou, X.-H. and Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves, *Statistics in Medicine*, **21**, 2045–2055.

Zhou, X-H.; Obuchowski, N. A. and McClisk, D. K. (2011). *Statistical Methods in Diagnostic Medicine*, Second Edition, John Wiley & Sons, New Jersey.

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

REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 21–41

A REVIEW ON ROC CURVES IN THE PRESENCE OF COVARIATES

Authors: Juan Carlos Pardo-Fernández
– Department of Statistics and Operations Research, Universidade de Vigo, Spain
juancp@uvigo.es

María Xosé Rodríguez-Álvarez
– Department of Statistics and Operations Research, Universidade de Vigo, Spain
mxrodriguez@uvigo.es

Ingrid Van Keilegom
– Institute of Statistics, Université catholique de Louvain, Belgium
ingrid.vankeilegom@uclouvain.be

Abstract:

- In this paper we review the literature on ROC curves in the presence of covariates. We discuss the different approaches that have been proposed in the literature to define, model, estimate and do asymptotics for ROC curves that incorporate covariates. For reasons of brevity, we mostly focus on nonparametric approaches, although some parametric and semiparametric methods are also discussed. We also analyze endocrinological data on the body mass index to illustrate the methodology. Finally, we mention some research topics that need further investigation or that are still unexplored.

Key-Words:

- *covariate; kernel smoothing; location-scale model; nonparametric inference; regression; ROC curve.*

AMS Subject Classification:

- 62-02, 62G08, 62H30.

1. INTRODUCTION

ROC curves are a very useful instrument to measure how well a variable or a diagnostic test is able to distinguish two populations from each other. It is therefore an essential element in the classification and discrimination literature, and it has interested and still interests many statisticians from a theoretical as well as from an applied point of view.

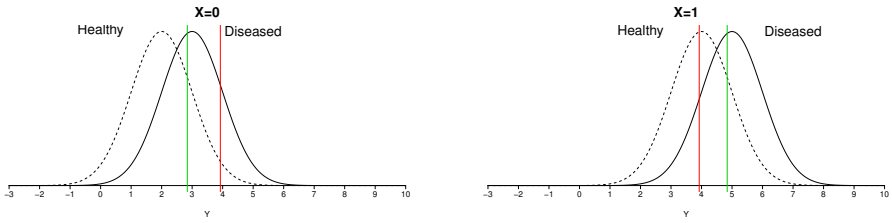
When covariates are present, it might be advisable to incorporate them in the ROC curve in order to make use of the additional information. In fact, in many situations the performance of a diagnostic test and, by extension, its discriminatory capacity can be affected by covariates. Pepe (2003, pp. 48–49) provides several examples of covariates that can affect a test result. For instance, patient characteristics, such as age and gender, are important covariates to be considered. Furthermore, when a diagnostic test is performed by a tester (*e.g.*, a radiologist engaged in interpreting images), a characteristic of the tester, such as experience or expertise, will often affect the test result. The incorporation of covariates into the ROC curve might be done for two purposes: (a) obtain covariate-specific ROC curves, or ROC curves that condition on a specific value of a covariate vector; and (b) get some kind of average ROC-curve, or covariate-adjusted ROC curve, which takes the covariate information of each data point into account in order to obtain a better measure of the discriminatory capacity than the rude 'marginal' or 'pooled' ROC curve.

In this paper we first explain in Section 2 why it is important to take covariate information into account by giving some concrete examples of situations where the covariates have an impact on the performance of the diagnostic test and/or its discriminatory capacity. We next consider in Section 3 both the covariate-specific and the covariate-adjusted ROC curve, and we give an overview of estimation methods that have been proposed for both concepts in Section 4. The focus lies on reviewing the literature and not on giving detailed derivations or lengthy discussions. They can be found in the respective papers. For reasons of brevity, we mostly focus on nonparametric approaches, although some parametric and semiparametric methods are also discussed. In Section 5 we analyze endocrinological data on the body mass index to illustrate the methodology. Finally, in Section 6 we mention some research topics that need further investigation or that are still unexplored.

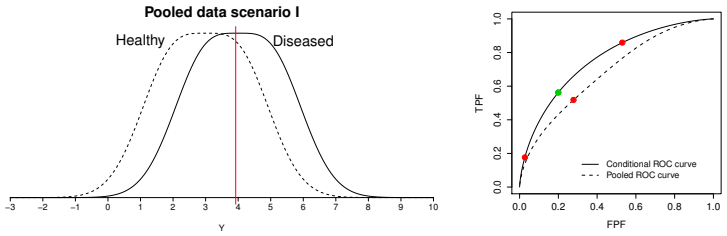
2. MOTIVATION

This section is devoted to motivating the need for incorporating covariates into the ROC analysis by means of illustrating the consequences that ignoring such information may have on the practical conclusions drawn from the study at hand. In brief, there are two different scenarios on which covariate information will have to be incorporated into the ROC analysis: (a) when the performance of the diagnostic test is affected by covariates, but not its discriminatory capacity; and, (b) when the discriminatory capacity itself is affected. A good overview of this aspect can be found in Janes and Pepe (2008, 2009a) and in fact, the examples given here are partially based on both papers. For a more detailed review of the subject, readers are urged to consult these references.

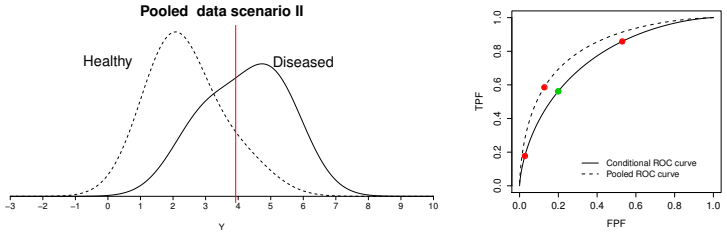
On the one hand, let us start with those situations in which the performance of the diagnostic test is affected by covariates, even where the discriminatory capacity of the test is unaffected. This situation is depicted in Figure 1(a), in which a covariate $\boldsymbol{X}$ (*e.g.*, patient gender) affects the result but not the discriminatory capacity of diagnostic test Y. In other words, the separation between the conditional distributions of the diagnostic test result in both healthy and diseased populations is the same, irrespective of the values of covariate $\boldsymbol{X}$. In Figure 1(b), covariate $\boldsymbol{X}$ is independent of disease status, which will be denoted by D (diseased) and $\bar{D}$ (healthy), *i.e.*, the result of the diagnostic test changes according to the gender of the patient but the prevalence of the disease is the same for both genders. In such a case, when data are pooled regardless of the gender of the patient, the obtained ROC is attenuated with respect to the ROC curve in each of the populations determined by covariate $\boldsymbol{X}$. However, if covariate $\boldsymbol{X}$ is associated with disease status, the pooled ROC curve will also 'incorporate' the portion of discrimination attributable to the covariate. This situation can lead to a pooled (or marginal) ROC curve that lies above or below the conditional ROC curve (see Figures 1(c) and 1(d)). It should be noted that, despite the fact that in the previous examples the discriminatory capacity of the diagnostic test is the same for both populations defined by covariate $\boldsymbol{X}$, the threshold that gives rise to a pair of values for the FPF (false positive fraction) and the TPF (true positive fraction) could not coincide in each population. This is also illustrated in Figure 1. The red lines and dots represent a common threshold used to define test positivity. As can be observed, this threshold provides a different pair of FPF and TPF on $\boldsymbol{X}=1$ and $\boldsymbol{X}=0$, as well as on the pooled data. On the other hand, the green lines and dots depict the threshold to be used to ensure a FPF$=0.2$ in both populations. Accordingly, studying the effect of covariates on the distribution of a diagnostic test in the healthy/diseased population will enable assessment of which factors affect the FPF/TPF when a specific threshold value is set. Conversely, different threshold values can be chosen for each of the populations determined by the covariates, in order to ensure that the FPF/TPF remains constant across all of them.



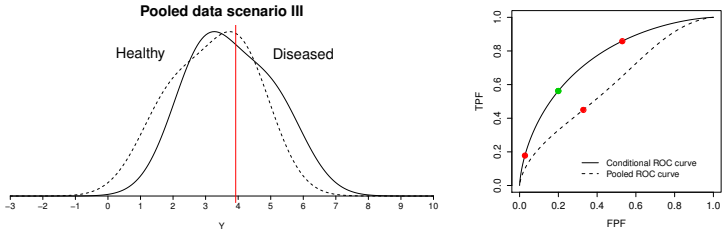
(a)



(b) Scenario I



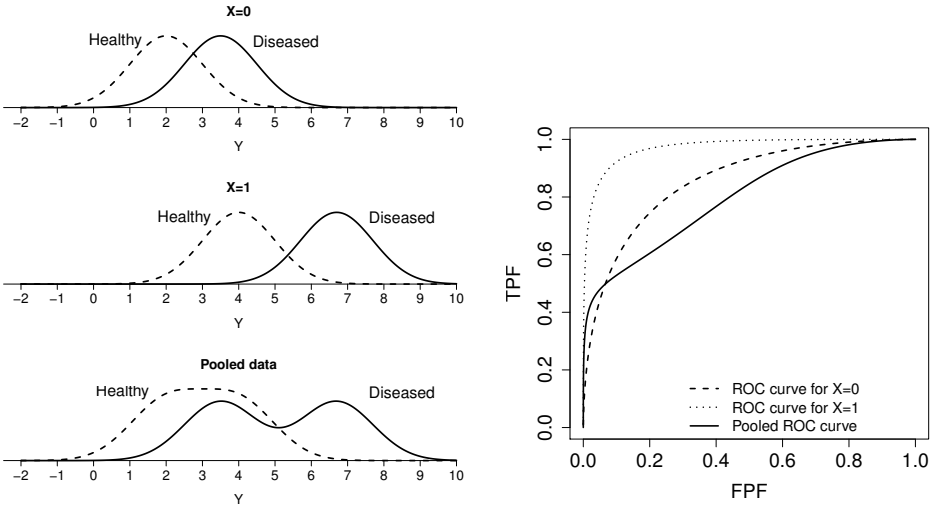
(c) Scenario II



(d) Scenario III

Figure 1: (a) Probability distributions of a hypothetical diagnostic test Y in diseased (solid line) and healthy (dashed line) populations conditional on a binary covariate $\boldsymbol{X} = 0, 1$.
Shown in (b), (c) and (d) are the pooled probability distributions (left panel), and the corresponding pooled ROC curves, along with the common conditional ROC curves (right panel).
Scenario I: disease status and covariate are independent, $P(\text{status } D \mid \boldsymbol{X} = 0) = 0.5$ and $P(\text{status } D \mid \boldsymbol{X} = 1) = 0.5$.
Scenario II: $P(\text{status } D \mid \boldsymbol{X} = 0) = 0.2$ and $P(\text{status } D \mid \boldsymbol{X} = 1) = 0.8$.
Scenario III: $P(\text{status } D \mid \boldsymbol{X} = 0) = 0.6$ and $P(\text{status } D \mid \boldsymbol{X} = 1) = 0.4$.
In all cases $P(\text{status } D) = 0.5$ and $P(\boldsymbol{X} = 1) = 0.5$ were considered. The performance of the common threshold 3.9 is also indicated (red lines and dots), as well as the common conditional threshold that gives rise to a FPF $= 0.2$ in both the populations determined by covariate X (green lines and dots).

On the other hand, in those situations where the accuracy of a diagnostic test is affected by covariates, failure to incorporate information furnished by them may lead, as in the previous cases, to erroneous conclusions. For instance, let us consider the example shown in Figure 2, where the accuracy of a diagnostic test changes according to a binary covariate $\boldsymbol{X}$ (with $\boldsymbol{X}$ again assumed to be patient gender). The conditional ROC curve shows that test Y is more accurate when $\boldsymbol{X} = 1$ than when $\boldsymbol{X} = 0$, though discriminatory capacity is high in both cases. Nevertheless, pooling the data regardless of the values of the covariate yields a ROC curve that is below the specific ROC curves for each of the populations determined by covariate $\boldsymbol{X}$. Taking into account the possible modifying effect of covariates on the accuracy of a diagnostic test, *i.e.*, on the ROC curve, will help identifying the optimal populations to whom or conditions under which the test should be applied, or alternatively, those where the test is unlikely to be useful. Furthermore, different thresholds for defining test positivity can be chosen to vary with covariate values.



(a) (b)

Figure 2: (a) Probability distributions of Y in diseased (solid line) and healthy (dashed line) populations conditional on $\boldsymbol{X}$ and pooled probability distributions.
(b) Conditional ROC curve in each of the populations determined by covariate $\boldsymbol{X}$, together with the pooled ROC curve.
The shown results were obtained assuming that the performance and discriminatory capacity of the diagnostic test depend on $\boldsymbol{X}$, but $\boldsymbol{X}$ is independent of true disease status: $P(\text{status } D \mid \boldsymbol{X} = 1) = P(\text{status } D \mid \boldsymbol{X} = 0) = 0.5$. Moreover, $P(\text{status } D) = 0.5$ and $P(\boldsymbol{X} = 1) = 0.5$ were considered.

Summarising, both in situations where the result of a diagnostic test, though not necessarily its discriminatory capacity, is affected by covariates, and in those where the discriminatory capacity itself is affected by covariates, this information must be incorporated into the ROC analysis. Failure to do so, by

pooling the data regardless of the values of the covariates and using a classification rule that relies on a common threshold value, will result in the test having a discriminatory capacity that is biased compared to its 'true potential' discriminatory capacity. Accordingly, optimistic or pessimistic results may be obtained and, by extension, erroneous conclusions with respect to the real discriminatory capacity of the diagnostic test, which in turn entails an 'incorrect' choice of threshold values to be used in practice.

The previous explanations motivate two possibilities when estimating ROC curves under the presence of covariates. If the discriminatory capacity of the diagnostic test is affected by covariates, then conditional or covariate-specific ROC curves must be considered. When the test diagnostic varies with the covariates, but its discriminatory capacity is not affected by them, then the covariate-adjusted ROC curve, introduced by Janes and Pepe (2009a), is recommended. Both concepts will be defined in the next Section.

3. NOTATION AND DEFINITIONS

Let us assume that along with the continuous diagnostic variables in the diseased population, Y_D, and in the healthy population, $Y_{\bar{D}}$, covariate vectors $\boldsymbol{X}_D$ and $\boldsymbol{X}_{\bar{D}}$ are also available. For the sake of clarity, in this paper we will further assume that the covariates of interest are the same in both healthy and diseased populations. It should be noted, however, that this is not always so. In some circumstances, it could be of interest to evaluate the discriminatory capacity of a diagnostic test with respect to population-specific covariates, as for instance disease stage.

As a natural extension of the ROC curve for continuous diagnostic tests, the conditional or covariate-specific ROC curve, given a covariate value $\boldsymbol{x}$, is defined as

$$\text{(3.1)} \qquad \text{ROC}_{\boldsymbol{x}}(p) = 1 - F_D\big(F_{\bar{D}}^{-1}(1-p \mid \boldsymbol{x}) \mid \boldsymbol{x}\big)\ , \qquad 0 \le p \le 1\ ,$$

where

$$F_D(y \mid \boldsymbol{x}) = P\big(Y_D \le y \mid \boldsymbol{X}_D = \boldsymbol{x}\big)\,,$$
$$F_{\bar{D}}(y \mid \boldsymbol{x}) = P\big(Y_{\bar{D}} \le y \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}\big)\,.$$

Note that in this case, a number of possible different ROC curves can be obtained for each value $\boldsymbol{x}$ in the range of the common part of the supports of $\boldsymbol{X}_D$ and $\boldsymbol{X}_{\bar{D}}$. Associated with the conditional ROC curve, some other measures of discriminatory performance can also be defined. The most widely used one is the area under the ROC curve (AUC), which in the conditional case is defined as $\text{AUC}_{\boldsymbol{x}} = \int_0^1 \text{ROC}_{\boldsymbol{x}}(p)\,\mathrm{d}p$. As for the unconditional case, the $\text{AUC}_{\boldsymbol{x}}$ takes values between 0.5 (for an uninformative test) and 1 (for a perfect test).

Both, the conditional ROC curve and the conditional AUC defined above, depict the discriminatory capacity of a test but for specific values of the covariate vector. It would nevertheless be of great interest to have global discriminatory measures that also take into account covariate information. In this context, the so-called covariate-adjusted ROC curve is defined as an average of conditional ROC curves weighted according to the distribution of the covariate in the diseased population, that is

$$\text{AROC}(p) = \int \text{ROC}_{\boldsymbol{x}}(p)\, \mathrm{d}H_D(\boldsymbol{x}) \ , \tag{3.2}$$

were $H_D(\boldsymbol{x}) = P(\boldsymbol{X}_D \leq \boldsymbol{x})$ is the multivariate distribution function of the vector $\boldsymbol{X}_D$. Despite of the intuitive definition given in the expression above, the covariate-adjusted ROC curve admits other equivalent representations. For instance, in Janes and Pepe (2009a) it is also expressed as

$$\text{AROC}(p) = P\Big(Y_D > F^{-1}_{\bar{D}}(1-p \,|\, \boldsymbol{X}_D)\Big) , \tag{3.3}$$

which means that the covariate-adjusted ROC curve for a FPF $= p$ can be seen as the overall TPF when the threshold used to define test positivity is covariate-specific. This latter expression will be useful when it comes to construct estimators for $\text{AROC}(p)$. Note that based on (3.2), in those situations where the accuracy of a diagnostic test is not affected by covariates, the covariate-adjusted ROC curve coincides with the covariate-specific ROC curve which is common for all covariate values.

4. ESTIMATION PROCEDURES

In order to introduce the estimators, let us assume that we have two independent samples of i.i.d. observations $(\boldsymbol{X}_{\bar{D}1}, Y_{\bar{D}1}), ..., (\boldsymbol{X}_{\bar{D}n_{\bar{D}}}, Y_{\bar{D}n_{\bar{D}}})$ from population $(\boldsymbol{X}_{\bar{D}}, Y_{\bar{D}})$ and $(\boldsymbol{X}_{D1}, Y_{D1})$, ..., $(\boldsymbol{X}_{Dn_D}, Y_{Dn_D})$ from population $(\boldsymbol{X}_D, Y_D)$. Some of the estimators that will be presented below only apply for one-dimensional covariates. However, by a slight abuse of notation, even in those cases we will keep the bold typography to denote the covariates.

4.1. Estimation of the conditional ROC curve

Several proposals for estimating the conditional ROC curve have been given in the statistical literature. Estimators can immediately be obtained by estimating the conditional distribution functions involved in the definition given in (3.1). Besides, other approaches within the general regression framework have been studied, namely the so-called induced and direct ROC-regression methodologies

(see, *e.g.*, Pepe, 1998, 2003; Rodríguez-Álvarez *et al.*, 2011). In this section, we will first present the general ideas behind both approaches, and then focus our attention on nonparametric estimation techniques.

Estimators based on conditional distribution functions. An obvious estimator of the conditional ROC curve follows directly from its definition. Given a covariate value, $\boldsymbol{x}$, the estimator can be constructed as

$$\widehat{\mathrm{ROC}}_{\boldsymbol{x}}(p) \,=\, 1 - \hat{F}_D\big(\hat{F}_{\bar{D}}^{-1}(1-p \mid \boldsymbol{x}) \mid \boldsymbol{x}\big)\,, \tag{4.1}$$

where $\hat{F}_D(\cdot \mid \boldsymbol{x})$ and $\hat{F}_{\bar{D}}(\cdot \mid \boldsymbol{x})$ are estimators of the conditional distributions $F_D(\cdot \mid \boldsymbol{x})$ and $F_{\bar{D}}(\cdot \mid \boldsymbol{x})$, respectively. When we restrict our attention to one-dimensional covariates, the conditional distributions can be estimated nonparametrically, for instance, by kernel-based estimators given in Stone (1977):

$$\hat{F}_{j,h_j}(y \mid \boldsymbol{x}) \,=\, \frac{\sum_{i=1}^{n_j} k\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_{ji}}{h_j}\Big)\, I(Y_{ji} \leq y)}{\sum_{i=1}^{n_j} k\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_{ji}}{h_j}\Big)}\,,$$

with $j \in \{\bar{D}, D\}$, where $I(\cdot)$ denotes the indicator function and where k is the kernel (usually a symmetric density) and h_D and $h_{\bar{D}}$ are the smoothing parameters. Under this approach, the estimator of the conditional ROC curve at a specific covariate value uses the information corresponding to individuals whose covariate values are close to $\boldsymbol{x}$.

The estimator given in (4.1) is of an empirical type, and therefore has discontinuities. In López-de Ullibarri *et al.* (2008) a nonparametric smooth estimator of the conditional ROC curve is obtained by applying the methodology that Peng and Zhou (2004) proposed in the unconditional case. The key idea of this method consists of smoothing the empirical ROC curve by means of kernel techniques. In the conditional case, the smoothed version of (4.1) given in López-de Ullibarri *et al.* (2008) is

$$\widehat{\mathrm{ROC}}_{\boldsymbol{x},h}(p) \,=\, 1 - \int \hat{F}_{D,h_D}\big(\hat{F}_{\bar{D},h_{\bar{D}}}^{-1}(1-p+hu \mid \boldsymbol{x}) \mid \boldsymbol{x}\big)\, k(u)\, \mathrm{d}u\,, \tag{4.2}$$

where the parameter h controls the amount of smoothing and k is a kernel function. The authors propose a bootstrap method to choose the smoothing parameters involved in (4.1) and (4.2).

Very recently, Inácio de Carvalho *et al.* (2013) presented a nonparametric Bayesian model to estimate the conditional distribution functions involved in (3.1). The main advantage of their approach, in contrast to the proposal of López-de Ullibarri *et al.* (2008), is the possibility of studying the effect of multidimensional covariates. Specifically, covariate-dependent Dirichlet processes (DDP) (MacEachern, 2000) defined in terms of i.i.d. Gaussian processes are proposed to

estimate $F_D(\cdot \mid \boldsymbol{x})$ and $F_{\bar{D}}(\cdot \mid \boldsymbol{x})$. Moreover, the computational burden associated with the proposal is overcome by approximating the Gaussian processes by B-splines basis functions, yielding the so-called B-splines DDP mixture model. The authors show by means of simulation the better performance of the proposed model in complex scenarios when compared to other nonparametric estimators of the conditional ROC curve (González-Manteiga *et al.*, 2011; Rodríguez-Álvarez *et al.*, 2011a).

Estimators based on induced-regression methodology. An alternative way to incorporate information from covariates to the ROC analysis is through regression models. The induced methodology in ROC analysis consists of modelling the effect of the covariates through regression models linking the classification variable and the covariates in each population separately. The regression models will then be used to compose the conditional ROC curve. In a general framework, the relationship between the covariate and the classification variable in each population is given by location-scale regression models

$$Y_{\bar{D}} = \mu_{\bar{D}}(\boldsymbol{X}_{\bar{D}}) + \sigma_{\bar{D}}(\boldsymbol{X}_{\bar{D}})\,\varepsilon_{\bar{D}}\ , \tag{4.3}$$

$$Y_D = \mu_D(\boldsymbol{X}_D) + \sigma_D(\boldsymbol{X}_D)\,\varepsilon_D\ , \tag{4.4}$$

where, for $j \in \{\bar{D}, D\}$, $\mu_j(\boldsymbol{x}) = E(Y_j \mid \boldsymbol{X}_j = \boldsymbol{x})$ and $\sigma_j^2(\boldsymbol{x}) = \mathrm{var}(Y_j \mid \boldsymbol{X}_j = \boldsymbol{x})$ are the conditional mean and the conditional variance of Y_j given $\boldsymbol{X}_j = \boldsymbol{x}$, respectively, and the error ε_j is independent of the covariate $\boldsymbol{X}_j$. The independence between the error and the covariate in the location-scale regression model allows us to rewrite the conditional distribution function of the classification variable in terms of the distribution of the regression error as follows:

$$\begin{aligned} F_j(y \mid \boldsymbol{x}) &= P\Big(Y_j \le y \mid \boldsymbol{X}_j = \boldsymbol{x}\Big) \\ &= P\Big(\mu_j(\boldsymbol{X}_j) + \sigma_j(\boldsymbol{X}_j)\,\varepsilon_j \le y \mid \boldsymbol{X}_j = \boldsymbol{x}\Big) \\ &= P\left(\varepsilon_j \le \frac{y - \mu_j(\boldsymbol{x})}{\sigma_j(\boldsymbol{x})}\right) = G_j\left(\frac{y - \mu_j(\boldsymbol{x})}{\sigma_j(\boldsymbol{x})}\right), \end{aligned}$$

where, for $j \in \{\bar{D}, D\}$, $G_j(y) = P(\varepsilon_j \le y)$ is the distribution function of the regression error. An analogous relationship can be established between the conditional quantile function of Y_j given $\boldsymbol{X}_j = \boldsymbol{x}$, $F_j^{-1}(\cdot \mid \boldsymbol{x})$, and the quantile function of ε_j, $G_j^{-1}(\cdot)$, through the expression $F_j^{-1}(p \mid \boldsymbol{x}) = \mu_j(\boldsymbol{x}) + \sigma_j(\boldsymbol{x})\,G_j^{-1}(p)$. Therefore, for a fixed covariate value $\boldsymbol{x}$, and for $0 < p < 1$, the conditional ROC curve can be expressed as

$$\begin{aligned} \mathrm{ROC}_{\boldsymbol{x}}(p) &= 1 - F_D\Big(F_{\bar{D}}^{-1}(1-p \mid \boldsymbol{x}) \mid \boldsymbol{x}\Big) \\ &= 1 - F_D\Big(\mu_{\bar{D}}(\boldsymbol{x}) + \sigma_{\bar{D}}(\boldsymbol{x})\,G_{\bar{D}}^{-1}(1-p) \mid \boldsymbol{x}\Big) \\ &= 1 - G_D\left(\frac{\mu_{\bar{D}}(\boldsymbol{x}) + \sigma_{\bar{D}}(\boldsymbol{x})\,G_{\bar{D}}^{-1}(1-p) - \mu_D(\boldsymbol{x})}{\sigma_D(\boldsymbol{x})}\right) \\ &= 1 - G_D\Big(G_{\bar{D}}^{-1}(1-p)\,b(\boldsymbol{x}) - a(\boldsymbol{x})\Big)\,, \end{aligned} \tag{4.5}$$

where $a(\boldsymbol{x}) = \big(\mu_D(\boldsymbol{x}) - \mu_{\bar{D}}(\boldsymbol{x})\big)/\sigma_D(\boldsymbol{x})$ and $b(\boldsymbol{x}) = \sigma_{\bar{D}}(\boldsymbol{x})/\sigma_D(\boldsymbol{x})$. This formulation allows us to express the conditional ROC curve in terms of the distribution function and quantile function of the regression errors, which are not conditional. Hence, from an estimation point of view, instead of estimating the conditional distribution of Y_D and $Y_{\bar{D}}$ given $\boldsymbol{x}$, one only needs to estimate the error distribution in each population. This is a main advantage with respect to the estimator given in (4.2).

The induced ROC methodology described above has been presented for the most general case. In fact, only particular cases have been addressed in the literature. In a parametric or semiparametric framework, Faraggi (2003) assumes an additive parametric model for the conditional means, with homoscedastic variances and normal errors, in both healthy and diseased populations. Pepe (1998) relaxes the distributional assumptions by not assuming a known probability distribution for the error terms, although the same distribution is considered for both populations. Zhou *et al.* (2002) extend the model in Pepe (1998) by allowing for heteroscedasticity. Finally, Zheng and Heagerty (2004) propose a semiparametric estimator for the conditional ROC curve, in which the distribution of the error terms is unknown and allowed to depend on the covariates, but, as in the previous articles, the effect of the covariates on the conditional means and variances is modelled parametrically. Very recently, Rodríguez and Martínez (2014) presented a Bayesian semiparametric model, where the error terms are assumed to be normally distributed, but nonparametric specifications of the conditional means and variances are allowed.

A different line of research has led to estimation in a fully nonparametric framework, although so far only one-dimensional covariates have been considered. We focus now on those approaches, introduced by González-Manteiga *et al.* (2011) and Rodríguez-Álvarez *et al.* (2011a). When models (4.3) and (4.4) are nonparametric, the estimator of the conditional ROC curve involves the following steps. First, for $j \in \{\bar{D}, D\}$, we need to estimate nonparametrically the location and scale functions in the regression models, say $\hat{\mu}_j(\boldsymbol{x})$ and $\hat{\sigma}_j(\boldsymbol{x})$ by means, for example, of Nadaraya–Watson or local-linear estimators (see, for example, Fan and Gijbels, 1996). Then the distribution of the errors in the two regression models are estimated by the corresponding empirical distribution function of the estimated residuals, *i.e.*, $\hat{G}_j(y) = n_j^{-1} \sum_{i=1}^{n_j} I(\hat{\varepsilon}_{ji} \leq y)$, where, for $j \in \{\bar{D}, D\}$, $\hat{\varepsilon}_{ji} = \big(Y_{ji} - \hat{\mu}_j(\boldsymbol{X}_{ji})\big)/\hat{\sigma}_j(\boldsymbol{X}_{ji})$, $i = 1, ..., n_j$. Finally, given the covariate value $\boldsymbol{x}$, an empirical estimator of the conditional ROC curve is

$$\widehat{\mathrm{ROC}}_{\boldsymbol{x}}(p) \;=\; 1 - \hat{G}_D\Big(\hat{G}_{\bar{D}}^{-1}(1-p)\,\hat{b}(\boldsymbol{x}) - \hat{a}(\boldsymbol{x})\Big)\,, \tag{4.6}$$

where $\hat{a}(\boldsymbol{x}) = \big(\hat{\mu}_D(\boldsymbol{x}) - \hat{\mu}_{\bar{D}}(\boldsymbol{x})\big)/\hat{\sigma}_D(\boldsymbol{x})$ and $\hat{b}(\boldsymbol{x}) = \hat{\sigma}_{\bar{D}}(\boldsymbol{x})/\hat{\sigma}_D(\boldsymbol{x})$. As in the case of (4.1), the previous estimator of the conditional ROC curve is not continuous. In order to obtain a smooth version, González-Manteiga *et al.* (2011) also apply

the methodology in Peng and Zhou (2004), which yields

$$\widehat{\mathrm{ROC}}_{\boldsymbol{x},h}(p) \;=\; 1 - \int \hat{G}_D\Big(\hat{G}_{\bar{D}}^{-1}(1-p+hu)\,\hat{b}(\boldsymbol{x}) - \hat{a}(\boldsymbol{x})\Big)\,k(u)\,\mathrm{d}u\;. \tag{4.7}$$

The authors show that the former estimator also admits the following explicit expression:

$$\widehat{\mathrm{ROC}}_{\boldsymbol{x},h}(p) \;=\; \frac{1}{n_D}\sum_{i=1}^{n_D} K\left(\frac{\hat{G}_{\bar{D}}\Big(\big\{\hat{\varepsilon}_{Di} + \hat{a}(\boldsymbol{x})\big\}/\hat{b}(\boldsymbol{x})\Big) - 1 + p}{h}\right),$$

where K is the distribution function corresponding to the density kernel k.

A detailed study of the asymptotic properties of the estimators given in (4.6) and (4.7) is provided in González-Manteiga *et al.* (2011). In Rodríguez-Álvarez *et al.* (2011a), a bootstrap-based test to check for the effect of the covariate over the conditional ROC curve is proposed. Although both papers focus on the estimation of the conditional ROC curve, an estimator of the conditional AUC is also presented, $\widehat{\mathrm{AUC}}_{\boldsymbol{x}} = \int_0^1 \widehat{\mathrm{ROC}}_{\boldsymbol{x}}(p)\,\mathrm{d}p$, with the integral being approximated by numerical integration methods. In that sense, the paper by Yao *et al.* (2010) goes one step further in proposing a nonparametric estimator for $\mathrm{AUC}_{\boldsymbol{x}}$ based also on induced modelling and local linear kernel smoothers. The authors exploit the relation between the Mann–Whitney statistic and the empirical estimator of the unconditional AUC (see, *e.g.*, Bamber, 1975) and propose a covariate-specific Mann–Whitney estimator for $\mathrm{AUC}_{\boldsymbol{x}}$.

Estimators based on direct-regression methodology. In contrast to the induced methodology, in the direct methodology the effect of the covariates is directly evaluated on the ROC curve. To motivate the standard formulation of direct methodology, let us re-express the conditional ROC curve as follows:

$$\begin{aligned}
\mathrm{ROC}_{\boldsymbol{x}}(p) \;&=\; 1 - F_D\Big(F_{\bar{D}}^{-1}(1-p \mid \boldsymbol{x}) \mid \boldsymbol{x}\Big) \\
&=\; 1 - P\Big(Y_D \le F_{\bar{D}}^{-1}(1-p \mid \boldsymbol{x}) \mid \boldsymbol{X}_D = \boldsymbol{x}\Big) \\
&=\; 1 - P\Big(F_{\bar{D}}(Y_D \mid \boldsymbol{x}) \le 1-p \mid \boldsymbol{X}_D = \boldsymbol{x}\Big) \\
&=\; P\Big(1 - F_{\bar{D}}(Y_D \mid \boldsymbol{x}) < p \mid \boldsymbol{X}_D = \boldsymbol{x}\Big) \qquad (4.8) \\
&=\; E\Big[I\Big(1 - F_{\bar{D}}(Y_D \mid \boldsymbol{x}) < p\Big) \mid \boldsymbol{X}_D = \boldsymbol{x}\Big]\,. \qquad (4.9)
\end{aligned}$$

As can be observed, the conditional ROC curve may be seen as: (a) the conditional distribution function of the random variable $1 - F_{\bar{D}}(Y_D \mid \boldsymbol{x})$ in expression (4.8), or (b) the conditional expected value of the binary variable $I\big(1 - F_{\bar{D}}(Y_D \mid \boldsymbol{x}) < p\big)$ in expression (4.9). The random variable $1 - F_{\bar{D}}(Y_D \mid \boldsymbol{x})$ is called 'placement value' in related literature (see, for example, Hanley and Hajian-Tilaki, 1997) and represents the standardization of the classification variable in the diseased population to the conditional distribution of the non-diseased population.

These two interpretations justify to express the conditional ROC curve as a sort of regression model of the form

$$\text{ROC}_{\boldsymbol{x}}(p) = g\big(\mu(\boldsymbol{x}), \gamma(p)\big)\,, \tag{4.10}$$

where g is a bivariate function on $[0,1]$ and γ is a function defined on the interval $[0,1]$. The function μ collects the effect of the covariates on the conditional ROC curve, and γ is a baseline function related to the shape of the ROC curve. In order to obtain a valid model of ROC curves, some restrictions need to be imposed on the elements of model (4.10). In particular, the function g needs to be monotone increasing in p, with $g(\mu(\boldsymbol{x}), \gamma(0)) = 0$ and $g(\mu(\boldsymbol{x}), \gamma(1)) = 1$ for all $\boldsymbol{x}$. As in the case of the induced methodology presented above, model (4.10) represents the most general formulation of the direct methodology. In fact, only the additive specification

$$\text{ROC}_{\boldsymbol{x}}(p) = g\big(\mu(\boldsymbol{x}) + \gamma(p)\big) \tag{4.11}$$

has been addressed in the statistical literature. Different proposals have been suggested, which differ in the assumptions made about the functions g, μ and γ. In Pepe (1997, 2000) and Alonzo and Pepe (2002), g is assumed to be known, the effect of the covariates on the conditional ROC curve is assumed to be linear, *i.e.*, $\mu(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$, and the baseline function γ is assumed to have a parametric form. Cai and Pepe (2002) and Cai (2004) leave γ completely unspecified, but the function μ is linear as well. In general, models such as (4.11) with parametric specifications for μ define the so-called class of ROC-GLMs due to its similarity with a generalized linear model (GLM, McCullagh and Nelder, 1989) in regression (Pepe, 2003). In all the aforementioned papers, the function g is assumed to be known. Huazhen *et al.* (2012) relax this assumption, by allowing a completely unknown function g. As for the approaches in Cai and Pepe (2002) and Cai (2004), the function γ remains unspecified and μ is assumed to have a parametric form. In a completely nonparametric framework, Rodríguez-Álvarez *et al.* (2011b) extend the class of ROC-GLM regression models, by assuming a generalized additive model (GAM, Hastie and Tibshirani, 1990) for the ROC curve, that is

$$\mu(\boldsymbol{x}) = \mu(x_1, ..., x_d) = \alpha + \sum_{k=1}^{d} f_k(x_k)\,,$$

where $f_1, ..., f_d$ are unknown nonparametric functions, and γ also remains unspecified.

Either if the specifications in (4.11) involve a GLM structure (as in Alonzo and Pepe, 2002) or a GAM structure (as in Rodríguez-Álvarez *et al.*, 2011b), the estimation process is similar and can be described as given in the following steps. First, choose a set of FPFs $0 \leq p_l \leq 1$, $l = 1, ..., n_P$, where the conditional ROC curves will be evaluated. Second, estimate $F_{\bar{D}}(\cdot \,|\, \boldsymbol{x})$, say $\hat{F}_{\bar{D}}(\cdot \,|\, \boldsymbol{x})$, on the basis of the sample $(\boldsymbol{X}_{\bar{D}i}, Y_{\bar{D}i})$, $i = 1, ..., n_{\bar{D}}$. Third, for each observation in the diseased population, calculate the estimated placement value $1 - \hat{F}_{\bar{D}}(Y_{Di} \,|\, \boldsymbol{x})$, $1 \leq i \leq n_D$.

Fourth, calculate the binary indicators $I\big(1-\hat{F}_{\bar{D}}(Y_{Di}\,|\,\boldsymbol{x})\leq p_l\big)$, for $1\leq i\leq n_D$ and $1\leq l\leq n_P$. And finally, fifth, fit the model $g\big(\mu(\boldsymbol{x})+\gamma(p)\big)$ as a regression model with the indicators $I\big(1-\hat{F}_{\bar{D}}(Y_{Di}\,|\,\boldsymbol{x})\leq p_l\big)$ as response and covariates $\boldsymbol{X}_{Di}$ and p_l, $i=1,...,n_D$, $l=1...,n_P$.

Depending on the chosen specifications for μ and γ, GLM or GAM techniques will be employed for fitting the model (4.11). For instance, in Rodríguez-Álvarez *et al.* (2011b) the proposed estimation procedure is based on a combination of local scoring and backfitting algorithms (Hastie and Tibshirani, 1990), and the nonparametric functions $f_1,...,f_d$ and γ are estimated using local linear kernel smoothers (see Fan and Gijbels, 1996). Note that in contrast to the nonparametric approaches based on induced modelling presented above, this proposal allows for the possibility of incorporating multidimensional covariates. However, the study of the theoretical properties of the estimator is so far lacking in the literature.

Throughout the above outline of induced and direct modelling, the covariates (whose effect on the ROC curve we seek to evaluate) were assumed to be common to both the healthy and the diseased population. As mentioned before, in practice this is not necessarily so. For instance, it may be of interest to evaluate the performance of the diagnostic variable with respect to disease stage. Induced methodology poses no problem when it comes to incorporating specific covariates of healthy or diseased populations, or both. On the other hand, direct methodology—as presented here—accepts no specific covariates of the healthy population. Yet, even in cases where this may seem a restriction, the need arises in few situations in practice.

4.2. Estimation of the covariate-adjusted ROC curve

As explained in the introduction, in some practical cases, although the diagnostic test varies along with the covariates, its discriminatory capacity may remain unalterable. In such a situation, instead of considering the conditional ROC curve, the covariate-adjusted ROC curve is more convenient. The definition given in (3.3)

$$\text{AROC}(p) \;=\; P\Big(Y_D > F_{\bar{D}}^{-1}\big(1-p\mid \boldsymbol{X}_D\big)\Big)$$

suggests estimating the covariate-adjusted ROC curve as sample proportion of individuals in the diseased population that exceed a certain covariate-specific threshold calculated with the conditional quantile function in the healthy population. Note that the conditional quantile function is an unknown function and therefore needs to be estimated. Janes and Pepe (2009a) propose estimators of the form

$$\widehat{\text{AROC}}(p) \;=\; \frac{1}{n_D}\sum_{i=1}^{n_D} I\Big(Y_{Di} > \hat{F}_{\bar{D}}^{-1}\big(1-p\mid \boldsymbol{X}_{Di}\big)\Big)\,,$$

where $\hat{F}_{\bar{D}}^{-1}(1-p \mid \boldsymbol{X}_{Di})$ can be estimated semiparametrically or nonparametrically. In the context of the induced methodology described in Subsection 4.1, Rodríguez-Álvarez *et al.* (2011a) used the relation between the conditional quantile and the quantile of the regression errors to obtain the following nonparametric estimator:

$$\widehat{\mathrm{AROC}}(p) = \frac{1}{n_D} \sum_{i=1}^{n_D} I\left(\frac{Y_{Di} - \hat{\mu}_{\bar{D}}(\boldsymbol{X}_{Di})}{\hat{\sigma}_{\bar{D}}(\boldsymbol{X}_{Di})} > \hat{G}_{\bar{D}}^{-1}(1-p) \right),$$

where $\hat{\mu}_{\bar{D}}$ and $\hat{\sigma}_{\bar{D}}$ are nonparametric estimators of $\mu_{\bar{D}}$ and $\sigma_{\bar{D}}$ in model (4.3), and $\hat{G}_{\bar{D}}^{-1}$ is the empirical quantile function of the estimated residuals. The theoretical properties of this estimator have not been studied yet.

5. ILLUSTRATION WITH REAL DATA

In this section, a real data illustration of the importance of including covariates into the ROC framework is presented. The data set comes from a cross-sectional study carried out by the Galician Endocrinology and Nutrition Foundation (FENGA), consisting of 2860 individuals representative of the adult population of Galicia (northwest of Spain). A detailed description of this data set can be found in Tomé *et al.* (2008). For confidentiality reasons, only a subsample of the global sample was used in this paper, where we aimed at assessing the performance of the body mass index (BMI) for predicting clusters of cardiovascular disease (CVD) risk factors. Accordingly, diseased subjects were defined as those having two or more CVD risk factors (raised triglycerides, reduced high-density lipoprotein cholesterol, raised blood pressure and raised fasting plasma glucose), following the International Diabetes Federation criteria (International Diabetes Federation, 2006). For the study here presented, a total of 1419 individuals were selected from the original data set, with an age range between 18 and 85 years. From those, 46.4% are men (449 healthy and 209 diseased) and the remaining 53.6% are women (625 healthy and 136 diseased). An in-depth study of the global data set is presented in Rodríguez-Álvarez *et al.* (2011a,b).

It is well known that anthropometric measures behave differently according to both age and gender. This can be observed in Table 1, where some summary statistics of the BMI for men and women, as well as for different age strata, are presented. As illustrated in Section 2, it is therefore advisable to incorporate both covariates into the ROC analysis. In this paper, we applied the nonparametric induced approach proposed by González-Manteiga *et al.* (2011) and Rodríguez-Álvarez *et al.* (2011a) and presented in Section 4.1. Since this proposal only admits one continuous covariate, separate analyses were conducted on men and women respectively.

Table 1: Median and interquartile range of the BMI for the global sample, for men and women, and for different age strata.

	1^{st} **Quartile**	**Median**	3^{rd} **Quartile**
Global sample	22.84	25.91	29.34
Gender			
Female	22.00	24.69	25.91
Male	24.16	26.88	27.14
Age strata			
<30 years	21.28	22.85	25.83
30–39 years	22.66	25.40	28.08
40–49 years	24.18	26.77	29.74
50–59 years	25.84	28.65	31.46
≥ 60 years	26.62	29.38	31.72

In addition to the estimated conditional ROC curves, other summary measures of accuracy, the conditional AUC and the age-adjusted ROC curve, were also obtained. In Figure 3, the estimated age-adjusted ROC curve for both men and women is shown, jointly with the estimated pooled ROC curve. As can be observed, in both cases the age-adjusted ROC curve lies below the pooled ROC curve, especially for men. It is worth remembering that the covariate-adjusted ROC curve is an average of conditional ROC curves, and can therefore be interpreted as a covariate-adjusted global discriminatory measure. Thus, for the endocrinology data, pooling the data regardless of age and gender would lead to an optimistic conclusion about the discriminatory capacity of the BMI when predicting the presence of CVD risk factors.

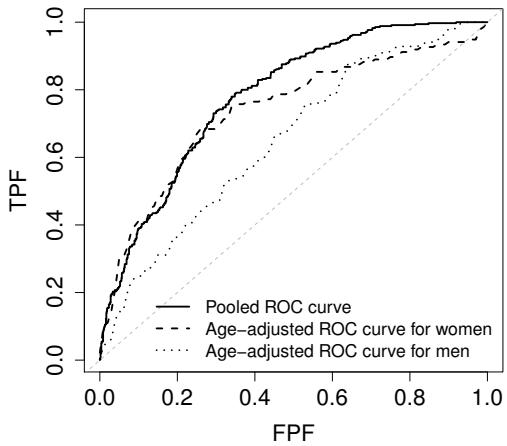


Figure 3: Estimated pooled ROC curve for the endocrinology data (solid line). The dashed and dotted lines represent the estimated age-adjusted ROC curve for women and men, respectively.

In Figure 4 the estimated conditional ROC curve and AUC for different age values are depicted, for both men and women. Note that whereas for men the accuracy of the BMI is more or less constant along age, for women, age displays a relevant effect on the discriminatory capacity of this anthropometric measure. This graphical conclusion was confirmed by applying the bootstrap-based test presented in Rodríguez-Álvarez *et al.* (2011a). The test enabled a significant age effect to be detected in the case of women. In the case of men, however, there was no evidence to suggest such an effect.

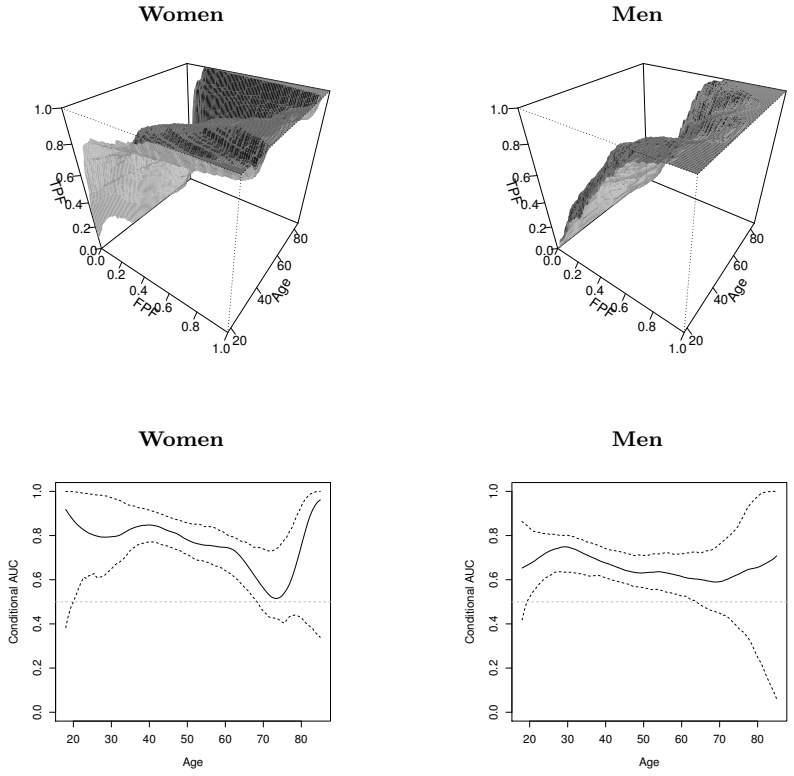


Figure 4: Estimated conditional ROC curves and AUCs for the endocrinology data for women and men. The dashed lines represent the 95 per cent pointwise bootstrap confidence interval.

The results presented in this section emphasize once again the importance and consequences of including the information provided by covariates when evaluating the discriminatory capacity of a diagnostic test. In the case of women, the conditional ROC curve should be reported since it has been proved that age has an effect on the accuracy of the BMI. For men, however, no age effect was detected. Nevertheless, even in this case, reporting the discriminatory capacity of the pooled data would lead to an optimistic conclusion, and therefore the age-adjusted ROC curve should be provided.

6. DISCUSSION

In this paper we explained why it is important to incorporate covariates in the ROC analysis and which effect it has on the curve. We also presented two different ways to take covariates into account, either by working with a conditional ROC curve or with a so-called covariate-adjusted ROC curve. Several estimation procedures were outlined for both approaches. Interested readers can find more details in the provided references.

Although we focused in this review on the estimation of the ROC curve in the presence of covariates, it is clear that apart from the ROC curve itself, interest also lies in summary statistics of the ROC curve, like *e.g.* the AUC, the Youden index and other related indices. Within a parametric or semiparametric framework, some attempts about this topic can be found in Faraggi (2003), in which the induced methodology is employed, and in Dodd and Pepe (2003a,b) and Cai and Dodd (2008), all based on the ROC-regression direct modelling approach.

An interesting extension of the ROC methodology is the extension to functional data. We mention the paper by Inácio *et al.* (2012), who consider the extension to functional covariates. To this end, semiparametric and nonparametric induced ROC-regression estimators are proposed and studied. Also, the extension of the ROC methodology from completely observed data to censored data is a promising field of research. For an overview article on this topic we refer to Pepe *et al.* (2008).

Another interesting point to note is that almost no theory has been done for the nonparametric estimators of the conditional and adjusted ROC curve, except in González-Manteiga *et al.* (2011), who obtain the asymptotic normality of nonparametric estimators of both the conditional ROC curve and the conditional AUC based on induced methodology. Their results are limited to a one-dimensional covariate, but they can be easily extended to multi-dimensional covariates by using Neumeyer and Van Keilegom (2010) in the proofs of the asymptotic results.

A number of issues remain unexplored in the context of ROC curves with covariates. For instance, a lot of work remains to be done to extend the concept of relative distributions to the inclusion of covariates (see Handcock and Morris, 1999, for a textbook on this topic). ROC curves are very much related to relative distributions or relative densities (see *e.g.* Li *et al.*, 1996), but their objective is different. In fact, the ROC curve in a point $0 < p < 1$ equals one minus the relative distribution evaluated in $1 - p$. Since the relative density of one population versus another population equals the uniform density in case both populations have the same distribution, it is clear that deviations from the uniform density give an indication of the way in which the two distributions differ from each other.

Hence, relative densities are more used in the context of comparing the distribution of two populations, whereas ROC curves are used for assessing the discriminatory capacity of a diagnostic test. As far as we are aware of, no formal and detailed study of the concept of relative distribution or relative density in the presence of covariates has been developed so far.

ACKNOWLEDGMENTS

The authors would like to thank the Galician Endocrinology and Nutrition Foundation (*Fundación de Endocrinoloxía e Nutrición Galega—FENGA*) for having supplied the database used in this study. J. C. Pardo-Fernández and M. X. Rodríguez-Álvarez express their gratitude for the support received in the form of the Spanish Ministry of Science and Innovation grants MTM2011-23204 (FEDER support included) and MTM2011-28285-C02-01. I. Van Keilegom acknowledges financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, from IAP research network P7/06 of the Belgian Government (Belgian Science Policy) and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

REFERENCES

Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques, *Biostatistics*, **3**, 421–432.

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

Cai, T. (2004). Semi-parametric ROC regression analysis with placement values, *Biostatistics*, **5**, 45–60.

Cai, T. and Dodd, L. E. (2008). Regression analysis for the partial area under the ROC curve, *Statistica Sinica*, **18**, 817–836.

Cai, T. and Pepe, M. S. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease, *Journal of the American Statistical Association*, **97**, 1099–1107.

Dodd, L. E. and Pepe, M. S. (2003a). Semiparametric regression for the area under the receiver operating characteristic curve, *Journal of the American Statistical Association*, **98**, 409–417.

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

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman & Hall/CRC, Boca Raton.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates, *The Statistician*, **52**, 179–192.

González-Manteiga, W.; Pardo-Fernández, J. C. and Van Keilegom, I. (2011). ROC curves in nonparametric location-scale regression models, *Scandinavian Journal of Statistics*, **38**, 169–184.

Handcock, M. S. and Morris, M. (1999). *Relative Distribution Methods in the Social Sciences*, Springer-Verlag, New York.

Hanley, J. A. and Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update, *Academic Radiology*, **4**, 49–58.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall/CRC, Boca Raton.

Huazhen, L.; Xiao-Hua, Z. and Gang, L. (2012). A direct semiparametric receiver operating characteristic curve regression with unknown link and baseline functions, *Statistica Sinica*, **22**, 1427–1456.

Inácio, V.; González-Manteiga, W.; Febrero-Bande, M.; Gude, F.; Alonzo, T. A. and Cadarso-Suárez, C. (2012). Extending induced ROC methodology to the functional context, *Biostatistics*, **13**, 594–608.

Inácio de Carvalho, V.; Jara, A.; Hanson, T. E. and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling, *Bayesian Analysis*, **3**, 623–646.

International Diabetes Federation (2006). The IDF consensus worldwide definition of the metabolic syndrome. Accessed January 2014. (`http://www.idf.org/webdata/docs/IDF_Meta_def_final.pdf`).

Janes, H. and Pepe, M. S. (2008). Adjusting for covariate in studies of diagnostic, screening, or prognosis markers: an old concept in a new setting, *American Journal of Epidemiology*, **168**, 89–97.

Janes, H. and Pepe, M. S. (2009a). Adjusting for covariate effects on classification accuracy using the covariate-adjusted ROC curve, *Biometrika*, **96**, 371–382.

Janes, H. and Pepe, M. S. (2009b). Accommodating covariates in ROC analysis, *Stata Journal*, **9**, 17–39.

Li, G.; Tiwari, R. C. and Wells, M. T. (1996). Quantile comparison functions in two sample problems with applications to comparisons of diagnostic markers, *Journal of the American Statistical Association*, **91**, 689–698.

López-de Ullibarri, I.; Cao, R.; Cadarso-Suárez, C. and Lado, M. J. (2008). Nonparametric estimation of conditional ROC curves: Application to discrimination tasks in computerized detection of early breast cancer, *Computational Statistics & Data Analysis*, **52**, 2623–2631.

MacEachern, S. N. (2000). *Dependent Dirichlet processes*, Technical report, Department of Statistics, The Ohio State University.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition, Chapman & Hall/CRC, Boca Raton.

Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, **101**, 1067–1078.

Peng, L. and Zhou, X.-H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves, *Journal of Statistical Planning and Inference*, **118**, 129–143.

Pepe, M. S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing, *Biometrika*, **84**, 595–608.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results, *Biometrics*, **54**, 124–135.

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures, *Biometrics*, **56**, 352–359.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.

Pepe, M. S; Zheng, Y.; Jin, Y.; Huang, Y.; Parikh, C. R. and Levy, W. C. (2008). Evaluating the ROC performance of markers for future events, *Lifetime Data Analysis*, **14**, 86–113.

Rodríguez, A. and Martínez, J. C. (2014). Bayesian semiparametric estimation of covariate-dependent ROC curves, *Biostatistics*, **15**, 353–369.

Rodríguez-Álvarez, M. X.; Roca-Pardiñas, J. and Cadarso-Suárez, C. (2011a). ROC curve and covariates: extending induced methodology to the non-parametric framework, *Statistics & Computing*, **21**, 483–499.

Rodríguez-Álvarez, M. X.; Roca-Pardiñas, J. and Cadarso-Suárez, C. (2011b). A new flexible direct ROC regression model: Application to the detection of cardiovascular risk factors by anthropometric measures, *Computational Statistics & Data Analysis*, **55**, 3257–3270.

Rodríguez-Álvarez, M. X.; Tahoces, P. G.; Cadarso-Suárez, C. and Lado, M. J. (2011). Comparative study of ROC regression techniques—Applications for the computer-aided diagnostic system in breast cancer detection, *Computational Statistics & Data Analysis*, **55**, 888–902.

Stone, C. J. (1977). Consistent nonparametric regression, *The Annals of Statistics*, **5**, 595–620.

Tomé, M. A; Botana, M. A.; Cadarso-Suárez, C.; Rego-Iraeta, A; Fernández-Mariño, A; Mato, J. A.; Solache, I. and Pérez-Fernández, R. (2008). Prevalence of metabolic syndrome in Galicia (NW Spain) on four alternative definitions and association with insulin resistance, *Journal of Endocrinological Investigation*, **32**, 505–511.

Yao, F.; Craiu, R. V. and Reiser, B. (2010). Nonparametric covariate adjustment for receiver operating characteristic curves, *The Canadian Journal of Statistics*, **38**, 27–46.

Zheng, Y. and Heagerty, P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data, *Biostatistics* **4**, 615–632.

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

REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 43–65

DEVELOPMENTS IN ROC SURFACE ANALYSIS AND ASSESSMENT OF DIAGNOSTIC MARKERS IN THREE-CLASS CLASSIFICATION PROBLEMS

Author: Christos T. Nakas
– Laboratory of Biometry, University of Thessaly,
School of Agricultural Sciences, Phytokou street, 38446 Volos, Greece
cnakas@uth.gr

Abstract:

- This article reviews current state of the art of ROC surface analysis and illustrates its use through an application on a pancreatic cancer diagnostic marker. Receiver Operating Characteristic (ROC) surfaces have been studied in the literature essentially only during the last decade and are considered as a natural generalization of ROC curves in three-class diagnostic problems. This article presents the definition, construction, modelling, and utility of the ROC surface while trying to provide an extensive reference list in the topic. It describes methodology for inference based on the Volume Under the ROC surface (VUS) and methodology for decision making through the selection of optimal cut-off points using the notion of the generalized Youden index as the optimality criterion of choice. It ends with a discussion regarding future directions for research in this field of knowledge.

Key-Words:

- *generalized Youden index; receiver operating characteristic (ROC) surface; true class fraction (TCF); volume under the ROC surface (VUS).*

AMS Subject Classification:

- 62C99, 62P10.

1. INTRODUCTION

Receiver Operating Characteristic (ROC) curve analysis has been an active area of research since the early 1950s. The ROC curve depicts the quality of a diagnostic marker in a two-class classification problem. It illustrates the trade-off between sensitivity and specificity as the cut-off point for decision making varies through possible values of the diagnostic marker. Put more formally, suppose that, in a two-class classification problem, a diagnostic marker results in measurements $X_1 \sim F_1$ from the first class under study and $X_2 \sim F_2$ from the second class under study. Suppose that, in general, values from X_2 are larger than values from X_1 but X_1 and X_2 are not perfectly separated, *i.e.* there is an amount of overlap between measurements from the two-classes.[1] A cut-off point c is selected for decision making which will result in the fractions of specificity, defined as $\text{spec}(c) = P(X_1 \leq c)$, and sensitivity, defined as $\text{sens}(c) = P(X_2 > c)$. The fractions of sensitivity (or else True Positive Fraction, TPF) and specificity (True Negative Fraction, TNF) vary as the cut-off point c varies. The ROC curve is defined as the graph depicting $(1 - P(X_1 \leq c), P(X_2 > c)) = (1 - \text{spec}(c), \text{sens}(c))$ in the unit square $[0, 1] \times [0, 1]$, as c varies. Equivalently, the ROC curve is the graph of the function $\text{ROC}(t) = 1 - F_2(F_1^{-1}(1 - t))$, where $t \in [0, 1]$. The Area Under the ROC Curve (AUC) is equivalent to $P(X_1 < X_2)$ and it is the most widely used index for the quantification of the performance of a diagnostic marker in the two-class setting. A useful diagnostic marker will result in an ROC curve with AUC close to 1. A diagnostic marker with AUC close to 0.5 will, in general, be considered as uninformative. The AUC takes on values in $[0.5, 1]$ if the condition that measurements from X_1 are in general smaller than those from X_2 actually holds. The non-parametric estimate of the AUC is equivalent to the Wilcoxon–Mann–Whitney statistic (Pepe, 2003). Formal assessment of the quality of a diagnostic marker based on the AUC consists of testing the null hypothesis, $H_0 : \text{AUC} = 0.5$ versus the alternative of interest through the statistic $z = \{(\text{AUC} - 0.5)/\text{se}(AUC)\} \sim N(0, 1)$, where se(AUC) is the standard error of AUC, estimated, *e.g.*, using the bootstrap. If H_0 is rejected, the diagnostic marker under study is considered to be useful and a cut-off point c must be chosen for decision-making purposes. Use of the maximum of the Youden index (J) is a widely adopted approach for cut-off point selection. The Youden index is defined as $J = \max_c\{\text{sens}(c) + \text{spec}(c) - 1\} = \max_c\{F_1(c) - F_2(c)\}$, as a result, the value of c that maximizes J is chosen. ROC curve analysis is presented in detail in a number of well-written books, such as Pepe (2003) and Zhou *et al.* (2011).

Notions of ROC curve analysis have been extended to accommodate problems of three-class and multiple-class classification. The ROC surface has been proposed as a natural generalization of the ROC curve for the assessment of diagnostic markers in three-class classification problems. The ROC surface was

[1] However, we do not impose any type of stochastic ordering by $X_1 < X_2$.

introduced by Scurfield (1996). The Volume Under the ROC Surface (VUS) was proposed as an index for the assessment of the diagnostic accuracy of the marker under consideration. Unfortunately, the latter article received very little attention probably because it only described the theoretical construction of the ROC surface and did not provide any related application. A similar construction was proposed independently, a few years later though, by Mossman (1999) which was implemented in Mathematica by Heckerling (2001). Inference regarding the VUS, based on Mossman's construction, using non-parametric statistics, was studied by Dreiseitl *et al.* (2000). The ROC surface construction, and the generalization of this construction in multiple-class classification problems, in a non-parametric context, was proposed in Nakas and Yiannoutsos (2004). Interestingly, the latter construction unifies the approaches of Mossman and Scurfield in a natural way and thus offered the framework and the theoretical basis for extending ROC curve analysis concepts in multiple-class classification problems. This construction has been reinvented at least a couple of times later on (*e.g.* Xiong *et al.*, 2006; Li and Fine, 2008), however, in Xiong *et al.* (2006) the parametric framework is studied extensively supplementing the work in Nakas and Yiannoutsos (2004). Given the theoretical basis for the ROC surface, several articles appeared in the literature during the last 10 years generalizing notions from ROC curve analysis. ROC surfaces are overviewed in the textbook on ROC analysis by Krzanowski and Hand (2009).

In the following sections the ROC surface analysis literature will be reviewed and unified, and an illustration offering insight on the use of ROC surfaces will be described. The Discussion in Section 5 will constitute an effort to provide guidance for future research to the interested reader.

2. ROC SURFACE ANALYSIS

2.1. Description of the problem

To define formally the general three-class classification problem, suppose that n_1 measurements from Class 1, denoted by X_1, follow a distribution with cumulative distribution function F_1 (*i.e.* $X_1 \sim F_1$), and similarly for n_2 measurements from Class 2, $X_2 \sim F_2$, and for n_3 measurements from Class 3, $X_3 \sim F_3$. A decision rule that classifies subjects in one of these classes can be defined using two ordered threshold points $c_1 < c_2$. Specifically, suppose that the ordering of interest is $X_1 < X_2 < X_3$. The researcher's goal is the assessment of the quality of a diagnostic marker in classifying correctly subjects from the three ordered classes.

2.2. Definition

The construction of the ROC surface is based on the following algorithm: Decide for Class 1 when a measurement is less than c_1, for Class 2 when it is between c_1 and c_2, for Class 3 otherwise. This decision rule will result in three True Class Fractions (TCFs) and six False Class Fractions (FCFs). Then, $\mathrm{TCF}_1 = P(X_1 \leq c_1)$, $\mathrm{TCF}_2 = P(c_1 < X_2 \leq c_2)$, and $\mathrm{TCF}_3 = P(X_3 > c_2)$. Also, $\mathrm{FCF}_{12} = P(c_1 \leq X_1 \leq c_2)$ and the remaining five possible FCF_{ij}, $i, j = 1, 2, 3$, $i \neq j$ are defined accordingly. Varying c_1, c_2 in the union of the supports of F_1, F_2, and F_3, $(\mathrm{TCF}_1, \mathrm{TCF}_2, \mathrm{TCF}_3)$ can be plotted in a three-dimensional coordinate system to produce the ROC surface in the unit cube. The True Class Fractions take values in $[0, 1]$ with corner coordinates $\{(1, 0, 0),\ (0, 1, 0),\ (0, 0, 1)\}$. Thus, the ROC surface is the 3-dimensional plot in the unit cube depicting $(F_1(c_1), F_2(c_2) - F_2(c_1), 1 - F_3(c_2))$, for all cut-off points (c_1, c_2), with $c_1 < c_2$. The functional form of the ROC surface is $\mathrm{ROC}_s(\mathrm{TCF}_1, \mathrm{TCF}_3) = F_2(F_3^{-1}(1 - \mathrm{TCF}_3)) - F_2(F_1^{-1}(\mathrm{TCF}_1))$ (Nakas and Yiannoutsos, 2004). It can be seen that this is a generalization of the ROC curve in three dimensions since projecting the ROC surface to the plane defined by TCF_2 versus TCF_1, *i.e.* setting $\mathrm{TCF}_3 = 0$, the ROC curve between Classes 1 and 2 is produced, *i.e.* $\mathrm{ROC}(\mathrm{TCF}_1) = 1 - F_2(F_1^{-1}(\mathrm{TCF}_1))$. The latter is the equivalent construction of the ROC curve depicting $(\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1))$ instead of $(\mathrm{FCF}_{12}(c_1), \mathrm{TCF}_2(c_1))$. Similarly, the projection of the ROC surface to the plane defined by the axes TCF_2, TCF_3, yields the ROC curve between Classes 2 and 3, *i.e.* $\mathrm{ROC}(\mathrm{TCF}_3) = F_2(F_3^{-1}(1 - \mathrm{TCF}_3))$, the latter being the functional form of TCF_2 versus TCF_3 analogous to specificity versus sensitivity rather than the other way around. For reasons of brevity, a pictorial representation will be provided in Section 4.

2.3. The Volume Under the ROC Surface (VUS)

The Volume Under the ROC Surface (VUS) is equal to $P(X_1 < X_2 < X_3)$. An unbiased non-parametric estimator of VUS is given by

$$\widehat{\mathrm{VUS}} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I(X_{1i}, X_{2j}, X_{3k}) \ ,$$

where $I(X_1, X_2, X_3)$ equals one if X_1, X_2, X_3 are in the correct order and zero otherwise (Dreiseitl *et al.*, 2000). The definition of $I(X_1, X_2, X_3)$ can be adapted to adjust for the presence of ties. Specifically, when ties are present, define: $I(X_1, X_2, X_3) = 1/2$ if $X_1 = X_2 < X_3$ or if $X_1 < X_2 = X_3$ and $I(X_1, X_2, X_3) = 1/6$ if $X_1 = X_2 = X_3$, and $I(X_1, X_2, X_3) = 0$ (or 1 if perfectly ordered) otherwise.

The expected value of VUS will be then

$$P(X_1 < X_2 < X_3) + \frac{1}{2}\,P(X_1 < X_2 = X_3) \\ + \frac{1}{2}\,P(X_1 = X_2 < X_3) + \frac{1}{6}\,P(X_1 = X_2 = X_3)\ .$$

The VUS takes the value $1/3! = 1/6$ when the three distributions completely overlap and the value one when the three classes are perfectly discriminated in the correct order. Parametric approaches for the estimation of VUS have been discussed in Xiong *et al.* (2006). Kang and Tian (2013) offer an extensive study comparing possible parametric and non-parametric approaches for the estimation of VUS in terms of bias and root mean square error.

In several situations in practice researchers may wish to limit the study of the ROC surface to a clinically relevant range of measurement values. In such cases the partial VUS has been defined in Xiong *et al.* (2006). The partial VUS generalizes the notion of the partial AUC in the two-class problem (see *e.g.* Zhou *et al.*, 2011).

2.4. ROC surface modelling

Restate the functional form of the ROC surface, by writing $\mathrm{TCF}_1 = p_1$ and $\mathrm{TCF}_3 = p_3$, as follows:

(2.1)

$$\mathrm{ROC_s}(p_1, p_3) = \begin{cases} F_2\big(F_3^{-1}(1-p_3)\big) - F_2\big(F_1^{-1}(p_1)\big)\,, & \text{if } \ F_1^{-1}(p_1) \le F_3^{-1}(1-p_3)\ , \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Then, VUS is defined as

$$\mathrm{VUS} \ = \int_0^1 \int_0^{1-F_3(F_1^{-1}(p_1))} \mathrm{ROC_s}(p_1, p_3)\ \mathrm{d}p_3\,\mathrm{d}p_1\ .$$

2.4.1. Empirical and non-parametric estimation

The empirical estimator of the ROC surface can be obtained by replacing the distribution functions in the definition of the ROC surface with their empirical counterparts. The empirical, non-parametric estimator of the ROC surface is

$$\widehat{\mathrm{ROC_s}}(p_1, p_3) = \begin{cases} \widehat{F}_2\big(\widehat{F}_3^{-1}(1-p_3)\big) - \widehat{F}_2\big(\widehat{F}_1^{-1}(p_1)\big)\,, & \text{if } \ \widehat{F}_1^{-1}(p_1) \le \widehat{F}_3^{-1}(1-p_3)\ , \\ 0\,, & \text{otherwise}\,, \end{cases}$$

where $\widehat{F}_1, \widehat{F}_2$ and $\widehat{F}_3$ are the empirical distribution functions for the measurements from the three classes.

Most recently, kernel approaches for the estimation of the ROC surface have been studied (Kang and Tian, 2013). Specifically, F_1, F_2, and F_3, can be modeled through Gaussian kernel estimators of the form $F_i(t) = 1/n_i \Sigma_{j=1}^{n_i} \Phi\{(t - X_{ij})/h_i\}$, for $i = 1, 2, 3$. For the bandwidth h_i, which controls the amount of smoothing, Kang and Tian (2013) have considered $h_i = \{4/(3n_i)\}^{1/5} \min(\mathrm{SD}_i, \mathrm{IQR}_i/1.349)$; here, SD_i and IQR_i are the standard deviation and interquartile range, respectively, for the X_i measurements.

Bayesian non-parametric estimation of the ROC surface based on Finite Polya Tree (FPT) prior distributions for the three-classes was studied by Inácio *et al.* (2011). The model is specified hierarchically and involves the specification of independent FPT prior distributions for F_i, for $i = 1, 2, 3$, conditional on a set of hyperparameters, *i.e.*

$$F_i \mid c_i, \theta_i \sim \mathrm{FPT}_{J_i}(F_{\theta_i}, c_i) \ , \qquad i = 1, 2, 3 \ .$$

Suppose, that the F_i are centered at $F_{\boldsymbol{\theta}_i} = N(\mu_i, \sigma_i)$, where $\boldsymbol{\theta}_i = (\mu_i, \sigma_i)$. The mixing parameters μ_i have independent normal prior distributions $N(a_{\mu_i}, b_{\mu_i})$, whereas σ_i have independent gamma prior distributions $\Gamma(a_{\sigma_i}, b_{\sigma_i})$. Hyperparameters are considered fixed. The levels of the finite Polya trees are set equal to J_i, and are used to determine the level of detail that is accommodated by the model; mathematical subtleties on the model can be found in Inácio *et al.* (2011).

2.4.2. Parametric estimation

Under the assumption of normality for F_1, F_2, and F_3 (*i.e.* $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, $X_3 \sim N(\mu_3, \sigma_3^2)$), Xiong *et al.* (2006) used the model in (2.1) to describe the general framework of the ROC surface and the VUS. The parametric form of the ROC surface is

$$\begin{aligned} \mathrm{ROC_s}(p_1, p_3) \ = \ & \Big\{ \Phi\big(\beta_1 + \beta_2 \Phi^{-1}(1 - p_3)\big) - \Phi\big(\beta_3 + \beta_4 \Phi^{-1}(p_1)\big) \Big\} \\ & \times \ \mathbb{1}_{\{\beta_3 + \beta_4 \Phi^{-1}(p_1) \leq \beta_1 + \beta_2 \Phi^{-1}(1 - p_3)\}}(p_1, p_3) \ , \end{aligned}$$

where $\mathbb{1}$ denotes the indicator function, Φ is the distribution function of the standard normal, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^{\mathrm{T}}$ specifies the parameters of the ROC surface. If the normality assumption is valid, the components of $\boldsymbol{\beta}$ may be expressed as functions of the means and variances of the three normal distributions which model F_1, F_2, and F_3, as follows:

$$\beta_1 = \frac{\mu_3 - \mu_2}{\sigma_2} \, , \qquad \beta_2 = \frac{\sigma_3}{\sigma_2} \, , \qquad \beta_3 = \frac{\mu_1 - \mu_2}{\sigma_2} \, , \qquad \beta_4 = \frac{\sigma_1}{\sigma_2} \ .$$

Kang and Tian (2013) have considered the use of the Box–Cox transformation for non-normally distributed data prior to the use of the parametric normal model and have compared with the kernel approach they proposed in terms of the bias and accuracy of the estimation of the VUS (see §2.4.1).

Under the Bayesian parametric paradigm, in order to find estimates for the beta parameters, a Markov Chain Monte Carlo approach is needed. A Metropolis–Hastings algorithm or a Gibbs sampler can be employed. The use of the Metropolis–Hastings algorithm with uninformative normal priors for the means and uninformative gamma prior distributions for the standard deviations is recommended in Inácio *et al.* (2011). However, studies focusing on Bayesian parametric approaches for the ROC surface have not appeared in the literature yet.

2.4.3. Semi-parametric estimation

Semi-parametric estimation of the ROC surface was studied by Li and Zhou (2009) generalizing the results of the two-class case in Hsieh and Turnbull (1996) and by Nze Ossima *et al.* (2013), generalizing the results of the two-class case in Gönen and Heller (2010). The estimation of the ROC surface of a diagnostic marker with continuous measurements given covariate information has been considered in Li *et al.* (2012). Specifically, suppose that the measurements of the diagnostic marker under study can be modeled through the following general regression model for a set of p covariates, $\boldsymbol{Z} = (Z_1, ..., Z_p)^{\mathrm{T}}$,

$$g(X_i) = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_i + \sigma_i\,\varepsilon\ , \qquad i = 1, 2, 3\ , \tag{2.2}$$

where g is a strictly monotone increasing function, $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{ip})^{\mathrm{T}}$ are the regression coefficients for Class i, σ_i is a class-specific scale parameter, and ε is the error following a common distribution function G with support $(-\infty, \infty)$ for all three classes. Then, the construction of the ROC surface is based on the rule: Decide for Class 1 when the diagnostic marker's measurement estimate from (2.2) is less than c_1, for Class 2 when it is between c_1 and c_2, for Class 3 otherwise.

2.5. Inference based on the VUS

Formal assessment of the diagnostic accuracy of a marker in a three-class classification problem via its VUS can be based on testing the null hypothesis $H_0 : \mathrm{VUS} = 1/6$ versus the alternative of interest. The test statistic is

$$Z_1 = \frac{\widehat{\mathrm{VUS}} - 1/6}{\sqrt{\mathrm{var}(\widehat{\mathrm{VUS}})}} \sim N(0, 1)\ . \tag{2.3}$$

The $\widehat{\text{VUS}}$ is the non-parametric estimate of VUS. Then, Z_1 is normally distributed based on results from U-statistics theory (Pepe, 2003). Variance of $\widehat{\text{VUS}}$ can be estimated by using U-statistics methodology or the bootstrap (Nakas and Yiannoutsos, 2004). The bootstrap approach consists of sampling with replacement n_1, n_2, n_3 subjects from the initial samples from X_1, X_2, X_3 respectively, and calculating the VUS for each of the b replications of this procedure. The bootstrap estimate of the variance of VUS is the sample variance of the b bootstrap VUSs (Nakas and Yiannoutsos, 2004). Properties of non-parametric estimators of the variance of $\widehat{\text{VUS}}$ have been studied by Guangming *et al.* (2013). Based on Z_1, 95% confidence intervals for VUS can be constructed in a straightforward fashion. Wan (2012) proposed an empirical likelihood confidence interval for the non-parametric estimate of VUS.

The parametric approach for confidence interval construction for VUS is studied in Xiong *et al.* (2006). Confidence intervals are constructed based on the Delta method, otherwise the bootstrap can be used, where for each bootstrap replication the parametric VUS is calculated. Non-parametric predictive inference for the ROC surface and the VUS is developed in Coolen-Maturi *et al.* (2013).

Regarding the comparison of VUSs, consider the case where two markers (A and B) are measured on the same $n = n_1 + n_2 + n_3$ specimens which are classified by a gold standard procedure into three ordered disease classes. Let $(X_1^{\text{A}}, X_2^{\text{A}}, X_3^{\text{A}})$ and $(X_1^{\text{B}}, X_2^{\text{B}}, X_3^{\text{B}})$ be the values for markers A and B, respectively.

To compare VUS^{A} and VUS^{B} via their non-parametric, empirical estimates, Dreiseitl *et al.* (2000) proposed a U-statistics approach. Specifically, the null hypothesis $H_0 : \text{VUS}^{\text{A}} = \text{VUS}^{\text{B}}$ is tested by calculating

$$Z_2 = \frac{\widehat{\text{VUS}}^{\text{A}} - \widehat{\text{VUS}}^{\text{B}}}{\sqrt{\text{var}\big(\widehat{\text{VUS}}^{\text{A}}\big) + \text{var}\big(\widehat{\text{VUS}}^{\text{B}}\big) - 2\,\text{cov}\big(\widehat{\text{VUS}}^{\text{A}}, \widehat{\text{VUS}}^{\text{B}}\big)}},$$

and then comparing this value to a standard normal distribution. The variance and covariance of $\widehat{\text{VUS}}$ can be estimated using the estimators provided in Dreiseitl *et al.* (2000). Alternatively, the bootstrap can be used to test H_0 as in Nakas and Yiannoutsos (2004). Xiong *et al.* (2007) have studied the parametric analogue for the comparison of VUSs based on the results in Xiong *et al.* (2006), while Tian *et al.* (2011) consider the parametric approach using notions of generalized pivots. Inference for specific TCFs is studied in Dong *et al.* (2011, 2013).

2.6. The ROC umbrella

The notion of the ROC surface has been generalized to accommodate cases with umbrella or tree orderings (*i.e.* $X_1 < X_3 > X_2$ or $X_2 > X_1 < X_3$, respec-

tively) between the three classes under study by Nakas and Alonzo (2007). The ROC surface and VUS reviewed in the previous sections are not applicable when such orderings are of interest. Specifically, these approaches do not allow one to assess the ability of a marker to differentiate two disease classes from a third disease class without requiring a specific monotone order for the three disease classes under study. The derivation of an ROC surface for the ordering $X_2 > X_1 < X_3$ is reviewed here, however, the derivation is analogous for the other ordering.

Using the fact that $(X_2 > X_1 < X_3) = (X_1 < X_2 < X_3) \cup (X_1 < X_3 < X_2)$, or equivalently $P(X_2 > X_1 < X_3) = P(X_1 < X_2 < X_3) + P(X_1 < X_3 < X_2)$, the construction of two ROC surfaces (say A and B) corresponding to the orderings $X_1 < X_2 < X_3$ and $X_1 < X_3 < X_2$, respectively, is possible. These are the plots of the points: $\left(\mathrm{TCF}_1^{\mathrm{A}}(c_1, c_2), \mathrm{TCF}_2^{\mathrm{A}}(c_1, c_2), \mathrm{TCF}_3^{\mathrm{A}}(c_1, c_2)\right)$ and $\left(\mathrm{TCF}_1^{\mathrm{B}}(c_1, c_2), \mathrm{TCF}_2^{\mathrm{B}}(c_1, c_2), \mathrm{TCF}_3^{\mathrm{B}}(c_1, c_2)\right)$, respectively, with $(c_1, c_2) \in \mathbb{R}^2$ and $c_1 < c_2$.

The umbrella ordering can be viewed however on a single graph in the unit cube by plotting on the same axes defined by $x = \mathrm{TCF}_1^{\mathrm{A}}$, $y = \mathrm{TCF}_2^{\mathrm{A}}$, $z = \mathrm{TCF}_3^{\mathrm{A}}$, in turn

$$\left(\mathrm{TCF}_1^{\mathrm{A}}(c_1, c_2),\ \mathrm{TCF}_2^{\mathrm{A}}(c_1, c_2),\ \mathrm{TCF}_3^{\mathrm{A}}(c_1, c_2)\right)$$

and

$$\left(1 - \mathrm{TCF}_1^{\mathrm{B}}(c_1, c_2),\ 1 - \mathrm{TCF}_2^{\mathrm{B}}(c_1, c_2),\ 1 - \mathrm{TCF}_3^{\mathrm{B}}(c_1, c_2)\right),$$

with $(c_1, c_2) \in \mathbb{R}^2$ and $c_1 < c_2$. It can be shown that surfaces A, B thus constructed on a single graph, are disjoint.

The resulting umbrella ROC graph is a diagnostic plot for the visual assessment of the degree of separation in the given ordering of the three populations based on the samples. The volume under surface A plus the volume over surface B can be used for inference. We refer to this summary measure as the umbrella volume (UV). UV is equivalently the sum of the volumes under the ROC surfaces A and B corresponding to the monotone orderings $X_1 < X_2 < X_3$ and $X_1 < X_3 < X_2$, respectively. The umbrella ROC graph contains both ordered ROC surfaces.

The non-parametric unbiased estimator of the volume of the umbrella ROC graph $P(X_2 > X_1 < X_3)$ is:

$$\widehat{UV} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I_U(X_{1i}, X_{2j}, X_{3k}) \ ,$$

where $I_U(X_1, X_2, X_3)$ equals one if $X_2 > X_1 < X_3$ and zero otherwise; the UV varies from zero to one and is equal to $P(X_1 < X_2 < X_3) + P(X_1 < X_3 < X_2) = 1/6 + 1/6 = 1/3$ when the three distributions completely overlap and equals one when the three classes are perfectly discriminated in the given ordering.

In practice, ties may occur between measurements in the three disease classes, in which case $I_U(X_1, X_2, X_3)$=1 if $X_1 < X_2 = X_3$, $I_U(X_1, X_2, X_3)$=1/2 if $X_1 = X_2 < X_3$ or if $X_1 = X_3 < X_2$, and $I_U(X_1, X_2, X_3)$=1/6 if $X_1 = X_2 = X_3$. The expected value of UV will then be

$$P(X_1 < X_2 < X_3) + P(X_1 < X_3 < X_2) + P(X_1 < X_2 = X_3) + \frac{1}{2} P(X_1 = X_2 < X_3) + \frac{1}{2} P(X_1 = X_3 < X_2) + \frac{1}{6} P(X_1 = X_2 = X_3) \ .$$

Comparison of umbrella ROC volumes in a non-parametric framework has been studied in Alonzo and Nakas (2007), while the umbrella ROC has not been studied in the parametric framework yet. Alonzo *et al.* (2009) provide a comparison of tests for restricted orderings in the three-class case, illustrating the usefulness of ROC surfaces and ROC umbrellas in different applied contexts.

2.7. The ROC manifold

For the k-class problem, with $k > 3$, based on a single diagnostic marker, an ROC manifold may be constructed as described in Nakas and Yiannoutsos (2004). Using $k-1$ ordered decision thresholds c_j, $j = 1, ..., k-1$, with $c_1 < \cdots < c_{k-1}$, define a decision rule as in the three-class case given above. Then k TCFs are defined in a k-dimensional space. The ROC manifold is produced by varying the $k-1$ ordered decision thresholds. The Hypervolume Under the ROC Manifold (HUM) is

$$\text{HUM} \ = \ P\Big\{(X_1 < X_2) \cap \cdots \cap (X_{k-1} < X_k)\Big\} .$$

The HUM will vary from $1/k!$ to 1, taking the value $1/k!$ for a completely uninformative marker and the value 1 when the k populations are perfectly separated.

A non-parametric unbiased estimate of HUM is

$$\widehat{\text{HUM}} \ = \ \frac{1}{n_1 \cdots n_k} \sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} I(X_{1i_1}, ..., X_{ki_k}) \ ,$$

where the n_i, for $i = 1, ..., k$, are the sample sizes from the k populations and the function $I(X_1, ..., X_k)$ is defined in analogy to the three-class case. The ROC manifold and HUM have not been studied in a parametric framework yet. Theoretical extensions relating to the general k-class problem are studied in Davidov and Herman (2012).

2.8. Other topics in three- and k-class ROC methodology

Computational aspects regarding the calculation of the VUS or HUM when computational complexity is an issue have also appeared in the literature (Waegeman *et al.*, 2008a,b; Clémençon *et al.*, 2013). Alternative approaches for the generalization of the ROC curve in three- and multiple-class classification problems have been proposed by Yang and Carlin (2000), Hand and Till (2001), Wan and Zhang (2009) and Yang and Zhao (2010). These approaches, however, address specific research questions in the sense that they do not offer a complete theoretical framework for the generalization of ROC curve analysis and will not be studied further in this review. Generalizations of ROC analysis notions when the gold standard is continuous-scale rather than categorical has been studied by Obuchowski (2006) and by Shiu and Gatsonis (2012).

In the two-class case, considerable amount of research has been conducted to address issues where no gold standard is available for the characterization of the true status of the subjects in the study, or when the gold standard information is available for a fraction of the subjects in the study, *i.e.*, in the presence of verification bias (see *e.g.* Pepe, 2003; Zhou *et al.*, 2011). Only a few papers have appeared that introduce these notions in ROC surface analysis (Chi and Zhou, 2009; Wang *et al.*, 2011; Kang *et al.*, 2013b). Bantis *et al.* (2013) have used a cubic spline smoothing approach to model the ROC surface when measurements are subject to a limit of detection.

Theoretical properties of the ROC surface and ROC manifold that span beyond the scopes of the current article have been studied in Scurfield *et al.* (1998), He and Frey (2006), He *et al.* (2006), Everson and Fieldsend (2006), Edwards and Metz (2007), Sahiner *et al.* (2008), He and Frey (2008), He and Frey (2009), He *et al.* (2010), Schubert *et al.* (2011), Edwards and Metz (2012), and Edwards (2013).

3. THE GENERALIZED YOUDEN INDEX

3.1. Definition

A three-class Youden index has been recently proposed for the assessment of accuracy and cut-off point selection in the three-class setting (Nakas *et al.*, 2010, 2013). Specifically, define:

$$\begin{aligned} J_3 &= \max_{c_1,c_2}\Big\{\mathrm{TCF}_1 + \mathrm{TCF}_2 + \mathrm{TCF}_3 - 1\Big\} \\ &= \max_{c_1,c_2}\Big\{F_1(c_1) + F_2(c_2) - F_2(c_1) - F_3(c_2)\Big\}. \end{aligned} \tag{3.1}$$

This is a constrained optimization problem with $c_1 < c_2$. This latter condition will always be true if a usual stochastic order of the form $P(X_1 > x) \leq P(X_2 > x) \leq P(X_3 > x)$ holds. The pair of cut-off points c_1, c_2 that corresponds to J_3 is considered optimal and can be used in practice for decision making in the three-class case. As in the two-class setting, weights can be added to the definition of J_3 to reflect the relative importance of the three TCFs.

3.2. Properties

The generalized Youden index lends itself to a natural unification of the two- and three-class analysis approaches. Denote by $J_{3;(1,2,3)}$ the J_3 index corresponding to the ordering $X_1 < X_2 < X_3$ and by $J_{2;(i,j)}$, the ordinary Youden index corresponding to the ordering $X_i < X_j$, for $i, j = 1, 2, 3$. Then, by the definitions of J_2 and J_3 above, it follows that

$$\begin{aligned} J_{3;(1,2,3)} &= \max_{c_1,c_2}\Big\{F_1(c_1) - F_2(c_1) + F_2(c_2) - F_3(c_2)\Big\} \\ &= \max_{c_1}\Big\{F_1(c_1) - F_2(c_1)\Big\} + \max_{c_2}\Big\{F_2(c_2) - F_3(c_2)\Big\} \\ &= J_{2;(1,2)} + J_{2;(2,3)} \ . \end{aligned}$$

Thus, J_3 is the sum of the Youden index for the two-class analysis of classes 1 and 2 and the Youden index for the two-class analysis of classes 2 and 3. This result holds if weights are introduced in the definition of J_3 since λ can be set to one and $\nu^* = \nu/\lambda$, $\mu^* = \mu/\lambda$ can be used instead of ν, μ in the definition of J_3^+. Then, $J^+_{3;(1,2,3)} = \max_{c_1,c_2}\{\nu^* \cdot \mathrm{TCF}_1 + \mu^* \cdot \mathrm{TCF}_2 + \mathrm{TCF}_3 - 1\} = J^+_{2;(1,2)} + J^+_{2;(2,3)}$. This result also holds whenever the ordering $X_1 < X_2 < X_3$ is true, thus $c_1 < c_2$. A counterexample, where the ordering is not true and, as a result, the property does not hold, can easily be constructed. As a rule of thumb, pairwise AUCs for adjacent classes can reveal the correct order, which in turn can be used for the three-class analysis. From the property above it follows that J_3 takes on values in $[0, 2]$. To define J_3 in $[0, 1]$, Luo and Xiong (2013) proposed using $J_3/2$.

3.3. Estimation

Note that J_3 can be estimated non-parametrically by using the empirical distribution functions in the definition in (3.1), *i.e.* $\widehat{J}_3 = \max_{c_1,c_2}\{\widehat{F}_1(c_1) + \widehat{F}_2(c_2) - \widehat{F}_2(c_1) - \widehat{F}_3(c_2)\}$, or parametrically based on distributional assumptions for the data. Empirical non-parametric estimation of the generalized Youden index has been considered in Nakas *et al.* (2010, 2013), while parametric estimation based on normality assumptions has been described in Luo and Xiong (2012, 2013). Luo and Xiong created an R-package (`DiagTest3Grp`) for the estimation of the VUS, generalized Youden index and respective optimal cut-off

points under the parametric normal model, of which further details can be found in Luo and Xiong (2012). Estimation and use of the generalized Youden index for non-parametric predictive inference is studied in Coolen-Maturi *et al.* (2013).

3.4. Other measures of discrimination ability

The generalized Youden index can serve as an index of the discrimination ability of a diagnostic marker for the purpose of selecting the cut-off points that may be used for decision making, while the VUS is the measure of choice for the evaluation of the discrimination ability of the marker under study *per se*. The reason for the selective use of different measures of discrimination ability is the interpretation of the measure itself. Other measures for the evaluation of the discrimination ability of a marker rising from the definition of the ROC surface has also been proposed in the literature (*e.g.* Van Calster *et al.*, 2012a,b) but have not received much attention from the research community. Use of a general cost function for the selection of cut-off points in multiple-class diagnostic testing has been studied in Skaltsa *et al.* (2012).

4. ILLUSTRATION OF ROC SURFACE ANALYSIS

CA19-9 is a standard pancreatic cancer diagnostic marker. Measurements on 40 pancreatic cancer patients, 23 pancreatitis patients, and 40 healthy controls were available. The dataset that is used here for illustrative purposes is part of the dataset in Leichtle *et al.* (2013). Evaluation of CA19-9 in terms of its diagnostic ability to differentiate between the three classes in the order

$$\texttt{Controls} < \texttt{Pancreatitis} < \texttt{Cancer}$$

is illustrated. Descriptive statistics are given in Table 1, while respective boxplots are depicted in Figure 1.

Table 1: Descriptive statistics for CA 19-9 marker measurements for the three classes under study.

	`Controls`	`Pancreatits`	`Cancer`
mean	6.94	22.50	200.46
sd	4.74	30.88	237.85
median	6.60	8.51	111.60
min	0.6	2.5	0.6
max	20.67	121.80	971.50
N	40	23	40

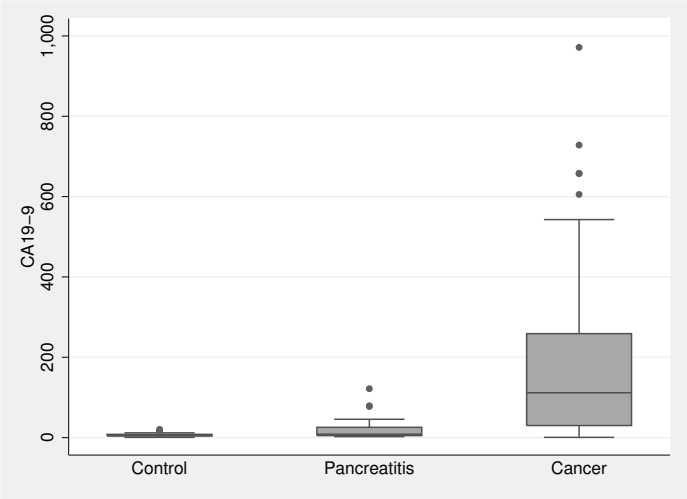


Figure 1: Boxplots of CA 19-9 marker measurements for the three classes under study.

Frequentist and Bayesian non-parametric ROC surfaces are depicted in Figure 2. The empirical non-parametric VUS is equal to 0.528 (95% CI: $0.403, 0.654$; $p < 0.001$), while the VUS based on the Bayesian non-parametric approach is equal to 0.550 (95% CI: $0.455, 0.652$; $p < 0.001$). The generalized Youden index J_3 is 0.929, resulting in $c_1 = 8.40$ and $c_2 = 25.60$. The cut-off point c_1 corresponds to the diagnosis between pancreatitis patients and healthy controls, while c_2 discriminates between pancreatic cancer patients and pancreatitis patients.

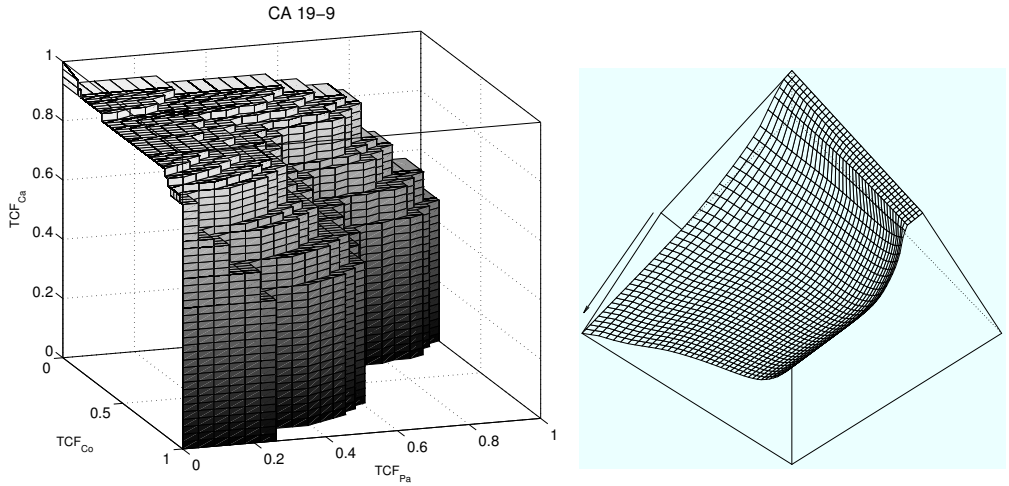


Figure 2: Non-parametric ROC surface for the CA 19-9 data (left panel) and Bayesian non-parametric model from a different viewpoint (right panel).

The corresponding TCF for healthy controls is equal to 80.00%. Regarding pancreatitis patients TCF is just 30.40%, while for pancreatic cancer patients TCF is 82.50%. Compare with the parametric approach that the `DiagTest3Grp` R-package employs: VUS $= 0.519$ (95% CI: $0.385, 0.653$; $p < 0.001$), $J_3 = 1.22 = 0.61 \times 2$, with $c_1 = 16.17$ and $c_2 = 86.62$, corresponding to the TCF triplet (0.974, $0.562, 0.684$) respectively. Unfortunately, the aforementioned R-package does not offer a graph for the ROC surface. However, the Shapiro–Wilk test rejects the normality assumption for all three groups in the study (with $p < 0.001$). Non-parametric approaches are thus considered as more reliable in our example.

Data analysis was conducted using `R version 3.0.0` (R Foundation for Statistical Computing, `http://www.R-project.org`), `Matlab R2013a` (MathWorks Inc., Natick, MA), and `Stata 11.2` (StataCorp LP, College Station, TX).

5. DISCUSSION

ROC surface analysis is a valuable tool for three-class classification problems as it generalizes ROC curve analysis in a natural way within the ROC framework. The utility of ROC surface analysis is demonstrated by the numerous applications that have already appeared in diverse scientific fields (*e.g.* Yu, 2012; Ratnasamy *et al.*, 2008; Yiannoutsos *et al.*, 2008; Abraham *et al.*, 2009; Wandishin and Mullen, 2009; Dalrymple-Alford *et al.*, 2010; Tremont *et al.*, 2011; Dunngalvin *et al.*, 2011; Bruña *et al.*, 2012; Cianferoni *et al.*, 2012; Coleman *et al.*, 2013; Migliaretti *et al.*, 2013; Leichtle *et al.*, 2013).

Until now, researchers have mainly dealt with geometric properties of the ROC surface itself and with generalizations of theoretical findings from the two-class case. Many issues remain to be resolved. Multiple-class classification within the ROC framework and the notion of the ROC umbrella have only scantly been dealt with. Based on the probabilistic properties of the VUS and UV, the claim that the ROC surface and VUS can also be used for three-class analysis when the classes are nominal instead of ordinal (*e.g.* Li and Fine, 2008) seems to be flawed. As a result, theoretical developments for the robustification of the framework of ROC surface analysis are still needed. Other topics of future research include further generalizations from the two-class case. Specifically, issues of future research include time-varying ROC surfaces and generalized linear modelling approaches for the ROC surface along the lines presented in Pepe (2003). The study of predictive values in the three-class and multiple-class case is also of interest. An initial attempt is presented in Yiannoutsos *et al.* (2008). Reclassification issues have just started attracting the interest of researchers in the field. Li *et al.* (2013) have extended the notions in Pencina *et al.* (2012) regarding the net reclassification improvement and integrated discrimination improvement for the k-class

case. Pepe and Thompson (2000) have studied the issue of combination of diagnostic markers in the two-class case via maximizing the area under the ROC curve and have compared this approach with the combination of the diagnostic markers measurements using logistic regression and linear discriminant analysis. Zhang and Li (2011) and Kang *et al.* (2013a) have generalized these results for ROC surface analysis by considering the combinations that maximize the VUS. Currently there is ongoing research on this topic regarding different approaches for VUS maximization using combinations of diagnostic markers.

The generalized Youden index is a simple, useful loss-function for the selection of the optimal cut-off points that can be used for decision-making based on a diagnostic marker of interest in the three-class case. Modelling approaches summarized here and in Kang and Tian (2013), could be employed to develop further practices for the choice of cut-off points after the construction of the ROC surface. Non-parametric predictive inference methods also offer a valuable framework for decision-making in three-class ROC analysis (Coolen *et al.*, 2013; Coolen-Maturi *et al.*, 2013).

R-packages for the implementation of ROC surface analysis tools are of great importance. Researchers interested in using ROC surface methodology should be able to use the Comprehensive R Archive Network repository for their research needs.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Alexander Leichtle for providing the CA19-9 data and Dr. Vanda Inácio de Carvalho for providing the R-code for the implementation of the Bayesian non-parametric approach.

REFERENCES

Abraham, A. G.; Duncan, D. D.; Gange, S. J. and West, S. (2009). Computer-aided assessment of diagnostic images for epidemiological research, *BMC Medical Research Methodology*, **9**, art. no. 74.

Alonzo, T. A. and Nakas, C. T. (2007). Comparison of ROC umbrella volumes with an application to the assessment of lung cancer diagnostic markers, *Biometrical Journal*, **49**, 654–664.

Alonzo, T. A.; Nakas, C. T.; Yiannoutsos, C. T. and Bucher, S. (2009). A comparison of tests for restricted orderings in the three-class case, *Statistics in Medicine*, **28**, 1144–1158.

Bantis, L. E.; Tsimikas, J. V. and Georgiou, S. D. (2013). Smooth ROC curves and surfaces for markers subject to a limit of detection using monotone natural cubic splines, *Biometrical Journal*, **55**, 719–740.

Bruña, R.; Poza, J.; Gòmez, C.; García, M.; Fernández, A. and Hornero, R. (2012). Analysis of spontaneous MEG activity in mild cognitive impairment and Alzheimer's disease using spectral entropies and statistical complexity measures, *Journal of Neural Engineering*, **9**, art. no. 036007.

Chi, Y. Y. and Zhou, X.-H. (2009). Receiver operating characteristic surfaces in the presence of verification bias, *Journal of the Royal Statistical Society. Ser.* C, **57**, 1–23.

Cianferoni, A.; Garrett, J. P.; Naimi, D. R.; Khullar, K. and Spergel, J. M. (2012). Predictive values for food challenge-induced severe reactions: Development of a simple food challenge score, *Israel Medical Association Journal*, **14**, 24–28.

Clémençon, S.; Robbiano, S. and Vayatis, N. (2013). Ranking data with ordinal labels: Optimality and pairwise aggregation, *Machine Learning*, **91**, 67–104.

Coleman, D. J.; Silverman, R. H.; Rondeau, M. J.; Lloyd, H. O.; Khanifar, A. A. and Chan, R. V. P. (2013). Age-related macular degeneration: Choroidal ischaemia?, *British Journal of Opthalmology*, **97**, 1020–1023.

Coolen, F. P. A.; Coolen-Schrijner, P.; Coolen-Maturi, T. and Elkhafifi, F. F. (2013). Nonparametric Predictive Inference for Ordinal Data, *Communications in Statistics—Theory and Methods*, **42**, 3478–3496.

Coolen-Maturi, T.; Elkhafifi, F. F. and Coolen, F. P. A. (2013). Nonparametric predictive inference for three-group ROC analysis, *Technical Report No 1307; Department of Mathematical Sciences, University of Durham, UK*, `http://www.npi-statistics.com/NPI-3ROC-report-1307.pdf`.

Davidov, O. and Herman, A. (2012). Ordinal dominance curve based inference for stochastically ordered distributions, *Journal of the Royal Statistical Society, Ser.* B:, **74**, 825–847.

Dalrymple-Alford, J.C.; MacAskill, M.R.; Nakas, C.T.; Livingston, L.; Graham, C.; Crucian, G. P.; Melzer, T. R.; Kirwan, J.; Keenan, R.; Wells, S.; Porter, R. J.; Watts, R. and Anderson, T. J. (2010). The MoCA: Well suited screen for cognitive impairment in Parkinson disease, *Neurology*, **75**, 1717–1725.

Dong, T.; Tian, L.; Hutson, A. and Xiong, C. (2011). Parametric and non-parametric confidence intervals of the probability of identifying early disease stage given sensitivity to full disease and specificity with three ordinal diagnostic groups, *Statistics in Medicine*, **30**, 3532–3545.

Dong, T.; Kang, L.; Hutson, A.; Xiong, C. and Tian, L. (2013). Confidence interval estimation of the difference between two sensitivities to the early disease stage, *Biometrical Journal*, DOI: 10.1002/bimj.201200012.

Dreiseitl, S.; Ohno-Machado, L. and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis, *Medical Decision Making*, **20**, 323–331.

Dunngalvin, A.; Daly, D.; Cullinane, C.; Stenke, E.; Keeton, D.; Erlewyn-Lajeunesse, M.; Roberts, G. C.; Lucas, J. and Hourihane, J. O. (2011). Highly accurate prediction of food challenge outcome using routinely available clinical data, *Journal of Allergy and Clinical Immunology*, **127**, 633–639.

Edwards, D. C. (2013). Validation of monte carlo estimates of three-class ideal observer operating points for normal data, *Academic Radiology*, **20**, 908–914.

Edwards, D. C. and Metz, C. E. (2007). Optimization of restricted ROC surfaces in three-class classification tasks, *IEEE Transactions on Medical Imaging*, **26**, 1345–1356.

Edwards, D. C. and Metz, C. E. (2012). The three-class ideal observer for univariate normal data: Decision variable and ROC surface properties, *Journal of Mathematical Psychology*, **56**, 256–273.

Everson, R. M. and Fieldsend, J. E. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective, *Pattern Recognition Letters*, **27**, 918–927.

Gönen, M. and Heller, G. (2010). Lehmann family of ROC curves, *Medical Decision Making*, **30**, 509–517.

Guangming, P.; Xiping, W. and Wang, Z. (2013). Nonparametric statistical inference for $P(X < Y < Z)$, *Sankhya* A, **75**, 118–138.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple-class classification problems, *Machine Learning*, **45**, 171–186.

He, X. and Frey, E. C. (2006). Three-class ROC analysis—the equal error utility assumption and the optimality of three-class ROC surface using the ideal observer, *IEEE Transactions on Medical Imaging*, **25**, 979–986.

He, X. and Frey, E. C. (2008). The meaning and use of the volume under a three-class ROC surface (VUS), *IEEE Transactions on Medical Imaging*, **27**, 577–588.

He, X. and Frey, E. C. (2009). The validity of three-class Hotelling trace (3-HT) in describing three-class task performance: Comparison of three-class volume under ROC surface (VUS) and 3-HT, *IEEE Transactions on Medical Imaging*, **28**, 185–193.

He, X.; Gallas, B. D. and Frey, E. C. (2010). Three-class ROC analysis: Toward a general decision theoretic solution, *IEEE Transactions on Medical Imaging*, **29**, 206–215.

He, X.; Metz, C.E.; Tsui, B. M. W.; Links, J. M. and Frey, E. C. (2006). Three-class ROC analysis - A decision theoretic approach under the ideal observer framework, *IEEE Transactions on Medical Imaging*, **25**, 571–581.

Heckerling, P. S. (2001). Parametric three-way Receiver Operating Characteristic surface analysis using Mathematica, *Medical Decision Making*, **21**, 409–417.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *Annals of Statistics*, **24**, 25–40.

Inácio, V.; Turkman, A. A.; Nakas, C. T. and Alonzo, T. A. (2011). Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface, *Biometrical Journal*, **53**, 1011–1024.

Kang, L. and Tian, L. (2013). Estimation of the volume under the ROC surface with three ordinal diagnostic categories, *Computational Statistics and Data Analysis*, **62**, 39–51.

Kang, L.; Xiong, C.; Crane, P. and Tian, L. (2013a). Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories, *Statistics in Medicine*, **32**, 631–643.

Kang, L.; Xiong, C. and Tian, L. (2013b). Estimating confidence intervals for the difference in diagnostic accuracy with three ordinal diagnostic categories without a gold standard, *Computational Statistics and Data Analysis*, **68**, 326–338.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*, Chapman & Hall/CRC, Boca Raton.

Leichtle, A. B.; Ceglarek, U.; Weinert, P.; Nakas, C. T.; Nuoffer, J. M.; Kase, J.; Conrad, T.; Witzigmann, H.; Thiery, J. and Fiedler, G. M. (2013). Pancreatic carcinoma, pancreatitis, and healthy controls: Metabolite models in a three-class diagnostic dilemma, *Metabolomics*, **9**, 677–687.

Li, J. and Fine, J. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies, *Biostatistics*, **9**, 566–576.

Li, J. and Zhou, X.-H. (2009). Nonparametric and semiparametric estimation of the three-way receiver operating characteristic surface, *Journal of Statistical Planning and Inference*, **139**, 4133–4142.

Li, J. L.; Zhou, X.-H. and Fine, J. P. (2012). A regression approach to ROC surface, with applications to Alzheimer's disease, *Science China Mathematics*, **55**, 1583–1595.

Li, J. L.; Jiang, B. and Fine, J. P. (2013). Multicategory reclassification statistics for assessing improvements in diagnostic accuracy, *Biostatistics*, **14**, 382–394.

Luo, J. and Xiong, C. (2012). `DiagTest3Grp`: An R package for analyzing diagnostic tests with three ordinal groups, *Journal of Statistical Software*, **51**, 1–24.

Luo, J. and Xiong, C. (2013). Youden index and associated cut-points for three ordinal diagnostic groups, *Communications in Statistics—Simulation and Computation*, **42**, 1213–1234.

Migliaretti, G.; Ciaramitaro, P.; Berchialla, P.; Scarinzi, C.; Andrini, R.; Orlando, A. Faccani, G. (2013). Teleconsulting for minor head injury: The piedmont experience, *Journal of Telemedicine and Telecare*, **19**, 33–35.

Mossman, D. (1999). Three-way ROCs, *Medical Decision Making*, **19**, 78–89.

Nakas, C. T. and Alonzo, T. A. (2007). ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering, *Biometrics*, **63**, 603–609.

Nakas, C. T.; Alonzo, T. A. and Yiannoutsos, C. T. (2010). Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index, *Statistics in Medicine*, **29**, 2946–2955.

Nakas, C. T.; Dalrymple-Alford, J. C.; Anderson, T. and Alonzo, T. A. (2013). Generalization of Youden index for multiple-class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening, *Statistics in Medicine*, **32**, 995–1003.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437–3449.

Nze Ossima, A. D.; Daurès, J. P.; Bessaoud, F. and Trétarre, B. (2013). The generalized Lehmann ROC curves: Lehmann family of ROC surfaces, *Journal of Statistical Computation and Simulation*, DOI: 10.1080/00949655.2013.831863.

Obuchowski, N. A. (2006). An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale, *Statistics in Medicine*, **25**, 481–493.

Pencina, M. J.; D'Agostino Sr, R. B. and Demler, O. V. (2012). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvements for normal variables and nested models, *Statistics in Medicine*, **31**, 101–113.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy, *Biostatistics*, **1**, 123–140.

Ratnasamy, C.; Kinnamon, D. D.; Lipshultz, S. E. and Rusconi, P. (2008). Associations between neurohormonal and inflammatory activation and heart failure in children, *American Heart Journal*, **155**, 527–533.

Sahiner, B.; Chan, H. P.ăand Hadjiiski, L. M. (2008). Performance analysis of three-class classifiers: Properties of a 3-D ROC surface and the normalized volume under the surface for the ideal observer, *IEEE Transactions on Medical Imaging*, **27**, 215–227.

Schubert, C. M.; Thorsen, S. N. and Oxley, M. E. (2011). The ROC manifold for classification systems, *Pattern Recognition*, **44**, 350–362.

Scurfield, B. K. (1998). Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks, *Journal of Mathematical Psychology*, **42**, 5–31.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253–269.

Skaltsa, K.; Jover, L.; Fuster, D. and Carrasco, J. L. (2012). Optimum threshold estimation based on cost function in a multistate diagnostic setting, *Statistics in Medicine*, **31**, 1098–1109.

Shiu, S. Y. and Gatsonis, C. (2012). On ROC analysis with nonbinary reference standard, *Biometrical Journal*, **54**, 457–480.

Tian, L.; Xiong, C.; Lai, C. Y. and Vexler, A. (2011). Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups, *Journal of Statistical Planning and Inference*, **141**, 549–558.

Tremont, G.; Papandonatos, G. D.; Springate, B.; Huminski, B.; McQuiggan, M. D.; Grace, J.; Frakey, L. and Ott, B. R. (2011). Use of the telephone-administered Minnesota Cognitive Acuity Screen to detect mild cognitive impairment, *American Journal of Alzheimer's Disease and other Dementias*, **26**, 555–562.

Van Calster, B.; Van Belle, V.; Vergouwe, Y. and Steyerberg, E. W. (2012a). Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure, *Biometrical Journal*, **54**, 674–685.

Van Calster, B.; Vergouwe, Y.; Looman, C. W. N.; Van Belle, V.; Timmerman, D. and Steyerberg, E. W. (2012b). Assessing the discriminative ability of risk models for more than two outcome categories, *European Journal of Epidemiology*, **27**, 761–770.

Waegeman, W.; De Baets, B.ăand Boullart, L. (2008a). Learning layered ranking functions with structured support vector machines, *Neural Networks*, **21**, 1511–1523.

Waegeman, W.; De Baets, B. and Boullart, L. (2008b). On the scalability of ordered multi-class ROC analysis, *Computational Statistics and Data Analysis*, **52**, 3371–3388.

Wan, S. (2012). An empirical likelihood confidence interval for the volume under ROC surface, *Statistics and Probability Letters*, **82**, 1463–1467.

Wan, S. and Zhang, B. (2009). Semiparametric ROC surfaces for continuous diagnostic tests based on two test measurements, *Statistics in Medicine*, **28**, 2370–2383.

Wandishin, M. S. and Mullen, S. J. (2009). Multiclass ROC analysis, *Weather and Forecasting*, **24**, 530–547.

Wang, Z.; Zhou, X.-H. and Wang, M. (2011). Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard, *Biostatistics*, **12**, 567–581.

Xiong, C.; van Belle, G.; Miller, J. P. and Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups, *Statistics in Medicine*, **25**, 1251–1273.

Xiong, C.; van Belle, G.; Miller, J. P.; Yan, Y.; Gao, F.; Yu, K. and Morris, J. C. (2007). A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups, *Biometrical Journal*, **49**, 682–693.

Yang, H. and Carlin, D. (2000). ROC surface: A generalization of ROC curve analysis, *Journal of Biopharmaceutical Statistics*, **10**, 183–196.

Yang, H. and Zhao, L. (2010). A Method of Estimating and Comparing Volumes Under Receiver Operating Characteristic (ROC) Surfaces, *Statistics in Biopharmaceutical Research*, **2**, 279–291.

Yiannoutsos, C. T.; Nakas, C. T. and Navia, B. A. (2008). Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: Application to proton MR Spectroscopy (MRS) in HIV-related neurological injury, *Neuroimage*, **40**, 248–255.

Yu, T. (2012). ROCS: Receiver operating characteristic surface for class-skewed high-throughput data, *PLoS ONE*, **7**, at. no. e40598.

Zhang, Y. and Li, J. (2011). Combining multiple markers for multi-category classification: An ROC surface approach, *Australian and New Zealand Journal of Statistics*, **53**, 63–78.

Zhou, X.-H.; Obuchowski, N. A. and McClish, D. K. (2011). *Statistical Methods in Diagnostic Medicine*, Second Edition, Wiley, New York.

REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 67–83

VERIFICATION BIAS—IMPACT AND METHODS FOR CORRECTION WHEN ASSESSING ACCURACY OF DIAGNOSTIC TESTS

Author: Todd A. Alonzo
– Department of Biostatistics, University of Southern California, Los Angeles, CA, USA
talonzo@childrensoncologygroup.org

Abstract:

- Sometimes it is not feasible to obtain disease status verification for all study subjects. Analysis of only those with disease ascertainment can result in biased estimates of the accuracy (sensitivity, specificity, ROC curve) of a diagnostic test, screening test, or biomarker if the estimation method does not properly account for the missing disease ascertainment. This paper discusses the impact of this bias, verification bias, when estimating the accuracy of dichotomous and continuous diagnostic tests. In addition, methods to correct for verification bias are described. Areas that require additional attention are also highlighted.

Key-Words:

- *imputation; inverse probability weighting; ROC curve; sensitivity; specificity.*

AMS Subject Classification:

- 62F10, 62F15, 62J12, 62P10.

1. INTRODUCTION

Estimating accuracy of a diagnostic test, screening test, or biomarker is ideally done by determining disease status using a gold standard test or reference test for all study subjects. However, sometimes disease status verification via the reference test is not obtained for all study subjects because the reference test is too costly or invasive to be applied to all study subjects. When this is the case, subjects who appear to be at high risk may be more likely to have disease status assessed via the reference test than those who appear to be at lower risk. Analysis of only those with disease ascertainment can result in biased estimates of accuracy if the estimation methods do not properly account for nonrandom disease ascertainment. This bias is known as work-up bias (Ransohoff and Feinstein, 1978) and verification bias (Begg and Greenes, 1983). Verification bias can yield investigators to incorrectly conclude that a diagnostic test is more accurate than it is or the reverse that the test is less accurate than it actually is. This can have significant implications if the diagnostic test is implemented in practice based on incorrect conclusions.

Incomplete disease verification can occur by design or be unplanned. As expected, designed partial verification is more likely to occur in prospective studies while retrospective studies more typically have unplanned partial verification. In some studies it is not feasible to obtain the reference standard on subjects thought to be at low risk so the study is designed with partial verification. For example, the Prostate Cancer Prevention Study (Thompson *et al.*, 2005) of the effects of prostate specific antigen (PSA) the reference standard, prostate biopsy, was recommended only if the PSA level was greater than 4.0 ng/ml or rectal examination result was abnormal.

Methods for assessing accuracy of diagnostic tests differ depending on how the test is measured. Diagnostic tests can yield dichotomous results indicating presence or absence of particular condition or disease. For example, stress echocardiography to detect significant coronary artery stenosis. Diagnostic tests can also yield results that are measured on a continuous scale, such as, prostate specific antigen (PSA) for detecting prostate cancer. Typically, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are used to assess the accuracy of dichotomous diagnostic tests. Conversely, receiver operating characteristic (ROC) curves and corresponding summary measures, such as area under the ROC curve (AUC), are used to assess the accuracy of continuous tests.

Correcting for verification bias can be framed as a missing data problem where true disease status is missing for a subset of study subjects. Each approach for bias correction makes an assumption about the mechanism for the missing-

ness of disease verification (Little and Rubin, 1987). Disease status is considered missing completely at random (MCAR) if disease verification is independent of observed and unobserved data. Disease status is considered missing at random (MAR) when disease verification is only a function of observed data and is considered nonignorable (NI) when disease verification depends on unobserved data.

In Section 2 the notation for this paper is introduced. Sections 3 and 4 discuss the impact of verification bias when estimating the accuracy of a dichotomous diagnostic test and a continuous diagnostic test, respectively, and summarize available bias correction methods. We end with a Discussion.

2. NOTATION

Consider a study with n subjects on which the diagnostic test T is measured. Let D be disease status, as measured by a gold standard or reference test, where $D = 1$ corresponds to presence of disease and $D = 0$ corresponds to absence of disease. Further, let V be verification status where $V = 1$ if disease status is verified and $V = 0$ otherwise. There are n_V subjects with disease verification and $n_{\bar{V}} = n - n_V$ without disease verification.

3. DICHOTOMOUS TEST

Consider a dichotomous test T where $T = 1$ indicates a positive test and $T = 0$ indicates a negative test. Table 1 summarizes the observed data from a study of $n = n_1 + n_0$ subjects in which disease verification is not obtained in u_1 test positives and u_0 test negatives.

Table 1: Observed data for the verification bias problem when T is dichotomous.

V	D	$T = 1$	$T = 0$
1	1	s_1	s_0
1	0	r_1	r_0
0	Missing	u_1	u_0
Total:		n_1	n_0

3.1. Impact of bias

Consider a study of 1000 subjects to assess the sensitivity and specificity of a dichotomous screening test with a true sensitivity of 80%, true specificity of 90%, and disease prevalence, $P(D = 1)$, of 10%. Data from this hypothetical study are summarized on the left-hand side of Table 2. If the study design is such that disease verification is obtained for all subjects who test positive and only 10% of subjects who test negative, this can result in observing the data on the right-hand side of Table 2.

Table 2: Left side: results when disease verification is obtained for everyone. Right side: observed data when disease verification is obtained for all subjects who test positive and only 10% of subjects who test negative.

V	D	$T = 1$	$T = 0$
1	1	80	20
1	0	90	810
0	Missing	0	0
Total:		170	830

V	D	$T = 1$	$T = 0$
1	1	80	2
1	0	90	81
0	Missing	0	747
Total:		170	830

If we only consider test results for those with disease verification, referred to as complete case estimators, the observed sensitivity is $s_1/(s_1 + s_0) = 80/82$ or 98% and the observed specificity is $r_0/(r_0 + r_1) = 81/171$ or 47%. This illustrates that if test positives are more likely to receive disease verification than test negatives, observed sensitivity overestimates true sensitivity (98% *vs.* 80%) and observed specificity underestimates true specificity (47% *vs.* 90%). This verification bias can cause investigators to make incorrect conclusions regarding the accuracy of a test under evaluation.

It can be shown that PPV, $P(D = 1 \mid T = 1)$, is 47% using the full data and 47% using only those who received disease verification. Similarly, NPV, $P(D = 0 \mid T = 0)$ is 98% using the full data and also when only those who received disease verification are used. There is no bias in the complete case estimators of PPV and NPV because disease verification is only a function of the test results T, and PPV and NPV are, by definition, calculated conditional on T. See Zhou (1994) for a detailed discussion of the effect of verification bias on positive and negative predictive values. Next, we discuss methods to correct for the biased sampling when estimating sensitivity and specificity.

3.2. Bias correction methods

3.2.1. MAR approaches

Begg and Greenes (1983) developed a bias correction method for sensitivity and specificity by using Bayes' Rule and assuming disease status is MAR. First, consider estimating the sensitivity of a test. Bayes' Rule can be used to re-write sensitivity as

$$\begin{aligned} P(T=1 \mid D=1) &= \frac{P(T=1, D=1)}{P(D=1)} \\ &= \frac{P(D=1 \mid T=1)\, P(T=1)}{P(D=1 \mid T=1)\, P(T=1) + P(D=1 \mid T=0)\, P(T=0)} \,. \end{aligned} \tag{3.1}$$

Each quantity on the right-hand-side of (3.1) can be directly estimated from the observed data using empirical estimates. In particular, $P(T)$ can be estimated using data from all subjects, and $P(D \mid T)$ can be estimated using the verification group since by the MAR assumption $P(D \mid T) = P(D \mid T, V = 1)$. Substituting empirical estimates of the probabilities in (3.1) results in the following unbiased estimate of sensitivity

$$\widehat{P}(T=1 \mid D=1) = \frac{\frac{s_1 n_1}{s_1 + r_1}}{\frac{s_1 n_1}{s_1 + r_1} + \frac{s_0 n_0}{s_0 + r_0}} \,. \tag{3.2}$$

A bias-corrected estimate of specificity can be calculated in a similar fashion.

$$\widehat{P}(T=0 \mid D=0) = \frac{\frac{r_0 n_0}{s_0 + r_0}}{\frac{r_0 n_0}{s_0 + r_0} + \frac{r_1 n_1}{s_1 + r_1}} \,. \tag{3.3}$$

It can be shown that these estimators of sensitivity and specificity are maximum likelihood estimators. Furthermore, this approach can be considered single imputation as compared with multiple imputation which is discussed later. The delta method can be used to develop variance estimators for sensitivity and specificity.

Iglesias-Garriz *et al.* (2005) performed a study to estimate the sensitivity and specificity of stress echocardiography to detect significant coronary artery disease (CAD). The study involved 487 consecutive patients presenting at a hospital emergency room with nontraumatic chest pain, and who were administered stress echocardiography. Table 3 presents a tabulation of the study data, where using our notation T represents stress echocardiography, D is CAD, and V is an indicator of whether CAD status was determined. Of the 487 patients with stress echocardiography results, only 78 (16%) received disease verification via

coronary angiography to determine presence or absence of CAD. Furthermore, a higher percentage of those who tested positive with stress echocardiography received disease verification than those who tested negative with stress echocardiography (62.5% *vs.* 6.9%).

Table 3: Tabulation of Stress Echocardiography (T), CAD status (D), and disease verification status (V) in the study by Iglesias-Garriz *et al.*

V	D	$T=1$	$T=0$
1	1	43	15
1	0	7	13
0	Missing	30	379
Total:		80	407

Using only those with CAD status obtained, the complete case estimate of sensitivity is 74.1% (43/58) and the complete case estimate of specificity is 65.0% (13/20). Applying Equations 3.2 and 3.3, the Begg and Greenes estimate of sensitivity is 24.0% and corrected estimate of specificity is 94.4%. In this study, the uncorrected estimate of sensitivity clearly overestimates the corrected estimate while the uncorrected specificity substantially underestimates the corrected estimate.

Harel and Zhou (2006) discuss the use of multiple imputation to estimate sensitivity and specificity of a binary diagnostic test in the presence of verification bias. Each missing disease status is replaced by M imputed values and then each of the M complete data sets is analyzed using complete data methods. The M point estimates of sensitivity and specificity and their corresponding variances are combined to provide final estimates. The predictive distribution of the missing data is derived given the observed data and sampling iteratively from multinomial distribution and posterior distribution. Harel and Zhou conclude that the proposed estimators are better than the estimators of Begg and Greenes (Equations 3.2 and 3.3). However, there has been debate about the validity of this conclusion (Hanley *et al.*, 2007; Harel and Zhou, 2007). Subsequently, De Groot and colleagues identified computational errors in the work of Harel and Zhou (2006) which make it difficult to accurately draw conclusions from their work. Therefore, a separate comparison of the multiple imputation estimator and Begg and Greenes estimator was performed (De Groot *et al.*, 2011). The conclusion of this comparison is that both estimation methods yield similar results when the missing data mechanism is straightforward, but multiple imputation is recommended when the missing data mechanism is less straightforward or unknown.

3.2.2. NI approaches

If the decision to obtain disease verification depends on unrecorded factors related to disease, then the MAR assumption is not satisfied and the estimators discussed above could be biased. Zhou (1993) extended Begg and Greenes' method to allow a more general model for the verification process and derived the maximum likelihood estimators for the sensitivity and specificity of a diagnostic test and their corresponding variances. This approach does not assume D is MAR, but assumes that

$$\lambda_1 = \frac{P\big(V=1 \mid D=1, T=1\big)}{P\big(V=1 \mid D=0, T=1\big)}\ , \qquad \lambda_0 = \frac{P\big(V=0 \mid D=1, T=1\big)}{P\big(V=0 \mid D=0, T=1\big)}\ ,$$

are known. In other words, the ratio of the probability of selecting for verification a diseased patient with a given test result to that of selecting for verification a non-diseased patient with the same test result is known. In practice, however, λ_1 and λ_0 are not usually known and may be difficult to estimate. If $\lambda_1 = \lambda_0$, then Zhou's estimators reduce to those of Begg and Greenes.

Kosinski and Barnhart (2003) derive a region of all sensitivity and specificity values consistent with the observed data. This region is referred to as the test ignorance region. Recall that disease verification is not determined for u_1 test positives and u_0 test negatives. Of the u_1 test positives, let u_{1D} correspond to those truly diseased so there are $u_1 - u_{1D}$ test positives that are truly non-diseased. Similarly, let u_{0D} correspond to the truly diseased test negatives so there are $u_0 - u_{0D}$ test negatives that are truly non-diseased. If these values were known, then sensitivity (sens) and specificity (spec) can be estimated as

$$\text{sens} = \frac{s_1 + u_{1D}}{s_1 + u_{1D} + s_0 + u_{0D}}\ , \qquad \text{spec} = \frac{r_0 + u_0 - u_{0D}}{r_0 + u_0 - u_{0D} + r_1 + u_1 - u_{1D}}\ .$$

The test ignorance region is a plot of all sensitivity and specificity values resulting by considering all possible values of u_{1D} and u_{0D} in these equations.

An interactive web-based tool has been developed (Richardson and Petscavage (2010)) to implement the global sensitivity analysis of Kosinski and Barnhart. This tool is available at `http://uwmsk.org/gsa`. We illustrate this tool by using the coronary artery disease data summarized in Table 3. The region between the two curves in Figure 1 corresponds to the test ignorance region of all sensitivity and specificity values consistent with the observed data. The Begg and Greenes estimates (labeled MAR) fall in this region while the complete case or unadjusted estimates (labeled MCAR) fall outside the region and are therefore not compatible with the data.

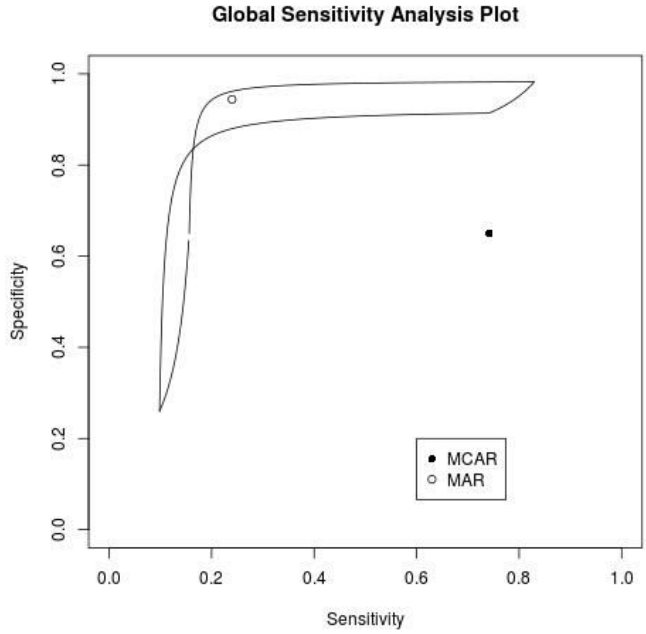


Figure 1: Global sensitivity analysis of the coronary artery disease data. MAR corresponds to Begg and Greenes estimates. MCAR corresponds to complete case estimates.

Baker (1995) and Kosinski and Barnhart (2003) propose likelihood-based regression approaches to deal with NI missingness when estimating the accuracy of a dichotomous test. These approaches require multiple diagnostic tests or covariates X. The approaches differ in how they factor the joint probability $P(V, T, D)$ as the product of conditional probabilities. Baker considered

$$P(V, T, D \mid X) \;=\; P(T \mid X)\, P(D \mid T, X)\, P(V \mid T, D, X)$$

while Kosinski and Barnhart considered

$$P(V, T, D \mid X) \;=\; P(D \mid X)\, P(T \mid D, X)\, P(V \mid T, D, X)\ .$$

The latter formulation is a product of the disease component $P(D \mid X)$, diagnostic test component $P(T \mid D, X)$, and missing data mechanism component $P(V \mid D, T, X)$. This formulation has the nice feature that sensitivity and specificity can be obtained directly from the diagnostic test component. Logistic regression models can be used to estimate parameters for each of the three components. When D is not included as a covariate in the missing data mechanism model, the missingness is MAR. Therefore, likelihood ratio, Wald, or Score tests can be used to test whether the MAR assumption is valid by testing whether the parameter is zero for D in the logistic regression model for $P(V \mid D, T, X)$. The expectation and maximization (EM) algorithm can be used to determine maximum likelihood estimates.

3.2.3. Bayesian approaches

Two Bayesian approaches have been developed to adjust for verification bias when estimating sensitivity and specificity of a binary diagnostic test. Both approaches allow for NI missingness. Martinez *et al.* (2006) describes an empirical Bayesian approach where Beta prior distributions are assumed for sensitivity, specificity, prevalence of disease, and the ratio of the probability of selecting for verification a diseased patient with a given test result to that of selecting for verification a non-diseased patient with the same test result is known (λ_1 and λ_0 considered by Zhou (1993)). Prior distributions for sensitivity and specificity are based on Begg and Greenes estimates of sensitivity and specificity and non-informative priors are used for the other parameters. The Gibbs sampling algorithm is used to estimate marginal posterior densities for all parameters.

Buzoianu and Kadane (2008) use the formulation $P(V,T,D)$ is equal to $P(D)\,P(T \mid D)\,P(V \mid T,D)$ considered by Kosinski and Barnhart (2003) to accommodate NI missingness. Similar to Kosinksi and Barnhart, logistic regression models can be used for each component. Prior distributions are used for the parameters in the logistic models.

4. CONTINUOUS TEST

Consider a continuous test T where higher values of T are more indicative of disease. The accuracy of a continuous diagnostic test is typically assessed using an ROC curve. An ROC curve is a plot of the true positive rate (TPR), sensitivity, versus the false positive rate (FPR), one minus the specificity, associated with all the dichotomous tests that can be formed by varying the cut point that defines a positive dichotomous test. When all subjects are verified, TPR and FPR can be estimated nonparametrically for a particular cutpoint c by using

$$\widehat{\mathrm{TPR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\, D_i}{\sum_{i=1}^{n} D_i}\ , \qquad \widehat{\mathrm{FPR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\,(1-D_i)}{\sum_{i=1}^{n}(1-D_i)}\ .$$

4.1. Impact of bias

Complete case estimators only use data from subjects who received disease verification. That is,

$$\widehat{\mathrm{TPR}}(c)_{CC} = \frac{\sum_{i=1}^{n} I(T_i \geq c)\, V_i D_i}{\sum_{i=1}^{n} V_i D_i}\ , \qquad \widehat{\mathrm{FPR}}(c)_{CC} = \frac{\sum_{i=1}^{n} I(T_i \geq c)\, V_i (1-D_i)}{\sum_{i=1}^{n} V_i (1-D_i)}\ .$$

The complete case estimator yields unbiased estimates of the ROC curve and corresponding AUC when disease verification is MCAR. If the missing data mechanism is not MCAR, the complete case estimator can yield biased estimates of the ROC curve by overestimating $\text{TPR}(c)$ and $\text{FPR}(c)$ for each cutpoint c that results in operating points on the ROC curve that are biased upwards relative to the full data curve and thus underestimates the ROC curve and corresponding AUC. However, the complete case approach can also overestimate the ROC curve and AUC depending on the verification mechanism and accuracy of T (Alonzo and Pepe, 2005).

4.2. Bias correction—ROC curve

4.2.1. MAR approaches

Alonzo and Pepe (2005) proposed several bias-corrected estimators of TPR and FPR that assume disease status is MAR. Bias-corrected ROC curves are obtained by plotting bias-corrected estimators of TPR and FPR for all cutpoints. One approach for bias correction is to use full imputation (FI) over the distribution $P(D \mid T, X)$. That is, FI imputes $\rho = P(D \mid T, X)$ for all subjects in the study which results in the following estimators

$$\widehat{\text{TPR}}_{\text{FI}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ \widehat{\rho}_i}{\sum_{i=1}^{n} \widehat{\rho}_i}\ , \qquad \widehat{\text{FPR}}_{\text{FI}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ (1-\widehat{\rho}_i)}{\sum_{i=1}^{n} (1-\widehat{\rho}_i)}\ ,$$

where $\widehat{\rho}_i$ is an estimate of $P(D_i = 1 \mid T_i, X_i)$ that can obtained using, for example, logistic regression. By the MAR assumption, the disease model $P(D = 1 \mid T, X)$ can be estimated using the verification sample. When T and X are discrete and a saturated model is used, these estimators of TPR and FPR reduce to the Begg and Greenes (1983) bias-corrected estimators of sensitivity and specificity presented in the previous section.

Another approach for bias correction is to use mean score imputation (MSI) where the observed disease status is used for those in the verification sample and disease status is imputed for subjects not in the verification sample. That is,

$$\widehat{\text{TPR}}_{\text{MSI}}(c) \ = \ \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i\, D_i + (1-V_i)\, \widehat{\rho}_i \right\}}{\sum_{i=1}^{n} \left\{ V_i\, D_i + (1-V_i)\, \widehat{\rho}_i \right\}}\ ,$$

$$\widehat{\text{FPR}}_{\text{MSI}}(c) \ = \ \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i\, (1-D_i) + (1-V_i)\, (1-\widehat{\rho}_i) \right\}}{\sum_{i=1}^{n} \left\{ V_i\, (1-D_i) + (1-V_i)\, (1-\widehat{\rho}_i) \right\}}\ .$$

Again, the MAR assumption implies that data from the verification sample can be used to obtain valid estimates of ρ_i.

Alonzo and Pepe (2005) also propose the following inverse probability weighting (IPW) estimators (Horvitz and Thompson, 1952) that weight each observation in the verification sample by the inverse of the sampling fraction (*i.e.* probability the subject was selected for verification)

$$\widehat{\text{TPR}}_{\text{IPW}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ V_i\, D_i \,/\, \widehat{\pi}_i}{\sum_{i=1}^{n} V_i\, D_i \,/\, \widehat{\pi}_i} ,$$

$$\widehat{\text{FPR}}_{\text{IPW}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ V_i(1-D_i) \,/\, \widehat{\pi}_i}{\sum_{i=1}^{n} V_i(1-D_i) \,/\, \widehat{\pi}_i} ,$$

where $\widehat{\pi}_i = P(V_i = 1 \,|\, T_i, X_i)$ may be known or may need to be estimated depending on the design of the study. The IPW estimators are similar to the CC estimators in that they use the observed disease status for the verification sample. Unlike the CC, however, they correct for the biased sampling by weighting the observed value by the probability the subject was verified.

The following doubly robust (DR) estimators have also been proposed:

$$\widehat{\text{TPR}}_{\text{DR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i D_i/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, \widehat{\rho}_i/\widehat{\pi}_i \right\}}{\sum_{i=1}^{n} \left\{ V_i D_i/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, \widehat{\rho}_i/\widehat{\pi}_i \right\}} ,$$

$$\widehat{\text{FPR}}_{\text{DR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i(1-D_i)/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, (1-\widehat{\rho}_i)/\widehat{\pi}_i \right\}}{\sum_{i=1}^{n} \left\{ V_i(1-D_i)/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, (1-\widehat{\rho}_i)/\widehat{\pi}_i \right\}} .$$

These estimators are referred to as doubly robust because they are consistent if either π_i or ρ_i is estimated consistently. That is, the verification model or disease model can be incorrectly specified and consistency is still guaranteed. These estimators have also been referred to as semiparametric because they require parametric conditional mean models to be specified for the disease model $P(D \,|\, T, X)$ and for the verification model $P(V \,|\, T, X)$ but are non-parametric with respect to the joint distribution of the data $P(D, T, X)$.

Alonzo and Pepe (2005) illustrated that misspecifying the verification model yields biased IPW estimates of the ROC curve and misspecifying the disease model results in biased FI and MSI. Furthermore, they showed the DR estimator of the ROC curve is unbiased if either the model for verification or the model for disease is correctly specified. Thus, they recommend the DR approach is used in practice.

The AUC can be estimated empirically for each of the bias-corrected ROC curves described above by using the Trapezoidal Rule (Bamber, 1975). Closed-form expressions for the AUC corresponding to the IPW and DR ROC estimators have been obtained as well as variance expressions (He *et al.*, 2009).

4.2.2. NI approaches

Rotnitzky *et al.* (2006) describe a DR estimator of the AUC. They note that AUC is identified under the untestable assumption

$$\log\left\{\frac{P(V=0\,|\,T,X,V)}{P(V=1\,|\,T,X,V)}\right\} \;=\; h(T,V)+q(T,V)\,X\;, \tag{4.1}$$

where $q(T,V)$ is an arbitrary specified function and $h(T,V)$ is an arbitrary unknown function. $q(T,V)$=0 for all T and V corresponds to the MAR assumption while $q(T,V)\neq 0$ corresponds to NI missingness. Fluss *et al.* (2009) extend the approach of Rotnitzky *et al.* (2006) to obtain a DR estimate of TPR and FPR and, thus, the empirical ROC curve that allows for NI missingness. They recommend performing a sensitivity analysis by repeating the estimation of TPR and FPR under a variety of reasonable choices for the selection bias function q. Conversely, Liu and Zhou (2010) use the likelihood approach to estimate a nonignorable parameter and obtain DR estimates of the ROC curve and AUC. They assume the disease verification model

$$P\big(V_i=1\,|\,D_i,T_i,X_i\big) \;=\; \frac{\exp(x)}{1+\exp(x)}\left\{h(T_i,X_i;\beta)+\alpha\,D_i\right\}\;,$$

where α is the NI parameter and $h(T_i,X_i;\beta)=\beta_0+\beta_1T_i+\beta_2X_i$. Since the nonignorable parameter cannot be tested nonparametrically, Liu and Zhou recommend that scientific knowledge is used to construct an appropriate disease verification model.

4.3. Covariate-adjusted ROC curves

The accuracy of a diagnostic test can be affected by factors such as disease severity, age, and gender. ROC curves have been adjusted for age in the assessment, for example, of the accuracy of fingerstick postprandial blood glucose measurements to discriminate between healthy and diseased subjects in the presence of verification bias (Fluss *et al.*, 2012).

Page and Rotnitzky (2009) discuss a parametric model for estimating the covariate-specific ROC curve in the presence of verification bias. They make the assumption that the ROC curve has an underlying binormal distribution and disease verification has NI missingness. Liu and Zhou (2011) discuss a likelihood approach to estimate the covariate-specific ROC curve in the presence of verification bias. Disease verification is assumed to be MAR and diagnostic test results are modeled using a location-scale model. Weighted estimating equations are used to estimate the parameters in the location-scale model. DR, IPW, and imputation approaches are compared for the estimation. Liu and Zhou conclude that

the DR estimator performed best in their simulation studies and their method is sensitive to the location-scale model assumption.

Fluss *et al.* (2012) develop a DR method for estimating the ROC curve adjusted for covariates for a NI missing data mechanism. Using the approach of Pepe (1998), they model the diagnostic test values distribution as a function of disease status and covariate values using a semi-parametric location-scale model. Since the proposed approach relies on the untestable specification of $q(T, V)$ (see Equation 4.1), the authors recommend a sensitivity analysis is performed to examine the sensitivity of the estimated ROC curve to the specified form of $q(T, V)$.

5. DISCUSSION

This paper highlights methods available for estimating the accuracy of dichotomous and continuous diagnostic tests in the presence of verification bias. More recently, this bias has also been referred to as partial verification bias so as not to be confused with differential disease verification in which a subset of study subjects have a different reference standard to determine disease status (Whiting, 2004).

As investigators design future studies of test accuracy, it is important to record all factors that may affect the decision to offer and receive disease verification. In cases where all factors are captured, then the MAR assumption will likely be satisfied and bias-correction methods that rely on this assumption can be used. When all factors that impact disease verification are not collected, it is preferred to use bias-correction methods that allow for NI missingness.

The focus of this paper is on the estimation of the sensitivity and specificity of a single dichotomous test and the ROC curve and AUC for a single continuous test in the presence of verification bias. Bias correction methods are also available for diagnostic tests measured on an ordinal scale (Gray *et al.*, 1984; Hunink *et al.*, 1990; Baker, 1995; Toledano and Gatsonis, 1996; Rodenberg and Zhou, 2000), such as a radiologist's interpretations of images to quantify the suspicion of cancer. In addition, methods have been developed to estimate the difference between two diagnostic tests in regards to bias-corrected sensitivity and specificity. Assuming disease verification is MAR, Zhou (1998) and Roldán Nofuentes and Luna del Castillo (2008) provide estimators for the difference in bias-corrected sensitivity and specificity.

This paper considers the setting when there are only two disease states (diseased and non-diseased). In some settings there can be more than two disease states. For example, Alzheimer's Disease dementia can be classified into more

than two categories. Chi and Zhou (2008) propose a non-parametric likelihood-based approach to construct the empirical ROC surface (extension of ROC curve to more than two disease states) and estimate the volume under the ROC surface in the presence of verification bias for ordinal diagnostic tests. Future work is needed to develop bias correction methods for estimating the ROC surface and volume under the ROC surface for continuous diagnostic tests.

The bias correction methods described in this paper, especially for continuous tests, would benefit from the development and distribution of code to apply the methods in practice. Increasing the availability of these methods in standard statistical packages would likely increase the use of the methods.

REFERENCES

Alonzo, T. A. and Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias, *Journal of the Royal Statistical Society, Ser.* C, **54**, 173–190.

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

Baker, S. G. (1995). Evaluating multiple diagnostic tests with partial verification, *Biometrics*, **51**, 330–337.

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease is subject to selection bias, *Biometrics*, **39**, 207–216.

Buzoianu, M. and Kadane, J. B. (2008). Adjusting for verification bias in diagnostic test evaluation: a Bayesian approach, *Statistics in Medicine*, **27**, 2453–2473.

Chi, Y.-Y. and Zhou, X.-H. (2008). Receiver operating characteristic surfaces in the presence of verification bias, *Journal of the Royal Statistical Society, Ser.* C, **57**, 1–23.

De Groot, J. A. H.; Janssen, K. J. M.; Kristel, J. M.; Zwinderman, A. H.; Bossuyt, P. M. M.; Reitsma, J. B. and Moons, K. G. M. (2011). Correcting for partial verification bias: a comparison of methods, *Annals of Epidemiology*, **21**, 139–148.

De Groot, J.A.H.; Janssen, K.J.M.; Zwinderman, A.H.; Moons, K.G.M. and Reitsma, J. B. (2008). Multiple imputation to correct for partial verification bias revisited, *Statistics in Medicine*, **27**, 5880–5889.

Fluss, R.; Reiser, B. and Faraggi, D. (2012). Adjusting ROC curves for covariates in the presence of verification bias, *Journal of Statistical Planning and Inference*, **142**, 1–11.

Fluss, R.; Reiser, B.; Faraggi, D. and Rotnitzky, A. (2009). Estimation of the ROC curve under verification bias, *Biometrical Journal*, **51**, 475–490.

Gray, R.; Begg, C. and Greenes, R. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias, *Medical Decision Making*, **4**, 151–164.

Hanley, J. A.; Dendukuri, N. and Begg, C. B. (2007). Multiple imputation for correcting verification bias by Ofer Harel and Xiao-Hua Zhou, *Statistics in Medicine*, **26**, 3046–3047.

Harel O. and Zhou X.-H. (2006). Multiple imputation for correcting verification bias, *Statistics in Medicine*, **25**, 3769–3786.

Harel O. and Zhou X.-H. (2007). Rejoinder to multiple imputation for correcting verification bias, *Statistics in Medicine*, **26**, 3047–3050.

He, H.; Lyness, J. M. and McDermott, M. P. (2009). Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias, *Statistics in Medicine*, **28**, 361–376.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Hunink, M. G. M.; Richardson, D. K.; Doubilet, P. M. and Begg, C. B. (1990). Testing for fetal pulmonary maturity ROC analysis involving covariates, verification bias, and combination testing, *Medical Decision Making*, **10**, 201–211.

Iglesias-Garriz, I.; Rodríguez, M. A.; García-Porrero, E.; Ereño, F.; Garrote, C. and Suarez, G. (2005). Emergency nontraumatic chest pain: use of stress echocardiography to detect significant coronary artery stenosis, *Journal of the American Society of Echocardiography*, **18**, 1181–1186.

Kosinski, A. S. and Barnhart, H. X. (2003). A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present, *Statistics in Medicine*, **22**, 2711–2721.

Kosinski, A. S. and Barnhart, H. X. (2003). Accounting for nonignorable verification bias in assessment of diagnostic tests, *Biometrics*, **59**, 163–171.

Little, R. J. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Liu, D. and Zhou, X.-H. (2010). A model for adjusting for nonignorable verification bias in estimation of the ROC curve and its area with likelihood-based approach, *Biometrics*, **66**, 1119–1128.

Liu, D. and Zhou, X.-H. (2011). Semiparametric Estimation of the Covariate-Specific ROC Curve in Presence of Ignorable Verification Bias, *Biometrics*, **67**, 906–916.

Martinez, E. Z.; Alberto Achcar, J. and Louzada-Neto, F. (2006). Estimators of sensitivity and specificity in the presence of verification bias: A Bayesian approach, *Computational Statistics and Data Analysis*, **51**, 601–611.

Page, J. H. and Rotnitzky, A. (2009). Estimation of the disease-specific diagnostic marker distribution under verification bias, *Computational Statistics and Data Analysis*, **53**, 707–717.

Pepe, M. S. (1998). Regression analysis of ROC curves, *Biometrics*, **54**, 124–135.

Ransohoff, D. F. and Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests, *New England Journal of Medicine*, **299**, 926–930.

Richardson, M. L. and Petscavage, J. M. (2010). An interactive web-based tool for detecting verification (work-up) bias in studies of the efficacy of diagnostic imaging, *Academic Radiology*, **17**, 1580–1583.

Rodenberg, C. and Zhou, X.-H. (2000). ROC curve estimation when covariates affect the verification process, *Biometrics*, **56**, 131–136.

Roldán Nofuentes, J. A. and Luna del Castillo, J. D. (2008). EM algorithm for comparing two binary diagnostic tests when not all the patients are verified, *Journal of Statistical Computation and Simulation*, **78**, 19–35.

Rotnitzky, A.; Faraggi, D. and Schisterman, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias, *Journal of the American Statistical Association*, **101**, 1276–1288.

Toledano, A. and Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data, *Statistics in Medicine*, **15**, 1807–1826.

Thompson, I. M.; Ankerst, D. P.; Chi, C.; Lucia, M. S.; Goodman, P. J.; Crowley, J. J.; Parnes, H. L. and Coltman Jr, C. A. (2005). Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower, *The Journal of the American Medical Association*, **294**, 66–70.

Whiting, P.; Rutjes, A. W. S.; Reitsma, J. B.; Glas, A. S.; Bossuyt, P. M. M. and Kleijnen, J. (2004). Sources of Variaton and Bias in Studies of Diagnostic Accuracy: A Systematic Review, *Annals of Internal Medicine*, **140**, 189–202.

Zhou, X.-H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias, *Communications in Statistics—Theory and Methods*, **22**, 3177–3198.

Zhou, X.-H. (1994). Effect of verification bias on positive and negative predictive values, *Statistics in Medicine*, **13**, 1737–1745.

Zhou, X.-H. (1998). Comparing accuracies of two screening tests in a two-phase study for dementia, *Journal of the Royal Statistical Society, Ser.* C, **47**, 135–147.

REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 85–99

MODELING WITHOUT A GOLD STANDARD: STRATIFICATION WITH STRATUM-DEPENDENT PARAMETERS

Authors: Francisco Louzada
– Department of Applied Mathematics and Statistics,
Institute of Mathematical Science and Computing,
University of São Paulo, São Carlos, Brazil
louzada@icmc.usp.br

Gilberto de Araujo Pereira
– Department of Nursing,
Federal University of Triângulo Mineiro, Uberaba, Brazil
pereira_gilberto@yahoo.com.br

Márcia M. Ferreira-Silva
– Federal University of Triângulo Mineiro, Uberaba, Brazil
marcia.mferreira@yahoo.com.br

Valdirene de Fátima Barbosa
– Research Group for Blood Transfusion Security,
Federal University of Triângulo Mineiro, Uberaba, Brazil
valdirene_fbarbosa@yahoo.com.br

Helio de Moraes-Souza
– Department of Medical Clinic,
Federal University of Triângulo Mineiro, Uberaba, Brazil
helio.moraes@dcm.uftm.edu.br

Gleici S. Castro Perdona
– Departament of Social Medicine,
University of São Paulo, Ribeirão Preto, Brazil
pgleici@fmrp.usp.br

Abstract:

- Bayesian latent-class models have been widely applied for assessing the performance of diagnostic tests in the absence of a gold standard. We provide a short discussion on identifiability issues appearing under the absence of a gold standard, and construct an extension of the well-known Hui–Walter stratification model which allows for stratum-dependent parameters. We illustrate our approach using a Chagas disease case study on blood donors from Brazil.

Key-Words:

- *absence of a gold standard; diagnostic test; identifiability; sample size; stratification.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

In the area of diagnostic medicine, it is common that the medical practitioner considers one or more complementary diagnostic tests for decision-making and detailed clinical analysis. Within this context, it is important that the physician knows the parameters of the test to be used, such as sensitivity and/or specificity, false-positive and/or false-negative rates, and positive and/or negative predictive values. The modeling structure for this estimation problem is relatively simple and straightforward when the subjects being investigated are submitted to the so-called gold standard test for confirmation, as they are usually 100% sensitive and specific (Kraemer, 1992).

However, in many practical situations no patient under investigation is submitted to a confirmatory test (Joseph *et al.*, 1995), either due to the lack of such a test or its high invasiveness, or to the high cost of its large scale implementation, or to the presence of subgroups with different prevalence rates (Hui and Walter, 1980).

Our main objectives here are to provide a short discussion on identifiability issues appearing under the absence of a gold standard, and to construct an extension of the Hui–Walter stratification model which allows for stratum-dependedent performance parameters. In the next section we discuss the modeling concepts and the inference techniques. In Section 3 we report details on numerical experiments, and we provide an illustration to Chagas disease data in Section 4.

2. MODELING WITHOUT A GOLD STANDARD

2.1. Absence of gold standard

In the case where the health condition of a subject (D) cannot be verified, due to the absence of a gold standard, the likelihood for a random sample of n subjects, can be written as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{l=1}^{L}\left\{\xi\,\mathrm{se}_l^{t_{i,l}}(1-\mathrm{se}_l)^{1-t_{i,l}} + (1-\xi)\,\mathrm{sp}_l^{1-t_{i,l}}(1-\mathrm{sp}_l)^{t_{i,l}}\right\}, \tag{2.1}$$

where, $\boldsymbol{\theta} = (\xi, \mathbf{se}, \mathbf{sp})^{\mathrm{T}}$, with ξ denoting the disease prevalence, and

$$\mathbf{se} = (\mathrm{se}_1, ..., \mathrm{se}_L)^{\mathrm{T}}\,, \qquad \mathbf{sp} = (\mathrm{sp}_1, ..., \mathrm{sp}_L)^{\mathrm{T}}\,,$$

Here se_l and sp_l are respectively the sensitivity and specificity of the lth test, and $t_{i,l}$ is the outcome of the lth diagnostic test on the ith subject (0: negative, 1: positive). In this model there are $2L+1$ parameters to be estimated, and $2^L - 1$ degrees of freedom.

A popular approach for modeling the performance of diagnostic tests, under the absence of a gold standard, is to consider latent classes. In this setting, the health condition y_i of the ith subject (healthy or diseased) can be modeled through a Bernoulli random variable, Y, with probability of success,

$$\tau_i = \frac{\xi \prod_{l=1}^{L} \text{se}_l^{t_{i,l}} (1-\text{se}_l)^{1-t_{i,l}}}{\xi \prod_{l=1}^{L} \text{se}_l^{t_{i,l}} (1-\text{se}_l)^{1-t_{i,l}} + (1-\xi) \prod_{l=1}^{L} \text{sp}_l^{1-t_{i,l}} (1-\text{sp}_l)^{t_{i,l}}} . \tag{2.2}$$

By combining the likelihood of the incomplete data (2.1) with the likelihood of the latent variable, Y, we can write the augmented likelihood (Dempster *et al.*, 1977; Tanner and Wong, 1987) for the case where L diagnostic tests are conducted, as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{l=1}^{L} \left[\left\{ \xi \, \text{se}_l^{t_{i,l}} (1-\text{se}_l)^{1-t_{i,l}} \right\}^{y_i} \left\{ (1-\xi) \, \text{sp}_l^{1-t_{i,l}} (1-\text{sp}_l)^{t_{i,l}} \right\}^{1-y_i} \right], \tag{2.3}$$

where y_i is the unobservable health condition of the ith subject (0: healthy; 1: diseased), which is modeled through a Bernoulli distribution with probability of success τ_i as given in (2.2). Estimation can then be conducted through numeric methods, such as the Expectation-Maximization algorithm (EM) (Dempster *et al.*, 1977), in the frequentist context, and Gibbs sampling (Gelfand and Smith, 1990) or a Metropolis–Hastings algorithm (Chib and Greenberg, 1995), in the Bayesian context.

According to Swartz *et al.* (2004), a primary difficulty regarding latent-class models is related to identifiability issues, and one of the practical lessons obtained by using them is that this issue becomes relatively less important as the dimension of the model increases.

2.2. Identifiability

The modeling approach discussed in §2.1 has been widely applied in the literature, for the case where the model obeys what we will call throughout as the *basic identifiability condition*,

$$\text{df} \geq p \, , \tag{2.4}$$

where df is the number of degrees of freedom, and p is the number of parameters to be estimated. For example, for the latent-class model in §2.1 to obey the basic identifiability condition, $\text{df} = 2^L - 1 \geq 2L + 1 = p$, a minimum of three conditionally independent tests is required.

Several procedures for assessing identifiability have been documented in the literature. For example, Goodman (1974) discusses a Jacobian-based criterion, whereas Garret and Zeger (2000) proposes a graphical method to assess weak identifiability, which is based on the idea that weak identifiability is associated with smaller sample sizes relatively to the number of latent classes, case in which the number of subjects may be insufficient to assign an element to each class.

The Bayesian approach offers here an important advantage: Although a certain model may not be identifiable, it is always valid as data can be suitably described from both its identifiable parameters and prior information (Lindley, 1971); this point is reinforced by Neath and Samaniego (1997), who support the view that Bayesian analysis may yield reasonable answers even for nonindentifiable models.

2.2.1. Hui–Walter stratification

To reestablish the basic identifiability condition many approaches have been considered, such as the introduction of constraints on the parameter space (Walter and Irwig, 1988), the choice of informative priors according to well defined criteria (Gustafson, 2005), or stratification-based approaches (Hui and Walter, 1980). These latter approaches are known as the Hui–Walter paradigm, and will be of particular interest for the remainder of this article; the Hui–Walter stratification paradigm has been widely discussed in the literature, and it has been modeled through a wealth of Bayesian and frequentist approaches (Singer *et al.*, 1998; Johnson *et al.* 2001; Nielsen *et al.*, 2002; Gustafson, 2005; Gardner, 2004; Toft *et al.*, 2005; Branscum *et al.*, 2005; Bertrand *et al.*, 2005; Toft *et al.*, 2007, among others).

The Hui–Walter stratification model is based on stratum-dependent disease prevalence rates, although it uses equal performance parameters across strata. Stratification increases the number of parameters to $2^L + V$ and the number of degrees of freedom to $2^L V - V$; hence, if the population is divided into two strata ($V = 2$), a minimum of two conditionally independent tests ($L = 2$) is sufficient to obey the basic condition for identifiability (2.4). As a byproduct, stratification also allows us estimate specific disease prevalence rates in each homogeneous subpopulation.

For the absence of gold standard, the likelihood of the Hui–Walter stratifi-

cation model can be written as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{v=1}^{V}\prod_{i=1}^{n_v}\left[\left\{\xi_v\prod_{l=1}^{L}\mathrm{se}_l^{t_{i,l}}(1-\mathrm{se}_l)^{1-t_{i,l}}\right\}^{y_{i,v}} \times\left\{(1-\xi_v)\prod_{l=1}^{L}\mathrm{sp}_l^{1-t_{i,l}}(1-\mathrm{sp}_l)^{t_{i,l}}\right\}^{1-y_{i,v}}\right],$$

where n_v and ξ_v are respectively the number of subjects and the prevalence rate in the vth stratum, whereas $y_{i,v}$ is the unobservable health condition of the ith subject in the vth stratum, modeled through a Bernoulli distribution with probability of success $\tau_{i,v}$.

Toft *et al.* (2005) pointed out some potential pitfalls of the Hui–Walter paradigm, particularly regarding the accuracy of estimates, which are strongly influenced by the magnitude of the difference in disease prevalence rates between strata, suggesting that the greater the difference of the prevalence rates the higher the estimation accuracy (smaller amplitude of 95% credibility interval) in the case of two tests ($L = 2$) and two strata ($V = 2$). Moreover, sensitivity and specificity may be overestimated.

2.2.2. Extended stratification

Since in most practical situations it is rather challenging to find a stratification factor in which both sensitivity and specificity of the tests are kept similar across strata, here we propose an extension of the Hui–Walter model which assumes that not only prevalences (ξ)—but also sensitivities and specificities—are stratum-dependent. Specifically, our setting is the following: We assume that L diagnostic tests are conducted—none of which being a gold standard—and we assume that the population is divided into V strata, with stratum-dependent prevalences

$$\left\{\xi_v = P_v(D=1)\colon\ v = 1, ..., V\right\},$$

and with stratum-dependent performance parameters,

$$\left\{(\mathrm{se}_{l,v}, \mathrm{sp}_{l,v})\colon\ l = 1, ..., L;\ v = 1, ..., V\right\}.$$

The unobservable health condition of a subject in the vth stratum, Y_v, can be modeled through a Bernoulli distribution, with probability of success τ_v. With our extension of the Hui–Walter model, the number of parameters increases to $2LV + V$, whereas the number of degrees of freedom remains unchanged ($\mathrm{df} = 2^L V - V$). This means that, for example, for a population stratified into two strata ($V = 2$), at least three tests need to be conducted ($L \geq 3$), so that the model obeys the basic condition for identifiability (2.4). (Compare with §2.2.1.)

The augmented data likelihood of the latent class model, for the general case of L conditionally independent tests and V strata, can be written as

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{v=1}^{V}\prod_{i=1}^{n_v}\left[\left\{\xi_v \prod_{l=1}^{L} \mathrm{se}_{l,v}^{t_{i,l,v}}(1-\mathrm{se}_{l,v})^{1-t_{i,l,v}}\right\}^{y_{i,v}} \times\left\{(1-\xi_v)\prod_{l=1}^{L}\mathrm{sp}_{l,v}^{1-t_{i,l,v}}(1-\mathrm{sp}_{l,v})^{t_{i,l,v}}\right\}^{1-y_{iv}}\right], \tag{2.5}$$

where, $\boldsymbol{\theta} = (\boldsymbol{\xi}, \mathbf{se}_1, ..., \mathbf{se}_V, \mathbf{sp}_1, ..., \mathbf{sp}_V)^{\mathrm{T}}$ with

$$\boldsymbol{\xi} = (\xi_1, ..., \xi_V)^{\mathrm{T}}, \qquad \mathbf{se}_l = (\mathrm{se}_{1,v}, ..., \mathrm{se}_{L,v})^{\mathrm{T}}, \qquad \mathbf{sp}_l = (\mathrm{sp}_{1,v}, ..., \mathrm{sp}_{L,v})^{\mathrm{T}},$$

for $v = 1, ..., V$. Here, ξ_v is prevalence rate in the vth stratum, whereas $\mathrm{se}_{l,v}$ and $\mathrm{sp}_{l,v}$ are the sensibility and specificity of lth test in the vth stratum, respectively; in addition, $t_{i,l,v}$ is the lth test result for the ith subject in the vth stratum, and $y_{i,v}$ is the unobservable health condition of the ith subject in the vth stratum, which is modeled through a Bernoulli with success probability,

$$\tau_{i,v} = \frac{\xi_v \prod_{l=1}^{L} \mathrm{se}_{l,v}^{t_{i,l,v}}(1-\mathrm{se}_{l,v})^{1-t_{i,l,v}}}{\xi_v \prod_{l=1}^{L} \mathrm{se}_{l,v}^{t_{i,l,v}}(1-\mathrm{se}_{l,v})^{1-t_{i,l,v}} + (1-\xi_v)\prod_{l=1}^{L}\mathrm{sp}_{l,v}^{1-t_{i,l,v}}(1-\mathrm{sp}_{l,v})^{t_{i,l,v}}},$$

for $i = 1, ..., n_v$ and $v = 1, ..., V$.

The non-stratified model ($V = 1$) in (2.3), and the Hui–Walter model in (2.2.1) are particular cases of our stratification model with stratum-dependent parameters.

2.2.3. Inference

A fully Bayesian approach is here used for conducting inference. This choice is based on the fact that each parameter in (2.5) is directly interpreted within the context of diagnostic tests, including the availability of expert opinions that can be modeled separately in terms of prior distribution for each parameter. We consider Beta$(1, 1)$ prior distributions for the components of $\boldsymbol{\theta}$, all independent among them; by combining the likelihood (2.5) with the joint prior of $\boldsymbol{\theta}$ we obtain the joint posterior and full conditionals, which can then be used in a Gibbs sampler, and which are given by

$$\begin{aligned} \xi_v \mid \boldsymbol{X}_{\xi_v} &\sim \mathrm{Beta}(\alpha_{\xi_v}, \beta_{\xi_v}), \\ \mathrm{se}_{l,v} \mid \boldsymbol{X}_{\mathrm{se}_{l,v}} &\sim \mathrm{Beta}(\alpha_{\mathrm{se}_{l,v}}, \beta_{\mathrm{se}_{l,v}}), \\ \mathrm{sp}_{l,v} \mid \boldsymbol{X}_{\mathrm{sp}_{l,v}} &\sim \mathrm{Beta}(\alpha_{\mathrm{sp}_{l,v}}, \beta_{\mathrm{sp}_{l,v}}), \end{aligned} \tag{2.6}$$

where, $\boldsymbol{X}_{\xi_v} = \{a_{\xi_v}, b_{\xi_v}, y_{i,v}, n_v\}$,

$$\boldsymbol{X}_{\mathrm{se}_{l,v}} = \{a_{\mathrm{se}_{l,v}}, b_{\mathrm{se}_{l,v}}, t_{i,l,v}, y_{i,v},\} \ , \qquad \boldsymbol{X}_{\mathrm{sp}_{l,v}} = \{a_{\mathrm{sp}_{l,v}}, b_{\mathrm{sp}_{l,v}}, t_{i,l,v}, y_{i,v}\} \ ,$$

and

$$\alpha_{\xi_v} = \sum_{i=1}^{n_v} y_{i,v} + a_{\xi_v} \ , \qquad \beta_{\xi_v} = n_v - \sum_{i=1}^{n_v} y_{i,v} + b_{\xi_v} \ ,$$

$$\alpha_{\mathrm{se}_{l,v}} = \sum_{i=1}^{n_v} t_{i,l,v}\, y_{i,v} + a_{\mathrm{se}_{l,v}} \ , \qquad \beta_{\mathrm{se}_{l,v}} = \sum_{i=1}^{n_v} (1-t_{i,l,v})\, y_{i,v} + b_{\mathrm{se}_{l,v}} \ ,$$

$$\alpha_{\mathrm{sp}_{l,v}} = \sum_{i=1}^{n_v} (1-t_{i,l,v})\,(1-y_{i,v}) + a_{\mathrm{sp}_{l,v}} \ , \qquad \beta_{\mathrm{sp}_{l,v}} = \sum_{i=1}^{n_v} t_{i,l,v}(1-y_{i,v}) + b_{\mathrm{sp}_{l,v}} \, .$$

3. SIMULATION STUDY

We consider a simulation study to compare the performance of our model with the Hui–Walter model. Following Georgiadis *et al.* (2003), we simulate data according to the following steps.

Step 1. Calculate the probabilities for each combination of outcomes of the L tests under investigation in vth stratum, given the health condition of a subject, $D \in \{0, 1\}$, i.e.

$$\begin{aligned} &P_{v|D=1}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 1\Big) \, , \\ &P_{v|D=0}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{Lv} \mid D = 0\Big) \, . \end{aligned} \tag{3.1}$$

Step 2. Calculate the amount of $X_{v|D}$ elements for each combination of outcomes of the L tests under investigation in vth stratum, given the health condition of a subject, $D \in \{0, 1\}$,

$$\begin{aligned} \mathrm{E}(X_{v|D}) \ = \ n_v \Big\{ &\xi_v \, P_{v|D=1}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 1\Big) \\ &+ (1-\xi_v)\, P_{v|D=0}\Big(T_{1,v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 0\Big)\Big\} \, . \end{aligned} \tag{3.2}$$

For the structure of conditional independence we have conditional probabilities (3.1) given by

$$\begin{aligned} P_{v|D=1}\Big(T_{1v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 1\Big) \ &= \ \prod_{l=1}^{L} \mathrm{se}_{l,v}^{\,t_{i,l,v}} (1-\mathrm{se}_{l,v})^{1-t_{i,l,v}} \, , \\ P_{v|D=0}\Big(T_{1v} = t_{1,v},\, ...,\, T_{L,v} = t_{L,v} \mid D = 0\Big) \ &= \ \prod_{l=1}^{L} \mathrm{sp}_{l,v}^{\,1-t_{i,l,v}} (1-\mathrm{sp}_{l,v})^{t_{i,l,v}} \, . \end{aligned} \tag{3.3}$$

Table 1: Settings under which data were simulated; here ξ denotes prevalence, whereas 'se' and 'sp' denote sensitivity and specificity. Data have been simulated with the following sample sizes: $n = 50, 100, 500, 1000$.

`Configuration (CONF)`	Stratum (v)								
	1			2			3		
	ξ_1	$\text{se}_{l,1}$	$\text{sp}_{l,1}$	ξ_2	$\text{se}_{l,2}$	$\text{sp}_{l,2}$	ξ_3	$\text{se}_{l,3}$	$\text{sp}_{l,3}$
I	0.30	0.93	0.99	0.70	0.99	0.93	0.50	0.95	0.95
II	0.35	0.93	0.99	0.65	0.99	0.93	0.50	0.95	0.95
III	0.40	0.93	0.99	0.60	0.99	0.93	0.50	0.95	0.95

We have compared the performance of two particular cases of our model: MODEL I (Hui–Walter stratification) and MODEL II (Hui–Walter extended stratification).

Table 2: AIC, BIC, and DIC for MODEL I and MODEL II, according to the settings in Table 1.

	`Configuration (CONF)`	n	AIC	BIC	DIC
MODEL I	I	50	1605.4	1626.5	2018.6
		100	3701.5	3727.4	4555.8
		500	24551.4	24588.6	28832.8
		1000	54053.8	54095.9	62572.0
	II	50	1587.0	1608.0	1955.9
		100	2471.8	2497.7	2583.8
		500	23544.8	23581.9	2743.7
		1000	51669.2	51711.3	59633.8
	III	50	1499.6	1520.7	1835.6
		100	3389.0	3414.9	4083.2
		500	21695.6	21732.7	25094.1
		1000	47547.6	4759.6	54538.5
MODEL II	I	50	1065.2	1146.5	2524.1
		100	2565.6	2665.6	6393.4
		500	11676.3	11819.7	15908.6
		1000	27800.0	27962.1	36265.5
	II	50	1094.4	1175.7	1395.8
		100	2578.1	2678.1	3279.6
		500	17765.8	17909.2	21554.1
		1000	40106.1	40268.3	47881.8
	III	50	1169.7	1251.0	1451.0
		100	2660.1	2760.1	3298.8
		500	17623.7	17767.2	21019.9
		1000	39340.7	39502.8	46184.2

Two MCMC parallel chains of 50.000 iterations were generated from posterior conditionals (2.6), discarding the first 5.000 iterations (burn-in) of each chain; after thining, we were left with a posterior sample of size $n = 2.000$. The convergence of posterior conditionals (2.6) to the posterior marginals of $\boldsymbol{\theta}$, was monitored by using the potential scale reduction factor (R) (Gelman and Rubin, 1992), and the posterior marginals were graphically evaluated in terms of symmetry, unimodality, and variability of estimates based on the amplitude of 95% credibility interval and mean standard errors. The AIC, BIC, and DIC criteria were used to evaluate the performance of the models (Iliopoulos *et al.*, 2007), and according to these criteria our model (MODEL II) overperforms MODEL I; see Table 2.

We observe estimates with smaller standard error as we increase the sample size and/or absolute mean difference in disease prevalence rates between the strata, with slightly smaller rates of sensitivity ($\text{se}_{l,v}$) and specificity ($\text{sp}_{l,v}$) being found in more prevalent and less prevalent strata, respectively; see Figure 2.

Despite presenting a slightly larger standard error to that of MODEL I, our model (MODEL II) had stationary marginals and estimates very close to the true ones; in addition, we note that sensitivity and specificity are always overestimated with MODEL I; see Figure 3.

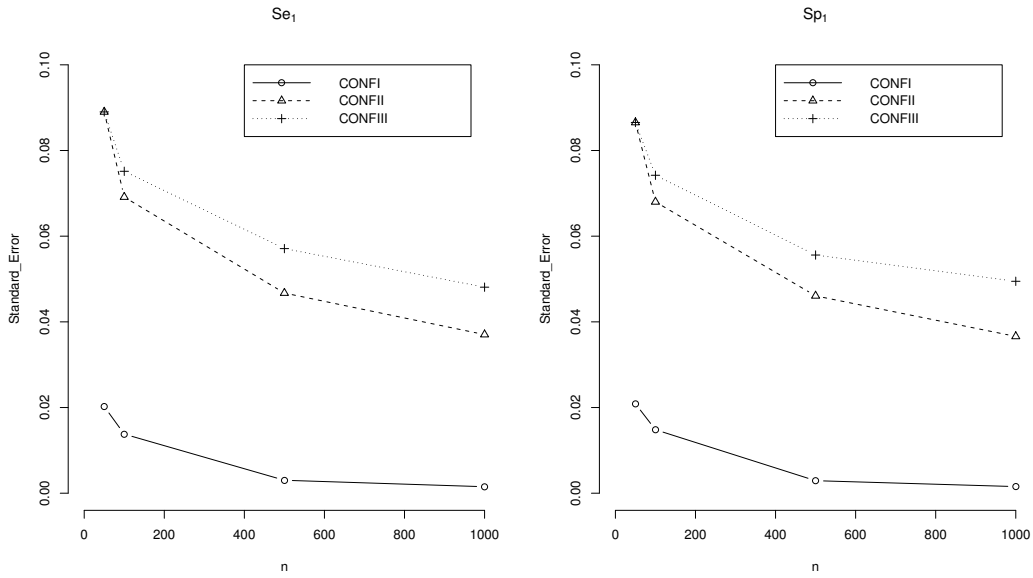


Figure 1: Standard error ($\times 10^{-2}$) to the sensitivities and specificities of the first test in MODEL I according to the settings in Table 1.

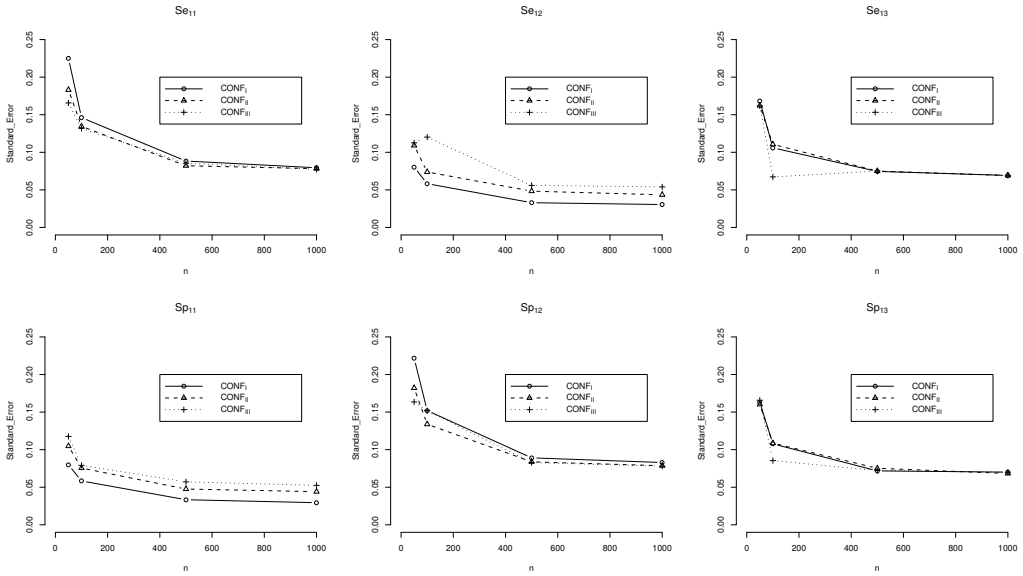


Figure 2: Standard error ($\times 10^{-2}$) to the sensitivities and specificities of the first test in MODEL II according to the settings in Table 1.

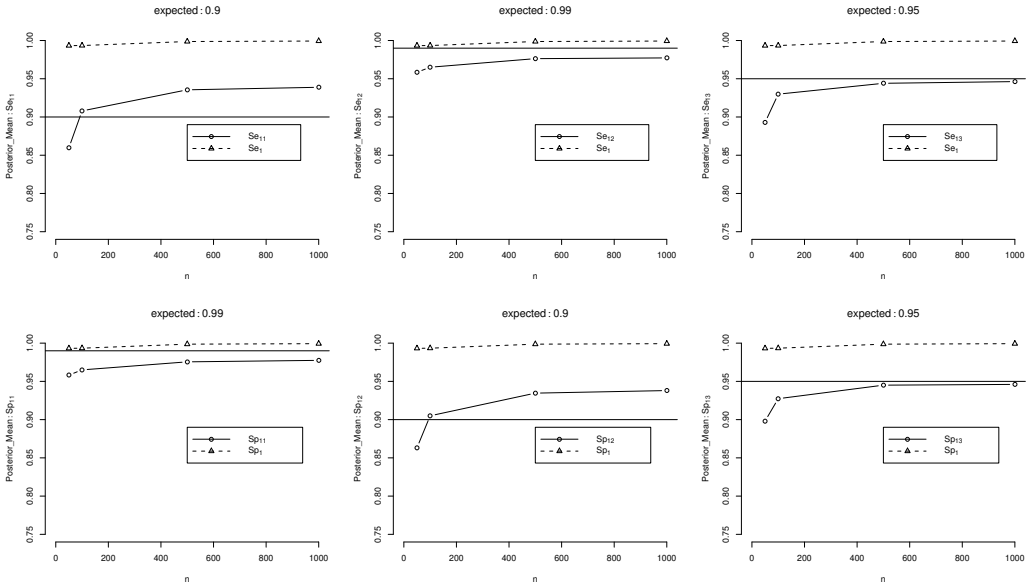


Figure 3: Posterior mean to the sensitivities and specificities of the first test in MODEL I and MODEL II according to the settings in Table 1.

4. ILLUSTRATION ON CHAGAS DISEASE DATA

We now consider an illustration using a Chagas disease case study in Brazil. The data were gathered from 238 blood donors attending a blood center in the region of Triângulo Mineiro, Brazil, who were randomly selected from two groups with different prevalences. Stratum I consists of 29 samples from blood bank donors with positive serology in three conventional serological reactions for Chagas' disease (positive control), and 30 blood samples with five or more negative donations (negative control). Stratum II consists of 179 samples from blood bank donors collected between 2005 and 2008, whose values were low, or within the region denominated 'gray zone' $\pm$ 20% of the reactivity threshold (undetermined serology). Several commercially available kits have been used to determine the diagnostic performance of the four tests, namely: One immunoblotting `TB` (TESA-blot), and three ELISA-based tests, viz.: ELISA Wienner total extract from the subclass `IgG1`, `E-BIO` (ELISA BioMérieux) and `E-WIE` (ELISA Winner recombinant).

Table 3: Results of four serological tests in two subgroups of blood donors.

Test				Group	
IgG1	E-BIO	E-WIE	TB	Control	Inc.
−	−	−	−	30	78
−	−	−	+	0	1
−	−	+	−	0	13
+	−	−	−	0	11
+	−	+	−	0	18
+	−	+	+	0	1
+	+	+	+	29	57
Total:				59	179

`IgG1`: ELISA Wienner total extract from the subclass `IgG1`;
`E-BIO`: ELISA BioMérieux;
`E-WIE`: ELISA Winner recombinant;
`TB`: *Imunoblotting* TESA-blot;
`Control`: negative and positive serology;
`Inc.`: inconclusive in screening serology;
'−': negative result;
'+': positive result.

In Table 4 we report the AIC, BIC, and DIC, for MODEL I and MODEL II; similarly to the simulation study in §3, we observe here that our model overperforms the Hui–Walter model.

Table 4: AIC, BIC, and DIC for MODEL I (Hui–Walter stratification) and MODEL II (Hui–Walter extended stratification); for purposes of presentation each of the entries was multiplied by $\times 10^{-4}$.

MODEL I				MODEL II			
p	DIC	BIC	AIC	p	DIC	BIC	AIC
10	67.4	52.0	52.0	18	36.9	31.4	31.4

In Table 5 we present the estimates obtained from the application of our model by using the group serology strata defined above.

Table 5: Estimates obtained from the application of our model by using group serology strata (Stratum I and Stratum II).

	Test	`Control`			`Inc.`		
		Mean	2.5 Pc	97.5 Pc	Mean	2.5 Pc	97.5 Pc
Sensitivity	`IgG1`	96.94	89.52	99.91	98.39	94.02	99.97
	`E-BIO`	96.88	88.52	99.90	96.64	90.97	99.71
	`E-WIE`	96.73	88.14	99.92	98.30	93.31	99.95
	`TB`	96.66	88.47	99.94	98.14	93.07	99.95
Specificity	`IgG1`	96.92	89.93	99.93	79.21	73.03	85.35
	`E-BIO`	96.92	88.84	99.90	99.34	97.53	99.98
	`E-WIE`	96.92	89.16	99.91	79.19	72.47	85.35
	`TB`	96.95	90.05	99.91	98.63	96.40	99.83
Prevalence		49.30	36.22	61.74	28.19	22.64	34.09

Pc: percentile;
`IgG1`: ELISA Wienner total extract from the subclass `IgG1`;
`E-BIO`: ELISA BioMérieux;
`E-WIE`: ELISA Winner recombinant;
`TB`: *Imunoblotting* TESA-blot;
`Control`: negative and positive serology;
`Inc.`: inconclusive in screening serology.

ACKNOWLEDGMENTS

We thank the Brazilian organizations FAPESP, CAPES and CNPq for financial support.

REFERENCES

Berkvens, D.; Speybroeck, N.; Praet, N.; Adel, A. and Lesaffre, E. (2006). Estimating disease prevalence in a Bayesian framework using probabilistic constraints, *Epidemiology*, **17**, 145–153.

Bertrand, P.; Bénichou, J.; Grenier, P. and Chastang, C. (2005). Hui and Walter's latent-class reference-free approach may be more useful assessing agreement than diagnostic performance, *Journal of Clinical Epidemiology*, **58**, 688–700.

Branscum, A. J.; Gardner, I. A. and Johnson, W. O. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling, *Preventive Veterinary Medicine*, **68**, 145–163.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings Algorithm, *The American Statistician*, **49**, 327–335.

Dempster, A. P.; Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Ser.* B, **39**, 1–38.

Georgiadis, M. P.; Johnson, W. O.; Gardner, I. A. and Singh, R. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests, *Journal of the Royal Statistical Society, Ser.* C, **52**, 63–76.

Gardner, I. A. (2004). An epidemiologic critique of current microbial risk assessment practices: the importance of prevalence and test acuracy data, *Journal of Food Protection*, **67**, 2000–2007.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis, *Biometrics*, **56**, 1055–1067.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of American Statistical Association*, **85**, 398–409.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, **61**, 215–231.

Gustafson, P. (2005). The utility of prior informative and stratification for parameter estimation with two screening tests but no gold standard, *Statistics in Medicine*, **24**, 1203–1217.

Hadgu, A.; Dendukuri, N. and Hielden, J. (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues, *Epidemiology*, **16**, 604–612.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests, *Biometrics*, **36**, 167–171.

Iliopoulos, G.; Kateria, M. and Ntzoufras, I. (2007). Bayesian order-restricted association models for a two-way contingency table, *Computational Statistics and Data Analysis*, **51**, 4643–4655.

Johnson, W. O.; Gastwirth, J. L. and Pearson, L. M. (2001). Screening without a 'gold standard':the Hui–Walter paradigm revisited, *American Journal of Epidemiology*, **153**, 921–924.

Jones, G.; Johnson, W. O.; Hanson, T. E. and Christensen, R. (2009). Identifiability of models for multiple diagnostic testing in the absence of a gold standard, *Biometrics*, **66**, 855–863.

Joseph, L.; Gyorkos, T. W. and Coupal, L. (1995). Bayesian estimation of disease prevalence and parameters for diagnostic tests in the absence of a gold standard, *American Journal of Epidemiology*, **141**, 263–272.

Kraemer, H. C. (1992). *Evaluating Medical Tests*, Lavoisier Librairie, Bervely Hills.

Lindley, D. V. (1971). *Bayesian Statistics: A Review*, Society for Industrial and Applied Mathematics, Philadelphia.

Neath, A. A. and Samaniego, F. J. (1997). On the efficacy of Bayesian inference for nonidentifiable models, *The American Statistician*, **51**, 225–232.

Nielsen, S. S.; Grønbaek, C.; Agger, J. F. and Houe, H. (2002). Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis, *Preventive Veterinary Medicine*, **53**, 191–204.

Singer, R. S.; Boyce, W. M.; Gardner, I. A.; Johnson, W. O. and Fisher, A. S. (1998). Evaluation of bluetongue virus diagnostic tests in free ranging bighorn sheep, *Preventive Veterinary Medicine*, **35**, 265–282.

Swartz, T.; Haitovsky, Y.; Vexler, A. and Yang, T. (2004). Bayesian identifiability and misclassification in multinomial data, *The Canadian Journal of Statistics*, **32**, 1–18.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

Toft, N.; Jørgensen, E. and Højsgaard, S. (2005). Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard, *Preventive Veterinary Medicine*, **68**, 19–33.

Toft, N.; Innocent, G. T.; Gettinby, G. and Reid, S. W. J. (2007). Assessing the convergence of Markov Chain Monte Carlo methods: An example from evaluation of diagnostic tests in absence of a gold standard *Preventive Veterinary Medicine*, **79**, 244–256.

Walter, S. D. and Irwing, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a Review, *Journal of Clinical Epidemiology*, **41**, 923–937.

REVSTAT – Statistical Journal

Background

Statistical Institute of Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT – STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

— The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

— All research articles will be refereed by at least two persons, one from the Editorial Board and another, external.

— The only working language allowed will be English.

— Three volumes are scheduled for publication, one in April, one in June and the other in November.

— On average, four articles will be published per issue.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract / indexed in

REVSTAT is expected to be abstracted / indexed at least in *Current Index to Statistics*, *Statistical Theory and Method Abstracts* and *Zentralblatt für Mathematik*.

Instructions to Authors, special-issue editors and publishers

Papers may be submitted in two different ways:

— By sending a paper copy to the Executive Editor and one copy to one of the two Editors or Associate Editors whose opinion the author(s) would like to be taken into account, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

— By sending a paper copy to the Executive Editor, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Submitted manuscripts (text, tables and figures) should be typed only in black, on one side, in double spacing, with a left margin of at least 3 cm and not have more than 30 pages.

The first page should include the name, affiliation and address of the author(s) and a short abstract with the maximum of 100 words, followed by the key words up to the limit of 6, and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style.

This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to *PC Windows System* (Zip format), *Mackintosh*, *Linux* and *Solaris Systems* (StuffIt format), and *Mackintosh System* (BinHex Format), are available in the REVSTAT link of the National Statistical Institute's Website: http://www.ine.pt/revstat/inicio.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: liliana.martins@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Also send always the final paper version to:

Maria José Carrilho
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística, I.P.
Av. António José de Almeida
1000-043 LISBOA
PORTUGAL

Copyright