Universidade do Minho
Escola de Engenharia
Departamento de Informática

Francisco Luís do Amaral Ribeiro Machado e Costa

RealROC

A Shiny based application for ROC curve study with covariate adjustment

October 2020

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Francisco Luís do Amaral Ribeiro Machado e Costa

RealROC

A Shiny based application for ROC curve study with covariate adjustment

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Ana Cristina Braga

October 2020

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

AGRADECIMENTOS

Se o documento presente for julgado pelo leitor como sucinto e objetivo será por esse mesmo impulso pessoal, de me manter breve e frio, que esta secção se torna a que mais tempo e reflexão pede para a escrever, no que será um eterno conflito, que adivinho partilhar com muitos colegas, do romance da minha língua mãe com a estéril realidade científica.

Requer o estereótipo de aqui catalogar família e amigos para que possa agradecer o apoio contínuo nos últimos anos. Não posso deixar de rejeitar esta exigência. O amor que sinto pelos mesmos é melhor expresso diariamente e oralmente que num documento imutável por arte de escrita que de minha tristeza será sempre inferior à eloquência da fala. Se porventura este gesto parecer radical direi preferir a comparação aos antigos, dado ser (em parte) graças à oralidade que mantemos até hoje histórias bíblicas, filosofia socrática e mitologia assíria.

Libertando-me no entanto dos clichés impostos de agradecer a pais e país, em soberania de alma e em rigor telegráfico expresso um obrigado,

À minha orientadora, Professora Doutora Ana Cristina Braga, por demonstrar um exemplo do humano e do académico que individualmente merecem elogio, mas cuja conjugação exige louvor,

Ao meu avô, Sr. Dr. Juiz José Machado e Costa, cuja eloquência de fala e escrita e procura da verdade tento replicar por diferentes caminhos,

À minha avó, Maria da Conceição Amaral, a quem poderei entregar este documento como prova que as suas inúmeras horas de ensino não foram em vão.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

RESUMO

A curva ROC (*Receiver operating characteristic*) é uma ferramenta analítica eficaz para testes clínicos. A análise permite visualizar a variação de sensibilidade e especificidade para uma dada região de corte através de um simples, mas robusto gráfico bidimensional.

Num contexto biológico, testes podem ser influenciados por múltiplas variáveis externas e como tal a analise ROC pode não ser a ideal ou gerar resultados incompletos. É então necessário saber que variáveis afetam determinado teste clínico de forma a determinar os melhores parâmetros para determinado teste ou até descartar determinada metodologia mediante a situação. O ajuste da curva ROC a covariáveis permite a normalização do efeito das mesmas ou diretamente ajustar a curva para os seus efeitos.

Software direcionado ao ajuste da curva ROC é, infelizmente, escasso e muitas vezes difícil de manusear por utilizadores não especializados. Recentemente o pacote AROC foi lançado para R que disponibiliza vários recursos para estes ajustamentos, no entanto a dificuldade de utilização mantém-se.

A combinação deste pacote com a estrutura *Shiny*, um pacote que permite o desenvolvimento de aplicações interativas, tem por objetivo a criação de um programa grátis e acessível que permita uma analise mais aprofundada disponível para todos os investigadores.

RealROC foi capaz de replicar resultados de um caso de estudo que analisou a influência do sexo no sistema de pontuação CRIB e respetiva previsão de mortalidade, demonstrando a usabilidade e acessibilidade do programa que será disponibilizado online e potencialmente contribuir para novos desenvolvimentos na área.

Palavras-chave: Curva ROC; AROC; Covariáveis; *Shiny*; Bioestatística; Informática Médica; Classificação estatística; Software

ABSTRACT

Receiver operating characteristic (ROC) curves are a powerful analytical tool for clinical tests. The analysis allows the visualization of varying sensitivity and specificity for a given threshold through a simple, yet robust, two-dimensional plot.

In a biological framework, tests can be influenced by multiple external variables, as such, standard ROC analysis may not be suitable or may provide incomplete data. It is then necessary to know which variables influence clinical test results to determine optimal conditions for trials or even to disregard a given method of evaluation in certain contexts. Adjusting for covariates allows ROC analysis to normalize the effects of the variable in question or to directly adjust the curve for its effects.

Unfortunately ROC software that is able to conduct such an adjustment is sparse and proven difficult to use for non technical users. Recently, the AROC package for R was released and provides a robust resource for such adjustments however with he same usability problems previously stated.

By combining this package with the Shiny framework, an R package that allows the creation of interactive applications, we hope to provide an accessible and free software that allows this extra depth of analysis to be available for all researchers.

RealROC was able to mimic the results of a case study analysing the affects of sex to the CRIB score and resulting mortality rates that proving its practicality and will be made available online and hopefully contribute to the advancement of software in this field.

Keywords: ROC curve; AROC; Covariates; Shiny; Biostatistics; Medical Informatics; Statistical classification; Software

CONTENTS

LIST OF FIGURES

ACRONYMS

AUC Area Under the Curve.

c threshold.

CI Confidence Intervals.

CRIB Clinical risk index for babies.

D Diseased.

FN False Negative.

FP False Positive.

FPF False Positive Fraction.

LDV Latent Decision Variable.

MVD Max Vertical Distance.

NICUs Neonatal Intensive Care Units.

NMR Neonatal Mortality Rates.

pAUC Partial Area Under the Curve.

ROC Receiver Operating Characteristic.

SDG Sustainable Development Goals.

se Standard Error.

SNAP Score for Neonatal Acute Physiology.

TN True Negative.

TP True Positive.

TPF True Positive Fraction.

UI User interface.

YI Youden Index.

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

From medical diagnosis describing an individual as sick or healthy, investment strategies dictating secure or otherwise risky investments or simply a spam filter on email services, there are limitless situations where one needs to set objects in belonging to a class rather than another. In fact there are instances where more than two classes may be employed, but in practice the dichotomy between sick/well, accept/reject, yes/no is by far the most popular. Further, as we will see, multi class systems can often be decomposed into two class cases.

Systems of assignment are not perfect, and errors can occur leading to false positive and false negative results, as such systems of evaluation are necessary to assess test performance. The Receiver Operating Characteristic (ROC) curve is a popular tool for such analysis that allows the visualization of varying sensitivity and specificity for a given threshold through a simple, yet robust, two-dimensional plot (x and y corresponding to [1- specificity] and sensitivity, respectively) (Egan, 1975).

A ROC curve illustrates the diagnostic ability of a binary classifier, for which there are four possible outcomes, True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Within a given threshold, these total number of observations allow the sensitivity (True Positive Rate) and specificity (True Negative Rate) to be calculated and built into the curve.

Tests can be influenced by multiple external variables and as such, standard ROC analysis may not be suitable or may provide incomplete data. It is then necessary to know which variables influence clinical test results to determine optimal conditions for trials or even to disregard a given method of evaluation in certain contexts (Rodríguez-Álvarez et al., 2011). Adjusting for covariates allows ROC analysis to normalize the effects of the variable in question or to directly adjust the curve for its effects (Pepe, 1997).

With the goal to provide an accessible, free software for non-technical users, this dissertation aims to develop and publish a web application using Shiny, an R extension that integrates web development into existing R language, that incorporates known adjustment methods to ROC curve analysis with covariates to contribute to resource building in the field, simplifying the process and making the analysis accessible for a non technical user.

1.2 GOALS

The main objective of this dissertation was the development of a user friendly Shiny application that allows building and accurate modelling ROC curves in the presence of covariates with applicable approaches, providing an intuitive system of step by step analysis, visualization of data, curve and summarization.

To achieve this goal theoretical principles of the ROC curve analysis and covariate adjustment as well as currently available software will be assessed and explored so as to develop and publish an appealing and robust software for both user comfort and accuracy of analysis.

1.3 DOCUMENT ORGANIZATION

Chapter 2 sees the proper theoretical concepts behind the ROC curve, it's applications and estimation as well as the current way to integrate covariates in the analysis.

Chapter 3 takes an in depth look into the the current R package for covariate adjusted ROC curves and its methods - AROC, that will serve as building blocks for the Shiny web application.

In Chapter 4 using the Shiny package in conjunction with AROC both backend and frontend development of the application will be explored, providing an example of its use in Chapter 5 concluding with future prospects in Chapter 6.

2

STATE OF THE ART

2.1 ROC CURVE CONCEPT

2.1.1 *ROC curve and applications*

In this section we will explore the theoretical concepts of the ROC curve construction and analysis that were mentioned briefly in the previous chapter as to provide a framework necessary for several statistical application we will examine in the next few sections.

As mentioned, the abstract problem of classifying an object as belonging to two independent classes, is limitless in application and the core concept behind decision statistics. Systems of classification are flawed however, and errors in classification can occur more often than not requiring the existence of performance classifiers such as the ROC curve.

While its name draws from its World War 2 origins for the analysis of radar signals and its operators it has since been applied in psychology, medicine, prominently in radiology (Obuchowski, 2003), epidemiology and diagnostic research, as well as machine learning (Bradley, 1997).

The widespread use of the ROC curve in various fields makes nomenclature a challenge, with notable discrepancies between major modern sources. While the focus of this section is exploring theoretical concepts, this dissertation concerns the application of these concepts in a biological/medical setting and will, therefore, assume these concepts being used in a diagnostics setting rather than mathematical abstraction. In a practical sense this means referring to Diseased (D) and Non-Diseased ($\bar{D}$) populations rather than Positive and Negative resembling notation used in Pepe (2003) while keeping to the definitions of Krazanowski and Hand (2009).

The ROC curve, defined as a plot of True Positive Fraction (TPF), or sensitivity, and False Positive Fraction (FPF), 1- specificity, pairs obtained by varying threshold (c) as (x,y) axis respectively. Defining Y_D and $Y_{\bar{D}}$ as continuous variables for diseased and non diseased groups respectively with cumulative distribution functions F_D and $F_{\bar{D}}$, we assume all test outcomes greater than c belong to the diseased group, with $c \in \mathbb{R}$. Subsequently, each given c will determine TPF,

$$TPF(c) = Pr(Y_D \geq c) = 1 - F_D(c) \tag{1}$$

and similarly FPF,

$$FPF(c) = Pr(Y_{\bar{D}} \geq c) = 1 - F_{\bar{D}}(c). \tag{2}$$

The ROC curve is then defined as all FPF-TPF pairs,

$$ROC(\cdot) = \{(FPF(c), TPF(c)), c \in \mathbb{R}\} \tag{3}$$

(Pepe, 2003). By converting FPF at threshold c to t, such as, $t = FPF(c) = 1 - F_{\bar{D}}(c)$, the ROC curve is defined as $\{(t, ROC(t)) : t \in [0,1]\}$ (Rodríguez-Álvarez et al., 2018), where

$$ROC(t) = P\{Y_D > F_{\bar{D}}^{-1}(1-t)\} = 1 - F_D\{F_{\bar{D}}^{-1}(1-t)\}, \quad 0 \leq t \leq 1 \tag{4}$$

2.1.2 *Summary Indexes*

In section 2.1.1 we saw the ROC curve can be seen as a summary of all possible FPF-TPF pairs describing a classifiers' performance in all possible threshold values, however this information itself can be complicated or otherwise difficult to digest sometimes, be it in communication, replication, or in comparing multiple classifiers. Scalar values are often used for such a task. In this section we will explore a few of the more popular indexes.

The Area Under the Curve (AUC) is by far the most widely used summary index and a household name in ROC curve analysis. AUC comes from a simple geometrical interpretation of the curve that allows one to say that the perfect test, ie the prefect distinction between D and $\bar{D}$ cases, is the upper borders of the graph (area of the square of side 1) while a random selection of results in indicated by the chance diagonal (area of the triangle of base and height 1) giving us the formal definition of AUC as

$$AUC = \int_0^1 y(x)dx \tag{5}$$

or, by using equation 3

$$AUC = \int_0^1 ROC(t)dt. \tag{6}$$

A simple and commonly used interpretation of AUC is as an average value of sensitivity for all possible values of specificity taking values between 0 and 1. An AUC value closest to 1 indicates a test with 100% accuracy bringing its practical lower limit to 0.5 or 50% accuracy which we refer to as the chance diagonal indicating a test relies on luck and is, fundamentally, not suitable (Park et al., 2004).

A more formal interpretation states the AUC is equal to the probability that tests results from a randomly selected pair of diseased and non-diseased subjects are correctly ordered, i.e. $P[Y_D > Y_{\bar{D}}]$.

Often times a particular TPF is of interest, this is particularly the case in medical contexts where the FPF is particularly small (< 0.05), in these cases and for specifying a range of threshold values Partial Area Under the Curve (pAUC) is used. Working with equation 6, we define PAUC as,

$$pAUC(t_0) = \int_0^{t_0} ROC(t)dt. \tag{7}$$

Both AUC and pAUC are used regularly, however a few others deserve mentioning. The Youden Index (YI) is the maximum difference between TP and FP fractions,

$$YI = max(tp - fp) = max(tp + tn - 1), \tag{8}$$

The threshold at the point on the ROC curve corresponding to YI is often taken to be the optimal classification threshold. Another such summary index is the Max Vertical Distance (MVD) between the chance diagonal and the ROC curve, $MVD = max|y(x) - x|$ a particularly useful statistic for its equivalence to both YI and Kolmogorov-Smirnov statistic in the ROC curve domain (Krzanowski and Hand, 2009).

2.1.3 *Binormal Model*

In ROC analysis, the binormal model refers to the assumption of normal distributions of both populations. This is the cornerstone of ROC analysis and a standard by which other specialized analysis can be judged.

Beginning with the aforementioned assumption

$$Y_D \sim N(\mu_D, \sigma^2_{\bar{D}}), \quad Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma^2_{\bar{D}}),$$

then

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)) \tag{9}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function (cdf) and,

$$a = \tfrac{\mu_D - \mu_{\bar{D}}}{\sigma_D}, \quad b = \tfrac{\sigma_{\bar{D}}}{\sigma_D}$$

We see the definition of the binormal ROC curve in equation 9 where we call a the intercept and b the slope for the curve where both values are positive if we abide by convention of larger values being indicative of disease (Pepe, 2003).
Krzanowski and Hand (2009) further expanded this equation form

$$y(t) = \Phi\left(\frac{\mu_D - \mu_{\bar{D}} + \sigma_{\bar{D}} \times z_t}{\sigma_D}\right) \tag{10}$$

where,

$$z_t = \Phi^{-1}[t(c)] = \frac{\mu_{\bar{D}} - c}{\sigma_{\bar{D}}}$$

For the binormal ROC curve the AUC is

$$AUC = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right), \tag{11}$$

an increasing function of a and decreasing of b. Partial AUC is not derivable from the previous expression and must be calculated using numerical integration or rational polynomial approximation.

As mentioned, the ROC curve is invariant to monotone increasing data transformations therefore to say Y_D and $Y_{\bar{D}}$ is to say that to for a strictly increasing transformation h, $h(Y_D)$ and $h(Y_{\bar{D}})$ have normal distributions, we can further impose that this function h transforms the data to normality. The assumption that such a function h makes both populations normally distributed exists is a weak one however empirical testing shows that for non-Gaussian distributions the binormal model still holds (Pepe, 2003). This emphasizes the concept that ROC curves quantify relationships between distributions and have no link to a particular one.

It is this standard model we will see adapted to allow covariate adjustment further ahead.

2.1.4 *ROC curve for Ordinal Tests*

So far, theoretical assessments have implied a numerical, continuous scale of data, however there are many cases where tests are not just discrete variables but are non numerical all together with subjective assessments that different assessors may "grade" differently. This has been an issue since the ROC curve's infancy and is a recurring problem in Radiology where, for instance, different categories (0-5) of assessments are given for the same image by different radiologists (Fig. 1). A common framework to tackle this problem is the latent variable framework.

Say, L is an unobserved latent continuous variable corresponding to the assessor's perception of the image. Much like mathematical thresholds, this assessor has its own, subjective, cut off points used to classify/rate the image. Y corresponds to the reported classifications, we say,

$$Y = y \iff c_{y-1} < L < c_y, \quad y = 1, ..., P$$

where $c_0 = -\infty$ and $c_p = \infty$. The reader classifies the image in the y^{th} category if L falls within the interval corresponding to his implicit definition for the y^{th} category (c_{y-1}, c_y).

To model the ROC curve we use L, the latent decision variable, while not an observable variable we are able to identify $P + 1$ points on the basis of Y and by interpolation the ROC curve for L. Since,

$$Y \geq y \iff L > c_{y-1}$$

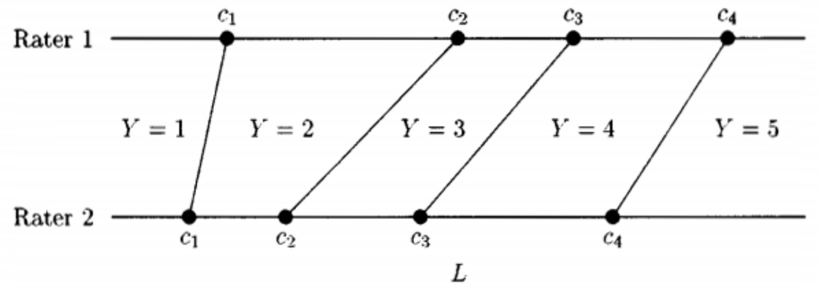


Figure 1: Example of rater bias in ordinal tests where rater 2 shows a more conservative threshold than rater 1 for c (Pepe, 2003).

We can identify TPF and FPF corresponding to the threshold c_{y-1} as $P[Y \geq y|D = 1]$ and $P[Y \geq y|D = 0]$ respectively (Pepe, 2003). If we are able to identify two non-degenerate points t_1 and t_2 we can apply the binormal model where a and b are given by

$$a = \Phi^{-1}(ROC(t_1)) - b\Phi^{-1}(t_1)$$

and

$$b = \frac{\Phi^{-1}(ROC(t_2) - \Phi^{-1}(ROC(t_1))}{\Phi^{-1}(t_2) - \Phi^{-1}(t_1)}.$$

We will see this framework mentioned in the next subsection.

2.2 ROC CURVE ESTIMATION

2.2.1 *Empirical Estimation*

Having explored the theoretical fundamentals of the ROC curve and its summary indexes we now turn to statistical methodology for inferring this curve from existing data rather than assuming the existence of sets of populations and classifiers that fit the mentioned criteria.

Three distinct approaches can be considered for this estimation:

1. Apply non parametric empirical methods to the data to obtain the empirical ROC curve, from which empirical summary indexes can be calculated;

2. Use statistical models for the distributions of cases and controls, parameters in these distributions are estimated and both induced ROC curve and summary indexes are calculated together;

3. Modeling the ROC curve, rather than the probability distributions, as a smooth parametric function.

All three approaches have their respective strengths and drawbacks however in this subsection we will be focusing on the empirical methods followed by the modeling option in subsection 2.2.2 for being the most popular when working with continuous and ordinal data respectively and for mirroring methods of covariate adjustment we will see in the following sections.

Of course when dealing with statistical estimation, one must take into account the accuracy and precision of these estimates, as well as estimator biases and sampling variability to construct reliable Confidence Intervals (CI), this will be explored along with the estimation procedure for both ROC curve and AUC.

Empirical estimation applies the ROC curve definition to the observed data thus empirical TPF and FPF are calculated as

$$\widehat{TPF}(c) = \sum_{i=1}^{n_D} I[Y_{D_i} \geq c]/n_D \tag{12}$$

$$\widehat{FPF}(c) = \sum_{j=1}^{n_{\bar{D}}} I[Y_{\bar{D}_j} \geq c]/n_{\bar{D}} \tag{13}$$

The empirical ROC curve, $\widehat{ROC}_e$ is a plot of $\widehat{TPF}(c)$ versus $\widehat{FPF}(c)$ for all $c \in [-\infty, \infty]$ writable as,

$$\widehat{ROC}_e(t) = \hat{S}_D(\hat{S}_{\bar{D}}^{-1}(t)), \tag{14}$$

where $\hat{S}_D$ and $\hat{S}_{\bar{D}}$ are the empirical survivor functions of Y_D and $Y_{\bar{D}}$ respectively. An example of the resulting curve can be seen in Figure 2.

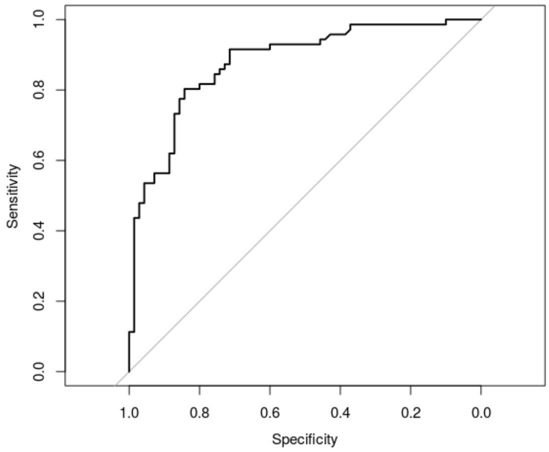


Figure 2: Example of an empirical ROC curve.

Sampling Variability and CIs

Several approaches are available for sampling variability in the empirical ROC curve ranging from fixing threshold to fixing FP and TP fractions as well as considering the entire curve with confidence bands. For our purposes we will consider the first and latter options.

Fixing the threshold, c, and calculating a joint confidence region using exact binomial or asymptotic methods or, alternatively, with bootstrap resampling if the samples are dependent, this method is particularly useful with knowledge of the threshold in advance, such as in blood tests where standard thresholds are already defined.

Considering the entire curve using a confidence band is particularly applicable when no assumptions can be made regarding the test and is more useful for describing the curve. One approach to calculate these bands is to base the calculation on the distribution of $sup|ROC_e(t) - ROC(t)|$, which can be calculated using two independent Brownian bridges or, alternatively, model the risk function $P[D = 1|Y]$ with logistic regression methods (Pepe, 2003).

Empirical AUC

As mentioned empirical indexes can be applied directly to the empirical curve thus,

$$\widehat{AUC}_e = \int_0^1 \widehat{ROC}_e(t)dt \tag{15}$$

$$p\widehat{AUC}_e(t_0) = \int_0^{t_0} \widehat{ROC}_e(t)dt.$$

The area under the empirical ROC curve is the Mann-Whitney U-statistic (Pepe, 2003).

Variability of AUC_e

Calculating variability in AUC_e or any other summary index is often complicated, involving analytic expressions for asymptotic variance however in practice one simply uses bootstrap methods to calculate CIs.

2.2.2 *Modeling the ROC curve*

In the beginning of this section we described using a smooth parametric function to model the ROC curve as a popular option when working with ordinal data, however with a few adjustments we can also work with continuous data and we will describe both methods in the remainder of this chapter.

Ordinal Tests

For this approach we adopt the Latent Decision Variable (LDV) framework mentioned in section 2.1.4. Let L denote the underlying decision variable for a single reader, recall the binormal model of the ROC curve is $ROC(t) = \Phi(a + b\Phi^{-1}(t))$ as seen in equation 9, we assume $L \sim N(0, 1)$ in the non diseased population, $\bar{D}$. With these conditions we imply that in the diseased population D, $L \sim N(a/b, (1, b)^2$, and calculate a and b using the already established relation of $Y = y \iff c_{y-1} < L < c_y$. We can then derive the probability of diseased and non diseased observation Y_D and $Y_{\bar{D}}$

$$P[Y_{\widehat{D}} = y] = \Phi(bc_y - a) - \Phi(bc_{y-1} - a). \tag{16}$$

$$P[Y_{\widehat{\bar{D}}} = y] = \Phi(c_y) - \Phi(c_{y-1}) \tag{17}$$

The log likelihood function is then constructed as

$$\sum_{i=1}^{n_D} logP[Y_D = Y_{Di}] + \sum_{j=1}^{n_{\bar{D}}} logP[Y_{\bar{D}} = Y_{\bar{D}j}] \tag{18}$$

and maximize with respect to parameters $\{a, b, c_1....c_{P-1}\}$. The resulting ROC curve is a smooth curve that follows the binormal model replacing a and b with it's estimators, $\widehat{a}$ and $\widehat{b}$. The Standard Error (se) for $\widehat{a} + \widehat{b}\Phi^{-1}(t)$ is

$$se = \left\{ var(\widehat{a}) + (\Phi^{-1}(t))^2 var(\widehat{b}) + 2\Phi^{-1}(t)cov(\widehat{a}, \widehat{b}) \right\}^{1/2}. \tag{19}$$

Corresponding confidence limits for $a + b\Phi^{-1}(t)$ written as

$$\widehat{a} + \widehat{b}\Phi^{-1} \pm \Phi^{-1}(1-\alpha)se, \tag{20}$$

generate the confidence limits for the binormal ROC curve.

Continuous Tests

Metz et al. (1998) proposed dealing with continuous data by categorizing them into a finite number of pre-defined categories and applying the previously explained fitting methods for ordinal data. The asymptotic properties of the estimator require the categories to be pre-defined, however (Metz et al., 1998) proposed defining theses categories using observable data using vertical and horizontal jumps in the empirical ROC to define the categories. Having defined the categories we apply the LDV framework.

This method is called LABROC and it's most appealing feature is, by working with ranks, being distribution free, making it invariant to monotone increasing data transformations mirroring the ROC curve (Pepe, 2003; Metz et al., 1998).

2.3 BAYESIAN METHODS

2.3.1 *Bayesian Approach and General ROC Analysis*

In the previous sections we discussed what is known as the frequentist or classical framework for statistical inference. In this framework populations are represented by probability models whose parameters are treated as fixed but unknown quantities about which inferences are to be made and, thus has no scope for extraneous information integration such as previous experiments or subjective elements.

The Bayesian approach treats population parameters as random variables with probability distributions reflecting the degree of belief the research has on the data, allowing the introduction and combination of prior knowledge and subjectivity to it. These prior distributions are combined with sample data to produce posterior distributions creating a basis for all inferences.

Bayesian methods have been available for many years however they were hampered by their intractability, general unattractiveness by statistics and bottlenecks in computer processing, this however changed by the mid 90's with the introduction of Markov-chain Monte Carlo methods (Krzanowski and Hand, 2009).

This method generates sample values from the posterior distributions and approximates integrals by the average of sample values of a function $f(\cdot)$, this features ensures each new proposal depends on the current one, and the sequence of proposals is guaranteed to converge to values from the desired distribution. Noticeable algorithms include Metropolis-Hastings, which uses joint distribution of new and current proposals and Gibbs sampling which uses a sequence of conditional distributions. We will see the latter mentioned in the next chapter.

Bayesian methodology shines when studies have uncertainty about underlying quantities and ROC analysis has benefited greatly for this addition to the statistical repertoire. A notorious case of this uncertainty is in labeling as D or $\bar{D}$ of subjects from which a ROC curve is to be constructed is either fallible or not available, this is often the case in medical studies when the **true** disease status of each sample member is either equivocal or unknown and ways.

This section will explore the fundamentals of Bayesian ROC analysis.

For continuous data very few proposals for a standard method exist, Erkanli et al. (2006) suggests the binormal model, presented in section 2.1.3, is a poor model for analysis in Bayesian methodology for its pretense of normal distributed populations using a mixture of normals as a more flexible alternative.

Ignoring for the time being population distinctions (D and $\bar{D}$), say C denotes the number of components in the normal mixture and K is a random variable which indicates the operative component for a classification score, Y. The formal Bayesian approach is then,

$$Y|K, \theta_K \sim N(\mu_K, \sigma_K^2) \tag{21}$$

where K and $\theta_k = (\mu_K, \sigma_K^2)$ are parameters and must be assigned prior distributions. The components of θ_K are assigned normal gamma baseline priors while K has an independent C-state multinomial distribution with probabilities $w_1, ..., w_C$ specified as,

$$w_1 = R_1, \quad w_k = (1 - R_1)(1 - R_2)...(1 - R_{k-1})R_k \quad for \quad k = 2, ..., C$$

where R_i are independent Beta(1,α) variables and $R_C = 1$ to ensure that the w_i sum to 1. This model is termed a mixed Dirichlet process.

With this model we can generate predicted values for Y given previous scores a density function, however the resulting integral is difficult to compute even with a small C, as such a Gibbs sampler is used to simulate observations and the expected value can be approximated by the average of these samples.

Having approximated the posterior predictive density we then obtain cumulative distribution function from which the posterior predicted true positive and false positive rate are calculated, from there, varying the threshold yields the predictive ROC curve.

Krzanowski and Hand (2009) dedicate a chapter to this model and its applications where this section was based on, for the sake of brevity some equations and proofs were excluded however this is a fundamental companion source of information to the original material of Erkanli et al. (2006) and should be consulted for further information.

2.3.2 *Parametric and Non-parametric Bayesian Methods*

It is essential to ensure the correct labeling of samples to populations D and $\bar{D}$ to conduct ROC analysis. We mentioned previously how, in medical tests, labeling is often either fallible or missing, this as lead to the adoption of the gold standard, a method of labeling that always delivers the correct result. This however is an imperfect solution because gold standards hardly exist for most diseases and even those credited as such have been shown to be quite fallible.

Bayesian methodology as made great contributions for this problem, including an extension of the previously mentioned mixed Dirichlet process by the same author (Erkanli et al., 2006) as well as parametric and non parametric estimations proposed by Choi et al. (2006) and Wang et al. (2007) in parametric and non parametric estimation respectively.

The general framework, described by Erkanli et al. (2006), for this approach describes a "group label" variable, gL, in the presence of a gold standard this is deterministic and assigns samples to populations D and $\bar{D}$ without error, however if no gold standard exists is a binary random variable with unknown (but defined) probability of assignment. If the test is not dichotomous one can derive gL by dichotomizing a second continuous classifier U. Deriving gL can also be achieved by thresholding Y, both situations derive Y from L and are handled by the method similarly however for simplicity we will denote the two continuous classifiers as Y_1 and Y_2.

Extending the framework mentioned in 2.3.1 we assume that the true group labels are given by a latent binary variable Z that can be related to gL in two ways,

$$gL|Z \sim Bernoulli(\pi) \tag{22}$$

where,

$$\begin{cases} log(\frac{\pi}{1-\pi}) = \beta_0, & Z = \bar{D} \\ log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1, & Z = D, \end{cases}$$

referred to as the classical model or,

$$Z|gL \sim Bernoulli(\zeta) \tag{23}$$

where, similarly,

$$\begin{cases} log(\frac{\zeta}{1-\zeta}) = \beta_0, & gL = \bar{D} \\ log(\frac{\zeta}{1-\zeta}) = \beta_0 + \beta_1, & gL = D, \end{cases}$$

known as the Berkson model. For full specification the former model also needs $Z \sim Bernoulli(\zeta)$ where $\zeta \sim Beta(a, b)$. The parameters β_0, β_1 can be fixed or assigned prior distributions to reflect uncertainty about the relationship between L and Z. The latent variable Z can also be simulated from its conditional posterior distribution as an additional step to the Gibbs sampler. The process proceeds as previously explained (Erkanli et al., 2006; Krzanowski and Hand, 2009).

Parametric and non parametric features to this method have been established and will be discussed in the remainder of the section.

Parametric estimation

For this method we first acknowledge the existence of a gold standard to correctly label individuals to either populations. Let Y_{1iD} and Y_{2iD} be the scores observed by two classifiers for the ith individual of a random sample of size m from population D and $Y_{1j\bar{D}}$,$Y_{2j\bar{D}}$ the scores from the same classifiers for the jth individual from a sample of size n from population $\bar{D}$. These two scores can be placed in a column vector $\mathbf{Y}_{iD} = (Y_{1iD}, Y_{2iD})^T$ and $\mathbf{Y}_{j\bar{D}} = (Y_{1j\bar{D}}, Y_{2j\bar{D}})^T$. The analysis will begin by using the binormal model discussed in section 2.1.3 with these two vectors so that,

$$\mathbf{Y}_{iD} \sim N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D), \quad \mathbf{Y}_{j\bar{D}} \sim N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}}).$$

Given the existence of a gold standard, all the usual features of ROC analysis can again be obtained. Since there are two classifiers there will be two possible ROC curves with their respective summary indexes that can be calculated with similar adjustments to the binormal model.

To move from the presence to the absence of a gold standard we, again, define Z as the latent variable so that if Z were observable,

$$\begin{cases} Z_j = 1 & j \in D \\ Z_j = 0 & j \in \bar{D}, \end{cases}$$

and let m be the number of individuals observed. Now we assume $Z_j \sim Bernoulli(\pi)$ meaning, $Pr(Z_j = 1) = 1 - Pr(Z_j = 0) = \pi$, making the probability density for the bivariate classification score of the jth individual $p(\cdot)^{Z_j} g(\cdot)^{1-Z_j}$ where $p(\cdot), g(\cdot)$ are $N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ and $N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$ densities respectively. The change adds only the need to estimate the prevalence of each population and an additional parameter π, which needs a prior distribution for which Choi et al. (2006) suggest a Beta(1,1) or a Beta(0.5,0.5) (Krzanowski and Hand, 2009).

Nonparametric estimation

For the nonparametric estimation Wang et al. (2007) assumes a K range of thresholds c, $c_1 < c_2 < ... < c_k$ and that Y has been applied to a sample of data to each threshold. If $(\alpha^{(i)}, \beta^{(i)})$ are the true (unknown) value of $(fp, 1 - tp)$ values at c_i for $i = 1, ..., K$, given the sorted thresholds, $\alpha^{(1)} \geq \alpha^{(2)} \geq ... \geq \alpha^{(K)}$ and similarly, $\beta^{(1)} \leq \beta^{(2)} \leq ... \leq \beta^{(K)}$ and defines the boundaries values as,

$$\alpha^{(K+1)} = \beta^{(0)} = 0, \quad \alpha^{(0)} = \beta^{(K+1)} = 0$$

and given the fallibility of L, the **true** fp and fn fractions are,

$$\alpha = P(D|\bar{D}), \quad \beta = P(\bar{D}|D)$$

respectively. Finally, they define θ be the true prevalence of population D so that θ is the probability that any individual in the group being sampled actually belongs to D (and $1 - \theta$ the probability that it belongs to $\bar{D}$) completing the unknown parameters of the problem.

Turning to the data, let n represent the individuals of the sample and n_D and $n_{\bar{D}}$ the classification of L to the populations D and $\bar{D}$ respectively.

Considering Y at each of the K thresholds, $(x_{iD}, x_{i\bar{D}})$ is the number of individuals labeled D, $\bar{D}$ by L respectively, classified as D at threshold c_{i-1} but $\bar{D}$ at threshold c_i for $i = 2, ..., K - 1$. By deriving the probability of each occurrence in terms of the model parameters, the likelihood of the sample can be shown to have the ordered multinomial form. However this model is unidentifiable for lack of degrees of freedom, Wang et al. (2007) overcomes this by assuming different prevalences $\theta_1, \theta_2, ..., \theta_G$ of population D but that the behaviour of Y is the same for each group so that if there are G groups and independent samples are obtained in each of them then the likelihood of the full set of data is the product of all the above expressions evaluated for each separate group making the model identifiable for $G \geq 2$.

Finally one must guarantee the monotocity of the curve by contraining all $\alpha^{(\cdot)}, \beta^{(\cdot)}$ to be positive by taking a Bayesian approach with Dirichlet priors for these parameters (Krzanowski and Hand, 2009).

2.4 ROC CURVE AND COVARIATES

2.4.1 *The Need to Adjust for Covariates*

Consider that to the previously mentioned continuous marker Y and binary outcome, is added a covariate X that affects the distribution of the marker. The traditional ROC curve combines all case observations and all control observations together regardless of covariate value. If this curve, describing the ability to of discrimination of a marker between cases and controls includes a discriminatory accuracy due to a covariate X, it is then biased. In Figure 3 scenario 1 we can see a binary covariate associated with both marker and outcome, we see that the classification accuracy for the marker is the same in the two centers, implying X is not an effect modifier resulting in an overly optimistic ROC curve when compared to the covariate adjusted curve due to failure in adjusting for the covariate.

The need to adjust in a broader sense, comes from the need to calibrate an evaluating marker when such marker's observations depend on a covariate (Janes and Pepe, 2008) and such calibration is obtained with the covariate-adjusted ROC curve, more recently called the $\mathcal{A}$ROC. Figure 3 shows that when a covariate affects the distribution of marker values among controls, covariate adjustment is necessary to appropriately compare case and control marker distributions (Pepe, 2003; Janes and Pepe, 2008) .

2.4.2 *Adjusting the ROC curve*

Incorporating covariates to the analysis follow two distinct methods, **indirect adjustment**, sometimes referred as induced, where the effects of the covariate are modeled in both diseased and non diseased populations and only afterwards is the ROC curve derived, and **direct adjustment** where the effect is modeled directly on the ROC curve itself (Krzanowski and Hand, 2009).

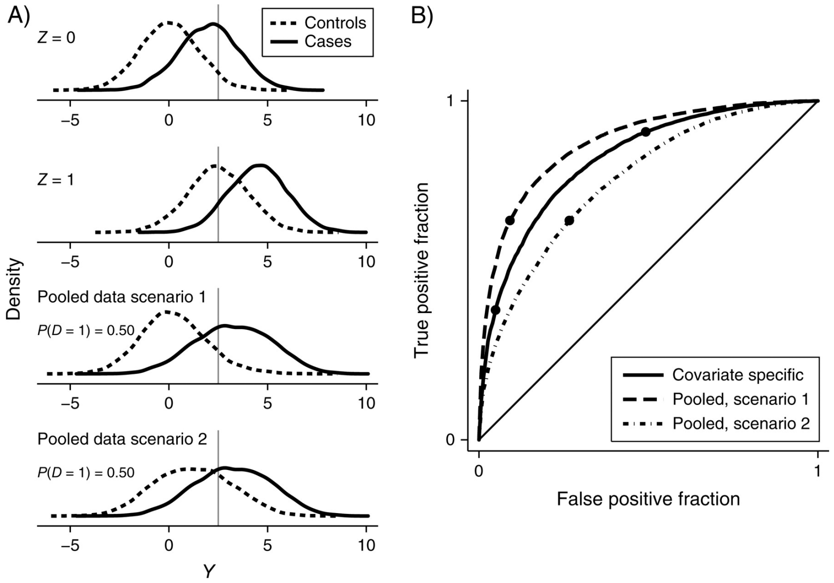


Figure 3: Simulated marker, Y , and binary covariate, $X = 0, 1$. In scenario 1, X is associated with the outcome: $P(D = 1|X = 0) = 0.36$ and $P(D = 1|X = 1) = 0.83$. In scenario 2, X is independent of the outcome: $P(D = 1|X = 0) = P(D = 1|X = 1) = 0.50$. **A)** Shows the densities of Y conditional on X = 0, and X = 1, followed by the pooled data under scenario 1, and scenario 2. A common threshold is indicated. **B)** Shows the common covariate specific ROC curve, the pooled ROC curve under scenario 1, and the pooled ROC curve under scenario 2. The performances of the common threshold rule are indicated (Janes and Pepe, 2008).

Indirect Method

Suppose there is a set of covariates $\boldsymbol{X}_D$ associated with population D and a set of covariates $\boldsymbol{X}_{\bar{D}}$ associated with population $\bar{D}$, in practical applications most, if not all, of these covariates will be common to both populations however this is not a requirement. Let α_D and $\alpha_{\bar{D}}$ be two scalars and $\boldsymbol{\beta}_D$ $\boldsymbol{\beta}_{\bar{D}}$ be element vectors that contain $\boldsymbol{X}_D$ and $\boldsymbol{X}_{\bar{D}}$. We are then able to model the mean of results for both populations ($\mu_D(\boldsymbol{X}_D)$, $\mu_{\bar{D}}(\boldsymbol{X}_{\bar{D}})$) for given values of covariates:

$$\mu_D(\boldsymbol{X}_D) = \alpha_D + \boldsymbol{\beta}_D^T \boldsymbol{X}_D \tag{24}$$

$$\mu_{\bar{D}}(\boldsymbol{X}_{\bar{D}}) = \alpha_{\bar{D}} + \boldsymbol{\beta}_{\bar{D}}^T \boldsymbol{X}_{\bar{D}} \tag{25}$$

And assuming normal distributions for D and $\bar{D}$ and standard deviations σ_D, $\sigma_{\bar{D}}$ respectively, we obtain the equation of the ROC curve as:

$$y = \Phi\left(\frac{\mu_D(\boldsymbol{X}_D) - \mu_{\bar{D}}(\boldsymbol{X}_{\bar{D}}) + \sigma_{\bar{D}} \times \Phi^{-1}(t)}{\sigma_D}\right) \quad (0 \leq t \leq 1) \tag{26}$$

Least-squares regression can estimate point values of α_D, $\alpha_{\bar{D}}$, $\boldsymbol{\beta}_D$, $\boldsymbol{\beta}_{\bar{D}}$, and substitution of these estimates into the above formula at given values $\boldsymbol{x}_D$, $\boldsymbol{x}_{\bar{D}}$ of $\boldsymbol{X}_D$, and $\boldsymbol{X}_{\bar{D}}$ grants us the covariate-specific ROC curves (Krzanowski and Hand, 2009). This model has been criticized over the years by a number of authors (Smith and Thompson, 1996; Faraggi, 2003), suffering changes on the regression approach and replacing the assumption of normality with a common scale parameters, this method however is quite restrictive due to inequalities in population scale and that this generalization could only be done at the cost of considerable mathematical complications. Pepe (1998) took a continuous-score analogy ordinal-score approach, and modeled D and $\bar{D}$, as coming from an arbitrary location-scale family but with the already specified means and standard deviation. If the distribution function of the chosen member of the location-scale family is $H(\cdot)$ on a standardized scale then the ROC curve equation is given by

$$y = 1 - H\left((\alpha_D - \alpha_{\bar{D}}) + \frac{\sigma_D}{\sigma_{\bar{D}}} H^{-1}(1-x) + \boldsymbol{c}^T \boldsymbol{X}\right) \quad (0 \leq x \leq 1) \tag{27}$$

where $\boldsymbol{c} = \frac{1}{\sigma_{\bar{D}}}(\boldsymbol{\beta}_D - \boldsymbol{\beta}_{\bar{D}})$.

To obtain the $\mathcal{A}$ROC in a practical application, these scalars, element vectors and standard deviations must first be estimated without a need to specify $H(\cdot)$ using quasi-likelihood methods (Pepe, 1998; Krzanowski and Hand, 2009).

Direct Method

While in the previous approach we modeled covariate effect on two separate populations, the direct method calculates covariate effect directly on the ROC curve. The main advantage of this methodology is that it allows a direct interpretation of the covariate effect on the ROC curve. In its core this approach requires the selection of a model that will capture the effects of the ROC curve whilst maintaining a flexibility to preserve the monotonically increasing ROC curve, its domain and range.

The need for this flexibility and constraints has popularized the use of the generalized linear models (GLM) (McCullagh and Nelder, 1989), specifically the ROC - GLM model given by,

$$h(y) = b(x) + \boldsymbol{\beta}^T \boldsymbol{X} \tag{28}$$

where,

- b(·) is an unknown baseline function monotonic on (0, 1);
- h(·) is the link function, specified as part of the model and also monotonic on (0, 1);
- $\boldsymbol{X}$ is the vector of covariates (however this time it's a single vector rather than associated to a specific population);
- $\boldsymbol{\beta}$ are the regression parameters associated with the covariates.

Link functions, h(·), vary but the more common in GLM are

- Probit: $h(y) = \Phi^{-1}(y)$
- Logistic: $h(y) = log\frac{y}{1-y} = logit(y)$
- Logarithmic: $h(y) = log(y)$

The same methods of obtaining the mean of both populations can be as applied in equations 24 and 25 (Pepe, 2003) by replacing the individual covariates vectors, $\boldsymbol{X}_D$ and $\boldsymbol{X}_{\bar{D}}$ by $\boldsymbol{X}$ as well as and respective σ_D, $\sigma_{\bar{D}}$ standard deviations which speaks to the flexibility of the approach (Krzanowski and Hand, 2009).

2.4.3 *AROC and Summary Indexes*

In some cases, covariate adjustment interest lies on the summary statistics rather than the ROC curve itself where most of the theoretical attention is focused on the AUC.

If the ROC curve was adjusted indirectly then the effects on AUC are obtained by simply expressing AUC in terms of the model parameters and substituting estimates of the parameters in the binormal expression for x_D and $\mathrm{x}_{\bar{D}}$ for covariates $\boldsymbol{X}_D$ and $\boldsymbol{X}_{\bar{D}}$ respectively:

$$AUC(\mathrm{x}_D, \mathrm{x}_{\bar{D}}) = \phi[\delta(\mathrm{x}_D, \mathrm{x}_{\bar{D}})], \tag{29}$$

where

$$\delta(\mathrm{x}_D, \mathrm{x}_{\bar{D}}) = \frac{\mu_D(\mathrm{x}_D) - \mu_{\bar{D}}(\mathrm{x}_{\bar{D}})}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}}$$

for

$$\mu_D(\mathrm{x}_D) = \alpha_D + \boldsymbol{\beta}_D^T \mathrm{x}_D$$
$$\mu_{\bar{D}}(\mathrm{x}_{\bar{D}}) = \alpha_{\bar{D}} + \boldsymbol{\beta}_{\bar{D}}^T \mathrm{x}_{\bar{D}}.$$

Estimation of these parameters by ordinary least squares and substituting into these expressions yields $\widehat{\delta}(\mathrm{x}_D, \mathrm{x}_{\bar{D}})$ and $\widehat{AUC}(\mathrm{x}_D, \mathrm{x}_{\bar{D}})$. This modulation will generally require strong or specific assumptions. This issue as well as specific cases where AUC values were available but not classification scores, prompted several authors to consider direct adjustment AUC.

Dodd and Pepe (2003) adapted an early proposed model of $\mathcal{A}$AUC calculation that would allow all types of covariates to be modeled with a binary estimation method. Once more, suppose there is a set of c_D covariates $\boldsymbol{X}_D$ associated with population D and a corresponding set $c_{\bar{D}}$ of $\boldsymbol{X}_{\bar{D}}$ for population $\bar{D}$ and we write $\boldsymbol{X}$ as the aggregation of all covariates. We have the AUC regression model,

$$E(h[\widehat{AUC}]) = \alpha + \boldsymbol{\beta}^T \mathrm{x} \tag{30}$$

for parameters α, β and monotone increasing link funtion $h(\cdot)$, with probit and logit again as natural link functions (Krzanowski and Hand, 2009).

3

SOFTWARE AND PACKAGES

3.1 INTRODUCTION

A large amount of software has been developed for ROC curve building and analysis across multiple platforms, enumerated in detail by a recent review by Obuchowski and Bullen (2018), some notable references include ″Metz ROC Software″, developed by the University of Chicago, ″Pepe Lab″ for the Stata platform, ″Analysing ROC curves with SAS″, developed by Mithat Gönen for SAS and the ″pROC″ package for R. These packages while popular either offer a somewhat shallow approach to covariate adjustment or are published in a non readily available platform that can prove difficult to use for non technical users.

To fulfill our goal of creating a user friendly application and bridging the gap between theoretical and practical application of the AROC in a widely available space the Shiny framework was selected. Shiny uses the R language to create an interactable application that can either be used on a local computer or online. However a framework in R requires packages to perform ROC adjustment.

Recently, a package published by Rodríguez-Álvarez et al. (2018) on CRAN, the R Archive Network, goes to great lengths to implement several regression approaches for the inclusion of covariate information on the traditional ROC framework, however, it still suffers from a previously mentioned problem of difficulty in usability.

By combining an interactable framework with this robust package for ROC curve adjustment the feasibility of our goal becomes clearer. Additionally, to provide the user an extra option of comparison between ROC curves to pre and post adjustment the Comp2ROC package was employed.

In this chapter we will take an indepth look at the AROC Package by Rodriguez-Alvarez and Inacio de Carvalho (2018), the Shiny framework as well as a quick overview of the Comp2ROC package by Braga et al. (2016).

3.2 THE AROC R PACKAGE

3.2.1 *Package Overview*

The AROC R-package developed by Rodríguez-Álvarez et al. (2018) implements different methods of computing covariate information in ROC curve construction and two additional methods of calculating marginal/polled ROC curve providing aditional tools for an underdeveloped sub-field. The methods include Bayesian, Kernel and Frequentist methods of adjustment and direct and indirect methods of regression, mentioned in the previous chapter, we will explore in the following subsections.

3.2.2 *Bayesian Methods*

The first methods presented by the package are two bayesian based estimations based on the Dirichlet process discussed in section 2.3.

The *AROC.bnp* function estimates the AROC curve using the Bayesian nonparametric method. Working with equation 4 we say,

$$\begin{aligned} AROC(t) &= Pr\{Y_D > F_{\bar{D}}^{-1}(1 - t|\mathbf{X}_D)\} \\ &= Pr\{1 - F_{\bar{D}}(Y_D|\mathbf{X}_D)) \leq t\} \\ &= Pr(U_D \leq t), \quad 0 \leq t \leq 1, \end{aligned} \tag{31}$$

where $U_D = 1 - F_{\bar{D}}(Y_D|\mathbf{X}_D)$, a placement value of the test outcome in the diseased population or the the standardization of Y_D to the conditional distribution of $Y_{\bar{D}}$ making the AROC a cumulative distribution function of U_D. This method first models the conditional distribution of test outcomes in the nondiseased group, $F_{\bar{D}}$ using a B-splines dependent Dirichlet process mixture of normals model followed by modeling U_D and it's cumulative distribution using a non parametric regression model through Bayesian bootstrap (Rodríguez-Álvarez et al., 2018).

The *AROC.bsp* method is very similar to the previous, non parametric method, in both construction and theory, where this models $F_{\bar{D}}$ using a normal linear regression, making it a counterpart to Janes and Pepe (2009) frequentist model and its AROC package method - AROC.sp (Rodriguez-Alvarez and Inacio de Carvalho, 2018).

3.2.3 *Kernel Methods*

The *AROC.kernel*, an earlier proposed model by Rodriguez-Alvarez *et al* (Rodríguez-Álvarez et al., 2011) is, as the name implies, a kernel based method. Test outcomes for the non diseased group are modeled with a location-scale regression model where both regression and variance functions are estimated using Nadaraa-Watson local estimators that in turn are used to compute standardised residuals to model U_D and estimate the AROC curve, $\widehat{AROC}(t)$ (Rodríguez-Álvarez et al., 2018; Rodriguez-Alvarez and Inacio de Carvalho, 2018; Rodríguez-Álvarez et al., 2011).

The package author notes that, for now, this method, unlike the previous, can only handle a single continuous covariate.

3.2.4 *Frequentist Method*

This semiparametric frequentist method arguments and construction are fairly similar to the previous methods but, as previously hinted, uses a semiparametric location regression model for $Y_{\bar{D}}$ to estimate $F_{\bar{D}}$ and estimates outer probability empirically (Rodriguez-Alvarez and Inacio de Carvalho, 2018; Janes and Pepe, 2008) making it less computationally heavy, i.e faster in comparison. Another advantage is being able to provide direct insight on covariate influence over the test/marker with the fit model's parameters.

The frequentist method for covariate adjustment was first implemented in rocreg from Stata 2013 complementing other Stata ROC commands such as roctab and roccomp (StataCorp, 2013). It arguably remained one of the most in depth options for covariate adjustment available in the market until the release of the AROC package, it is however only available through Stata which requires an yearly licence and can prove challenging for inexperienced users.

3.2.5 *Other Methods*

Posterior Predictive Checks (PPC)

All previous methods are meant to construct and analyze the AROC curve and the behavior of the incorporated covariates, for the remainder of this section the methods focus on posterior predictive checks and pooled ROC estimation.

Both *predictive.checks.AROC.bnp* and *predictive.checks.AROC.bsp* are implementations of PPCs on their respective Bayesian based method. The premise behind PPCs is evaluating the generated model on how well it is able to generate data similar to the data observed (Gabry et al., 2019) utilizing in this case the B-splines dependent Dirichlet process and Bayesian normal linear regression model for the *AROC.bnp* and *AROC.bsp* objects respectively (Rodriguez-Alvarez and Inacio de Carvalho, 2018). To exemplify these methods we use the previously generated objects.

While the mandatory argument is solely the AROC object, the *devnew* was changed from the default *TRUE* argument to display all depicted graphics on the same device. These graphics are histograms of the desired test statistics, by default minimum, maximum, median and skewness and Kernel density estimates showing diagnostic test outcome in the nondiseased group as well.

Pooled ROC

The package also provides two methods of polled ROC estimation, *pooledROC.emp* an implementation of empirical estimation proposed by Hsieh and Turnbull (Hsieh and Turnbull, 1996; Rodriguez-Alvarez and Inacio de Carvalho, 2018) and *pooledROC.BB* for Bayesian bootstrap estimation proposed by Gu et al. (2008) (Rodriguez-Alvarez and Inacio de Carvalho, 2018). Both methods are similar to construct and have a similar output structure, mandatory arguments identify diseased and non diseased groups in the test or marker column which can be achieved with a straightforward indexing in R.

3.3 R SHINY

Shiny is an R package which provides a framework to develop interactive web-based applications such as data summaries and queries to end users through a standard web browser. A recent addition to the R repository, its applications have provided an excellent resource for users to work with complex R packages or perform data analysis through a extensible visual framework based around HTML and CSS with further JavaScript and jQuery integration to extend the scope of possible applications.

The package is based around reactive programming, a programming paradigm that facilitates the automatic propagation of change of dataflows. A practical understanding of this concept is, while working with several inputs generating a specific output, be it plots, tables or text, any modification on input will automatically generate and update the output without requiring a new command or refresh action on the user's side. Another user friendly implementation is the customization of the shiny interface using widgets and code chunks promoted by RStudio and implemented across several other apps available on their archive

The standard architecture of a shiny app is two scripts in the same directory, *ui.R* for the User interface (UI) and *server.R* for internal calculations and app behavior. These two files can also be joined in a single *app.R* however this option is best used for applications such as data visualization (Beeley, 2013).

3.4 COMP2ROC

The final package required to build the RealROC app is Comp2ROC. To establish a statistical significant confounding affect of a given covariate on the ROC curve we have explored the possibility of using the fit model's parameters as seen in section 3.2.4 however another well established alternative is available in some scenarios, ROC curve comparison.

When comparing ROC curves, Area Under the Curve (AUC) is frequently used to indicate greater performance. The Z-test, a non parametric approximation of the Wilcoxon-Mann-

Whitney test is often used for this comparison, however performance misreadings can occur due to curve crossing, further argued is the ROC curves inclusion of areas with very little interest that can influence the result perchance claiming that, for instance, two ROC curves have no statistical differences overall disregarding partial areas of interest (Braga et al., 2013).

For a more robust comparison of the two ROC curves the Braga methodology (Braga et al., 2013) and package Comp2ROC from the same author was selected. The method uses a collection of sampling lines similar to multi-objective distinct optimization algorithms allowing ROC curve comparison in several regions of space evaluating statistical performance and generating confidence intervals with non parametric bootstrap re-sampling method Braga et al. (2013).

Due to compatibility issues we cannot use this package to compare curves with and without accounting for a covariate affect however in instances where we can separate the populations assigned with a given covariate value, such as in the case of a categorical variable, this method can be used to calculate the differences between curves, inferring the covariate affect which, if used in tandem with the AROC method, can give further credence to the conclusions drawn when using RealROC.

This dissertation did not thoroughly explore ROC curve comparison however both the package original article (Braga et al., 2013) and documentation (Braga et al., 2016) as well as Krzanowski and Hand (2009) provide valuable resources on this topic.

4

REALROC WEB APPLICATION

4.1 APPLICATION FLOW

RealROC follows a linear flow intended to replicate an analysis progression from building the standard ROC curve to adjusting it followed by a comparison of both and a report with all relevant statistics that can be consulted at any time. The user is also able to skip or return to any of the previous steps to change any parameters as needed.

The app is comprised of seven distinct modules that will be individually explored in the next section.

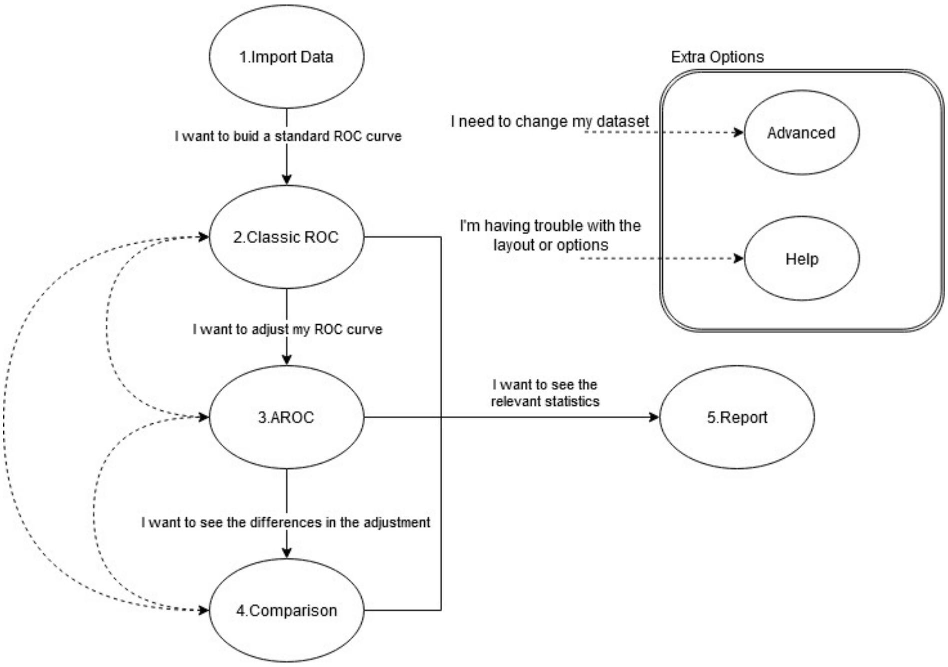

1.Import Data
I want to buid a standard ROC curve
2.Classic ROC
I want to adjust my ROC curve
3.AROC
I want to see the differences in the adjustment
4.Comparison
I want to see the relevant statistics
5.Report
Extra Options
I need to change my dataset
Advanced
I'm having trouble with the layout or options
Help

Figure 4: RealROC flowchart.

The application is available online at https://frmachadoecosta.shinyapps.io/RealROC/, additionally it can be downloaded or run locally by downloading it at https://github.com/frmachadoecosta/RealROC where the source code is also available.

4.2 APPLICATION MODULES

4.2.1 *Home Screen*

When RealROC opens, the user is presented with an option of uploading the desired data, in csv or xml format, or using the sample data provided by the app. The sample data, *endosim.csv*, is present in another package developed by Rodriguez-Alvarez and Javier Roca-Pardinas (2017) and used Body Mass Index (BMI) to detect patients having a higher risk of cardiovascular problems, with age and biological sex as covariates. This data was selected not just because BMI is a well established index, making both values and results easier to understand but also for the presence of both a continuous and binary variables, age and sex respectively, to allow for different approaches when adjusting the ROC curve.

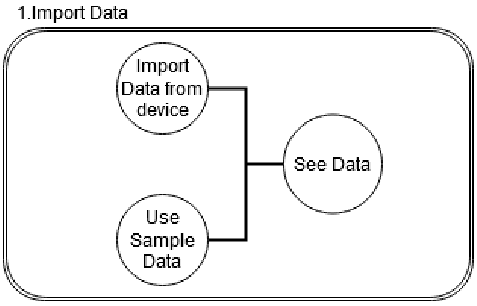
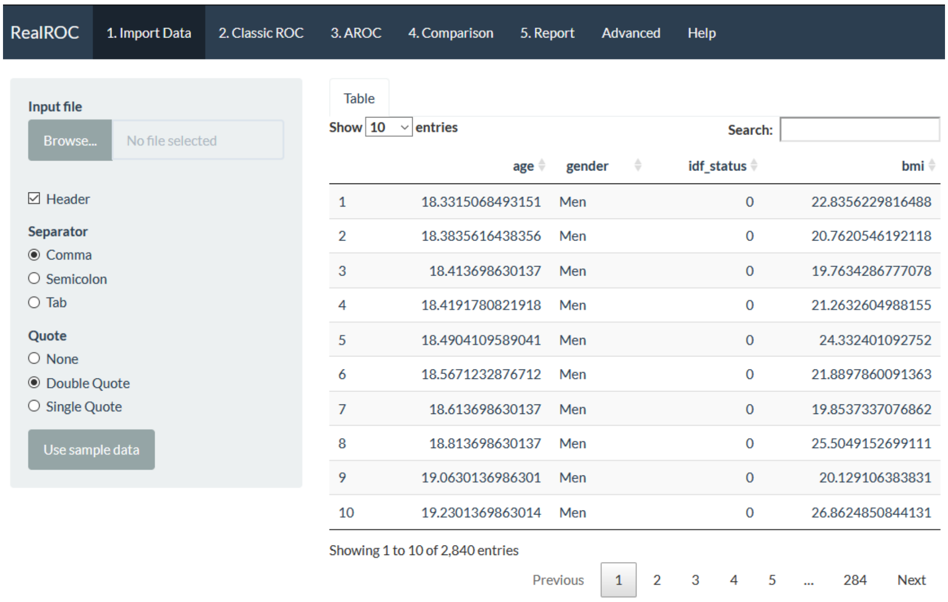


Figure 5: Home Screen options flowchart.

Either option will trigger the appearance of a data-table where the user can consult their data and adjust the options for import if necessary as can be seen in Figure 6.



Input file

No file selected

☑ Header

Separator

● Comma

○ Semicolon

○ Tab

Quote

○ None

● Double Quote

○ Single Quote

Show 10 entries

Search:

	age	gender	idf_status	bmi
1	18.3315068493151	Men	0	22.8356229816488
2	18.3835616438356	Men	0	20.7620546192118
3	18.413698630137	Men	0	19.7634286777078
4	18.4191780821918	Men	0	21.2632604988155
5	18.4904109589041	Men	0	24.332401092752
6	18.5671232876712	Men	0	21.8897860091363
7	18.613698630137	Men	0	19.8537337076862
8	18.813698630137	Men	0	25.5049152699111
9	19.0630136986301	Men	0	20.129106383831
10	19.2301369863014	Men	0	26.8624850844131

Showing 1 to 10 of 2,840 entries

Previous 1 2 3 4 5 ... 284 Next

Figure 6: Home Screen using sample data.

4.2.2 *Classic ROC Analysis*

Classic ROC is the first of 3 modules that present the same overall behaviour, the user intends to perform a given analysis, selects the appropriate method and inputs the parameters needed to display an output, a simplified option flowchart is available in Figure 7.

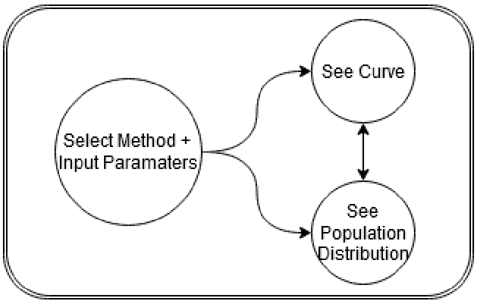


Figure 7: Classic ROC and AROC module options.

Classic ROC allows the user to perform the standard or classic ROC analysis with no covariate attached, parameters are input in the lateral panel where marker and result tabs should be specified and are dependent on the names given in the original dataset as well as the control/healthy and case/disease tags.

Methods included are empirical and pooled estimators present in pROC and AROC package respectively, that generate the ROC curve after the command is given by pressing the bottom button of the lateral screen, this will in turn collect the inputs and construct the curve and population density plots that can be seen in Figure 8.

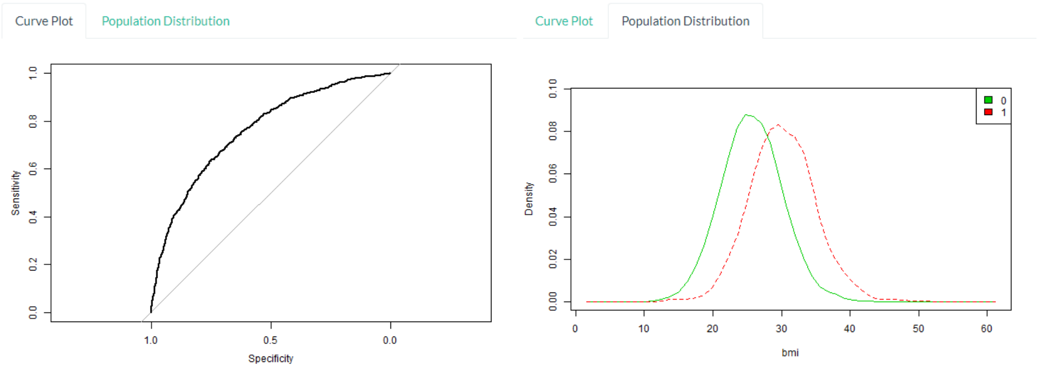


Figure 8: Classic ROC module output using sample data.

4.2.3 *AROC*

The AROC module allows the user to perform AROC analysis. This has a similar flow to Classic ROC, seen in Figure 7, however an extra option for selection of the covariate is present in the lateral panel. Users that began their analysis in the previous module will note that the options selected previously are saved between modules to enhance the experience, needing only to select their desired covariate if all remaining options are the same. As of release this is limited to one covariate per analysis to cover the majority of cases with no added complexity of the influence of interdependent variables, or values of different interactions with the marker.

The AROC methods mentioned in section 3.2 with the exception of the kernel estimator since it only allows for continuous covariates, a specification that restricts the user selection process, that was hence dropped for the time being.

The output curve should be similar to the previous example however in the case of population distribution a fork occurs, if the covariate selected is continuous a dotplot will be displayed exhibiting the density distribution of cases and controls across the covariate, however if the covariate happens to be binary a boxplot will appear displaying the behaviour of the marker in separate covariate values, this can be seen in Figure 9.

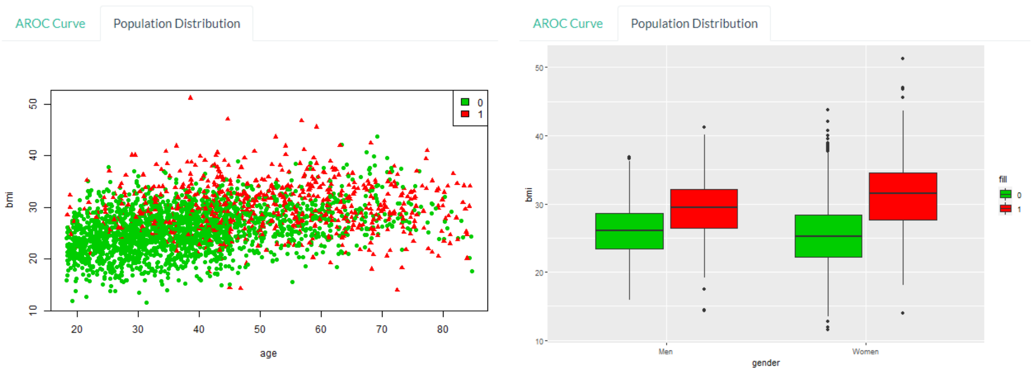


Figure 9: AROC module Population Distribution output using a continuous covariate - *age* (left) and binary covariate - *gender* (right).

4.2.4 *Comparing adjustment*

The Comparison module is more complex than the previous modules, the user will again note that the previous choices are again present in the selection tab however an extra option is available once again, a method selection of comparison mentioned in Chapter 3. If the user wishes to perform a comparison on a continuous covariate they must choose the AROC method of comparison, however if the covariate is of a binary nature it can opt to use either the Comp2ROC method or the AROC method, a simplified version of this selection can be seen in Figure 10.

For output, the AROC method presents the user with a superimposed plot of pooled empirical and frequentist AROC curves in a ggplot environment, these methods are chosen to provide not just the most straight forward comparison but also a larger amount of information on the covariate effect we will see in the Report section. Provided the user selected the Comp2ROC method with a binary covariate they will be presented with two covariate-specific ROC curves and information about their differences can be found in the Report section. These outputs can be seen in Figure 11.

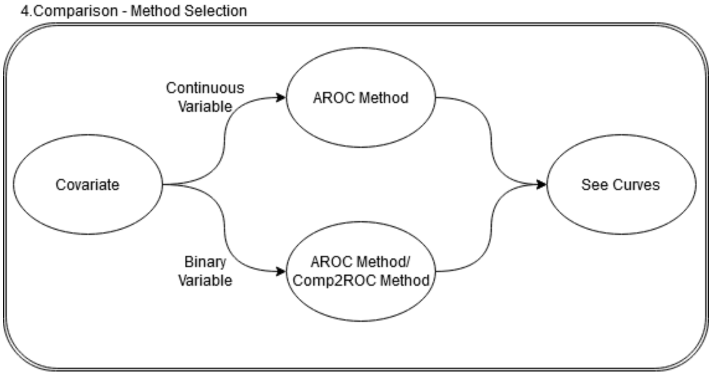


Figure 10: Comparison module selection options.

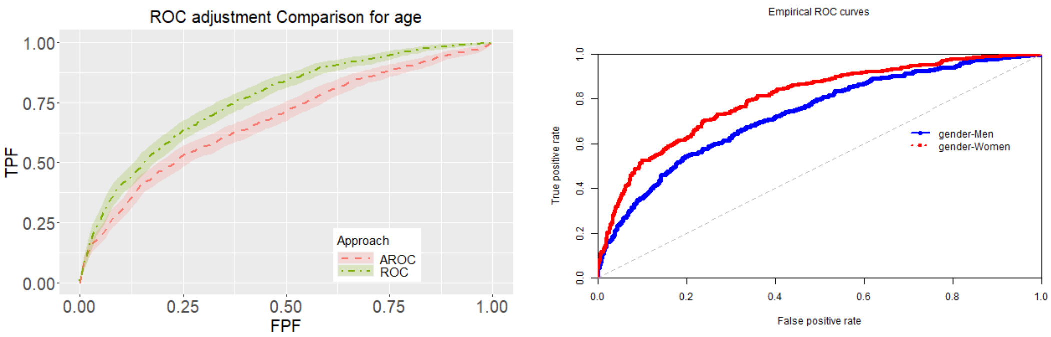


Figure 11: Comparison module plot outputs for AROC method (left) and Comp2ROC method (right).

4.2.5 *Report Screen*

The report section logs all actions taken in the application and is the closest option to an R console output however the outputs for each action are parsed to present the most straightforward and clear information possible. This section can be accessed at any time by the user to get extra information on the generated ROC curves, such as AUC and CI values, in the case of covariate adjustment and the comparison module this is displayed with extra information on the fitted regression model and Z-statistic value and p-value, that will help the user determine not just through visual output if the covariate has add a confounding effect on their data.

4.2.6 *Advanced Options*

The Advanced section is the most sensitive section within the app. This module was designed to allow some modification to the uploaded data, often required by analysts, to try and avoid the need for extra modifications outside the application. These modifications can be inverting the signal of a given variable, such as markers that do not follow the standard ROC assumption of $Y_D > Y_{\bar{D}}$, a simple logarithmic transformation or even fixing common mistakes of the data such as a binary variable being recorded as a numeric value that can have implications on output.

These small adjustments are key to circumvent some of the app's limitations regarding data manipulation, that would ideally not be required, however this change of data is unsafe as it can erase input selections and modify the data in detrimental ways to the analysis. These changes can be seen in the data table in the Home screen however they should be limited and performed before any calculations are made so as to not compromise results.

4.2.7 *Help Section*

The Help section is designed to guide new users in the functionality of each module in a structure similar to the one found throughout this chapter. This is achieved using the Readme file generated by Github and displaying it in the app ensuring both methods of accessing the application and reading the documentation grant the same information with the same structure.

5

CASE STUDY - CRIB SCORE AND THE SEX COVARIATE

5.1 NEONATAL MORTALITY AND CRIB SCORE

The following case study will show an application of RealROC comparing it with the conventional R scripting method to show if biological sex has a confounding effect in a neonatal mortality assessment system.

A common healthcare goal across countries is the reduction of neonatal mortality, in fact it is a listed concern of the United Nations' Sustainable Development Goals (SDG), which targets Neonatal Mortality Rates (NMR) reduction to 12 per 1000 births by 2030 (Harahap et al., 2019). Preterm birth is listed as the most important factor of mortality in developed countries that accompanied by birth weight and gestational age are significant univariant predictors of mortality risk (Terzic and Heljić, 2012). Along with advancements on perinatal medicine, increase support techniques in Neonatal Intensive Care Units (NICUs) and development of adequate scoring systems for assessing mortality risk have lead to a global 51% decrease on NMR from 1990 to 2017 (Brito et al., 2003; Hug et al., 2019).

Clinical risk index for babies (CRIB) is a risk assessment tool used in NICUs for infants born with less than 32 week gestation or 1500g or lower birth weight. Along with the Score for Neonatal Acute Physiology (SNAP), these scoring systems and their updates have served as prediction tools more accurate to the previous weight or gestational age univariate predictors (Brito et al., 2003).

CRIB score uses six different variables obtained routinely during the first 12 hours of life, namely, birth-weight, gestational age, the presence of congenital malformation (excluding inevitably lethal congenital malformations) and indices of physiological status (maximum base deficit, minimum and maximum appropriate fraction of inspired oxygen (FiO_2), all measured in the first 12 hours) resulting in a score between 0 and 24 where higher values denote a higher risk of death. This scoring system is a staple in NICUs for both statistical and practical reasons, being not just an overall better scoring system than most of its contemporaries (Braga et al., 2013; Bastos et al., 1997) but also for the ease of data collection and score calculation making the test last only 5 minutes per infant (Dorling et al., 2005).

The CRIB score was derived using data from infants admitted to UK tertiary neonatal units from 1988 to 1990, this lead to Parry et al. (2003) raising concerns over poor calibration to contemporary data, the inclusion of FiO_2 also warranted criticism since it was considered a subjective entry determined by the care team rather than a physiological measurement. Similarly the addition of data up to 12h after admission also lead to concern over early treatment bias (Parry et al., 2003). These issues eventually resulted in an updated CRIB score, CRIB-II.

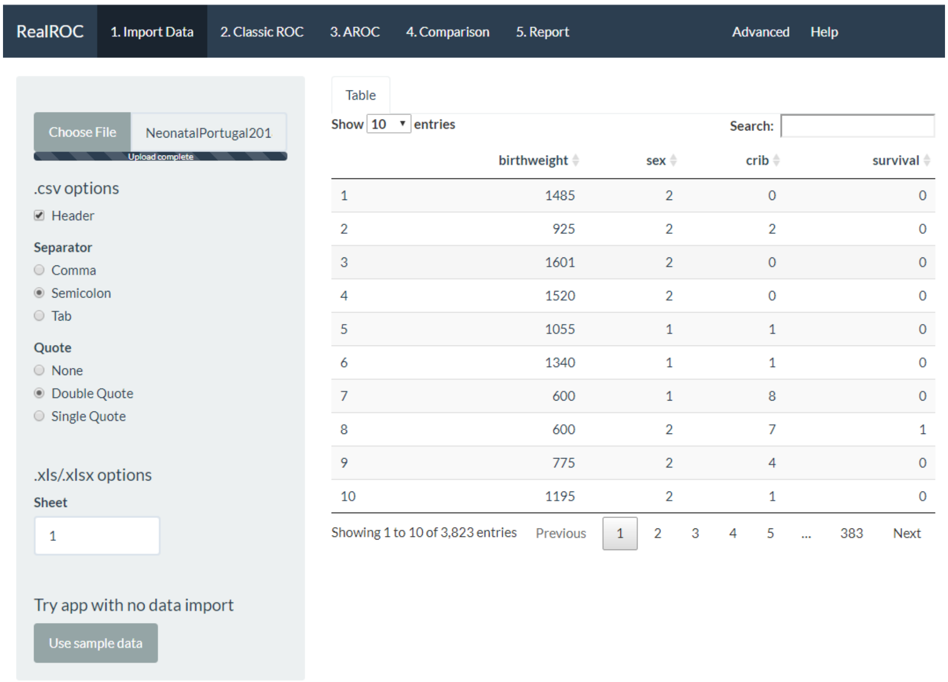
CRIB-II score system maintains a relatively low number of variables needed for calculation, maintaining the advantage of its predecessor, gestational age, birth weight, admission temperature and base excess are used to predict mortality. This new prediction tool was met with some skepticism with reports of studies comparing both systems, showing no statistical difference (Gagliardi et al., 2004; Felice et al., 2005) nevertheless CRIB-II is now a known and recognized neonatal scoring system used in some hospitals.

5.2 DATASET

The dataset used is part of the Portuguese National Registry on low weight newborns between 2013 and 2018 and was made available for research purposes. The original data included possible confounding observations such as repeated ids and twins that were removed to ensure no unaccounted variable. After ensuring all id's were unique, these were promptly removed along with any possible identifiable features to abide by European Union's anonymity and data protection standards. The resulting dataset composed of 3823 unique entries registering gestational age in weeks, the mothers age, in years, biological sex of the infant (1-Male; 2-Female), CRIB score (0-21), survival (0-Survival; 1-Death) and other possibly relevant covariates were used for the remainder of the study. An abridged dataset with relevant information was published and is available for consultation (Machado E Costa, 2019).

5.3 REALROC VS SCRIPTING

In both methodologies the dataset must be imported to the environment, this is achieved by choosing a file in the application of by using the *read.csv()* function in R, note that server-side RealROC is performing this same command however it simplifies the process and replaces R syntax with simple button presses. A successful data import can be seen in Figure 12 with a table output where the user can consult their data unlike in native R.



NeonatalPortugal201

.csv options

☑ Header

Separator

○ Comma
● Semicolon
○ Tab

Quote

○ None
● Double Quote
○ Single Quote

.xls/.xlsx options

Sheet

1

Try app with no data import

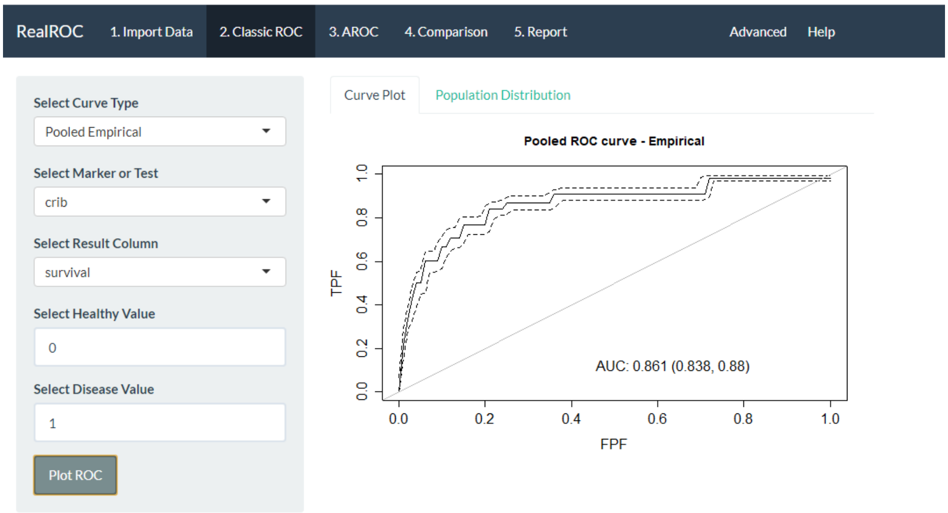
Show 10 entries

	birthweight	sex	crib	survival
1	1485	2	0	0
2	925	2	2	0
3	1601	2	0	0
4	1520	2	0	0
5	1055	1	1	0
6	1340	1	1	0
7	600	1	8	0
8	600	2	7	1
9	775	2	4	0
10	1195	2	1	0

Showing 1 to 10 of 3,823 entries Previous 1 2 3 4 5 ... 383 Next

Figure 12: Successful data import with table output.

Moving to the Classic ROC section, we can begin selecting the parameters and ploting several different ROC curves, in R this requires loading and potentially installing the desired libraries however these come preloaded in the application. Once again the need to know each package correct syntax is replaced by simple data inputs that can display the empirical or pooled ROC curve as seen in Figure 13.



Select Curve Type
Pooled Empirical

Select Marker or Test
crib

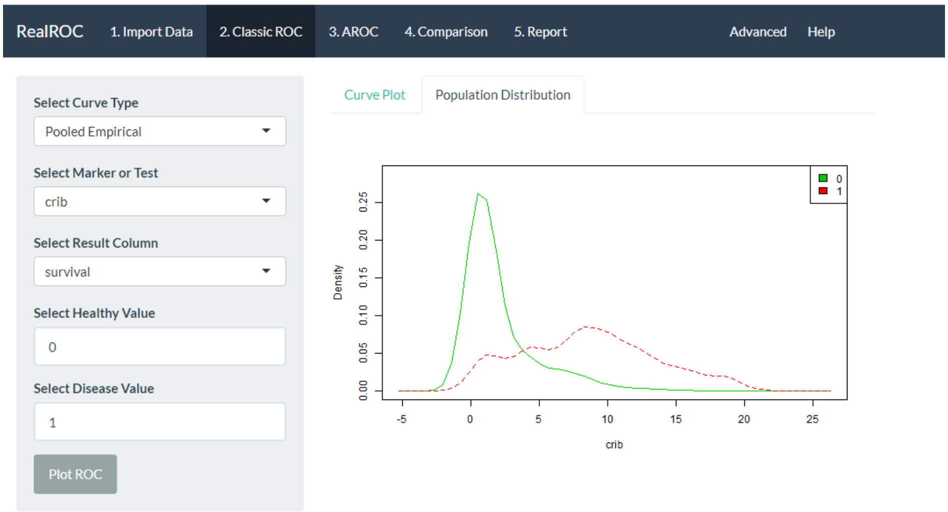
Select Result Column
survival

Select Healthy Value
0

Select Disease Value
1

Figure 13: Pooled empirical ROC curve for CRIB score.

Extra information is even present in the ”Population Distribution” tab displayed in Figure 14 where we can see the population densities for controls and cases, giving a broader view of the data.



Select Curve Type
Pooled Empirical

Select Marker or Test
crib

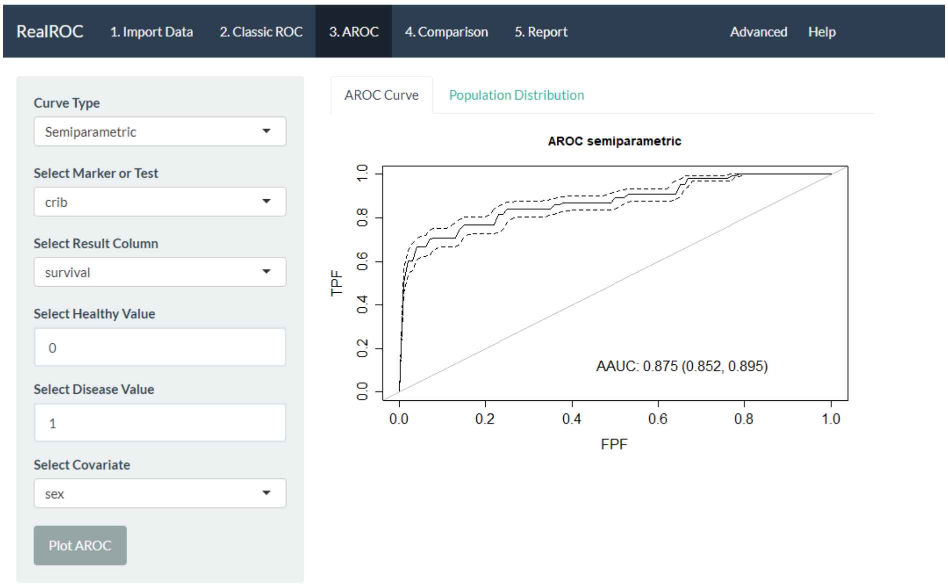
Select Result Column
survival

Select Healthy Value
0

Select Disease Value
1

Figure 14: Population densities of mortality by CRIB score.

Moving to the AROC section, the user will see all previous selected inputs are saved in memory and all that is needed is to correctly select the covariate, in this case ”sex”. The curve type or method of adjustment selected is the frequentist method mentioned in section 3.2.4 for a more in depth summary on the effect of the covariate. While the AROC curve can be seen, the population distribution tab will not display the correct result, this is because the data was originally submitted with the ”sex” column as numeric rather than a factor, this however can be easily fixed in the Advanced section where the nature of the column can be changed. Results for both outputs can be seen in Figures 15 and 16.



Curve Type: Semiparametric

Select Marker or Test: crib

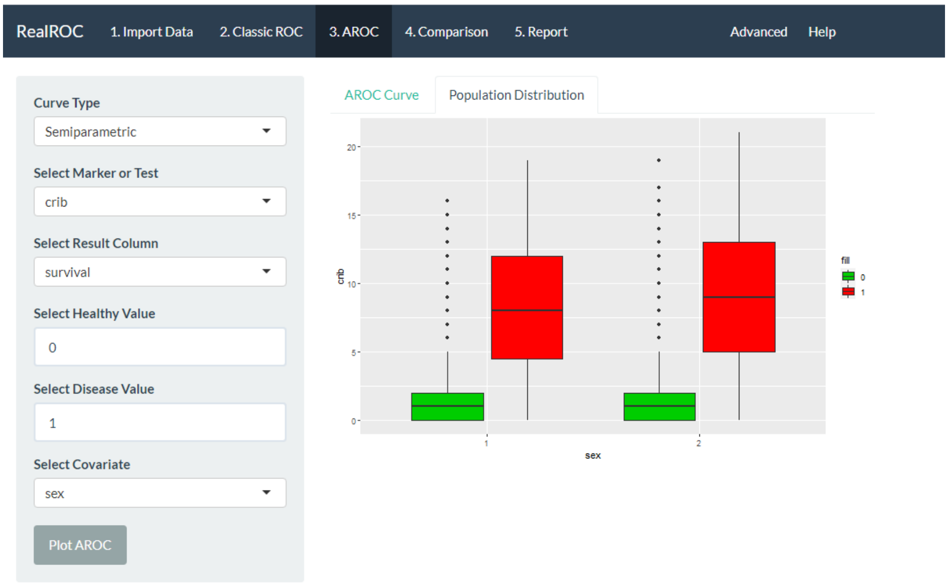
Select Result Column: survival

Select Healthy Value: 0

Select Disease Value: 1

Select Covariate: sex

Figure 15: AROC curve for CRIB score adjusted for sex.



Curve Type: Semiparametric

Select Marker or Test: crib

Select Result Column: survival

Select Healthy Value: 0

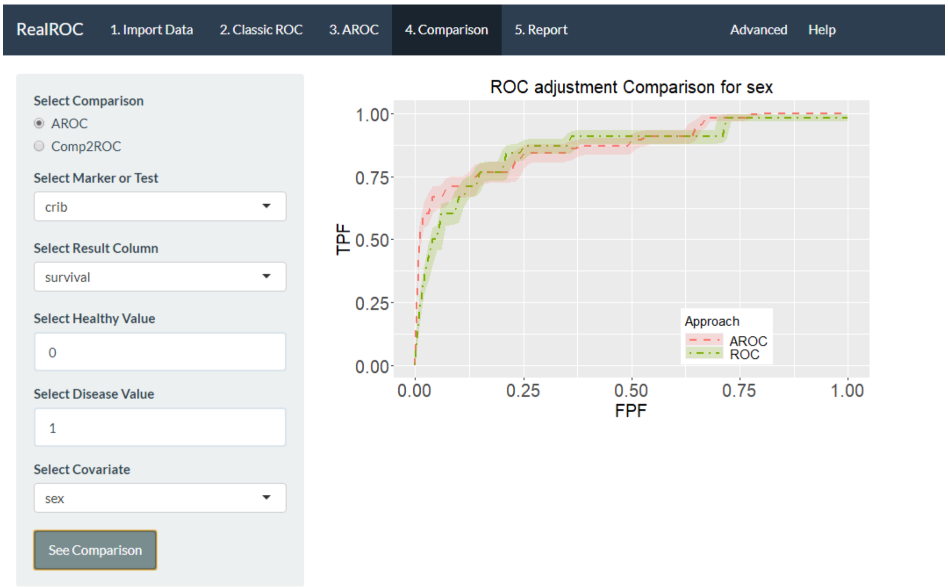
Select Disease Value: 1

Select Covariate: sex

Figure 16: Boxplot of covariate specific CRIB controls and cases.

Already we can see the behaviour of the covariate and might opt to go straight to the Report section to get our summary however we can go further and directly compare the two curves in the Comparison section. Given the nature of how covariate we can use both AROC and Comp2ROC methods mentioned in Chapter 3.

In the Comparison section the gap of usability between app and script is far more noticeable, the AROC method has no method for superimpose two curves generated by the package so for scripting the user would require extensive knowledge of R plot syntax to be able to manually collect each curve points and plot them. This, like other features of the application, is achieved with simple button presses after selection of the method of comparison, displaying a *ggplot* object with CI present as can be seen in Figure 17. For the Comp2ROC method, the package requires a particular data structure to operate, and the user would need to manually rearrange the dataset and re-import it, in RealROC however this is seamless to the previous method and results can be seen in Figure 18.



Select Comparison
● AROC
○ Comp2ROC

Select Marker or Test: crib

Select Result Column: survival

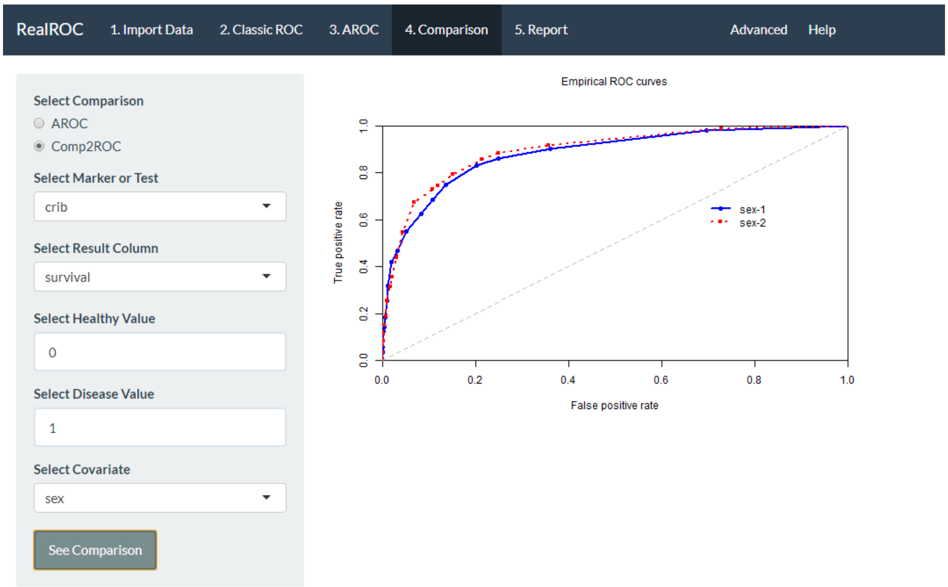
Select Healthy Value: 0

Select Disease Value: 1

Select Covariate: sex

Figure 17: AROC and ROC curve comparison.

With the data thoroughly analysed, the user can now see the summary of their findings in the Report section that can be seen in Figure 19. This section displays several summaries for the many operations made, and a equivalent in the script method would be several different summary commands. The section organizes the many summaries offered by each package in an easily digestible format containing all relevant information.



Select Comparison
○ AROC
● Comp2ROC

Select Marker or Test
crib

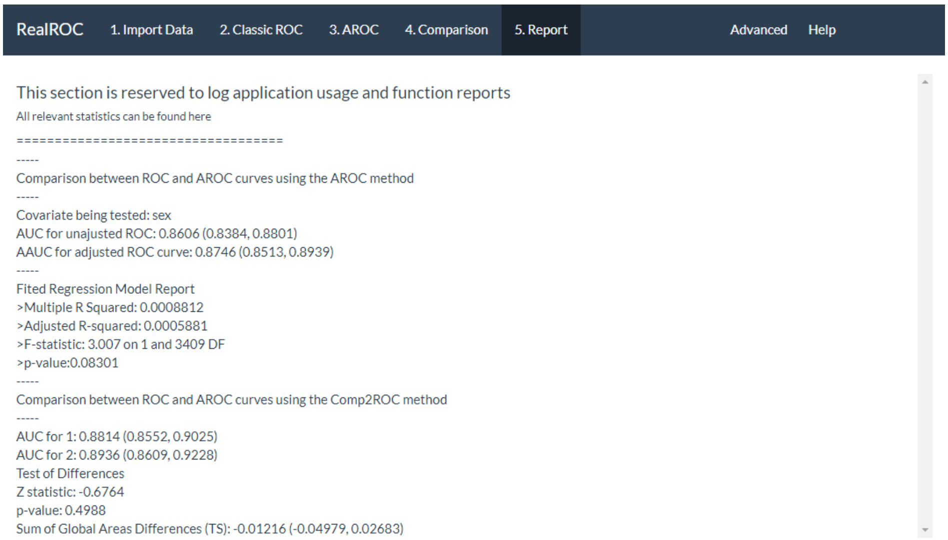
Select Result Column
survival

Select Healthy Value
0

Select Disease Value
1

Select Covariate
sex

Figure 18: Comp2ROC comparison of covariate specific ROC curves.



```
This section is reserved to log application usage and function reports
All relevant statistics can be found here
==================================
-----
Comparison between ROC and AROC curves using the AROC method
-----
Covariate being tested: sex
AUC for unajusted ROC: 0.8606 (0.8384, 0.8801)
AAUC for adjusted ROC curve: 0.8746 (0.8513, 0.8939)
-----
Fited Regression Model Report
>Multiple R Squared: 0.0008812
>Adjusted R-squared: 0.0005881
>F-statistic: 3.007 on 1 and 3409 DF
>p-value:0.08301
-----
Comparison between ROC and AROC curves using the Comp2ROC method
-----
AUC for 1: 0.8814 (0.8552, 0.9025)
AUC for 2: 0.8936 (0.8609, 0.9228)
Test of Differences
Z statistic: -0.6764
p-value: 0.4988
Sum of Global Areas Differences (TS): -0.01216 (-0.04979, 0.02683)
```

Figure 19: Report section for sex adjusted CRIB.

In this instance we can see that the AROC method shows a p-value > 0.05 which coupled with the Z statistic p-value > 0.05 and the sum of global areas crossing 0 along with the CI intersections clearly shows no statistical significance to the sex covariate in the CRIB score system.

5.4 FINAL REMARKS

Results from both covariate adjustment and ROC curve comparison methods clearly indicate sex is not a confounding covariate in the CRIB score meaning the sex of the infant has no statistical relevance to the score system or the mortality outcome. These findings validate previous research on the matter (Terzic and Heljić, 2012; Mourão et al., 2014) from more contemporary sources but also come from a substantially larger pool of data.

RealROC displays an intuitive option selection menu and clear output which allows any user to perform a thorough analysis without R code syntax details. It is, of course, worth noting that all results displayed by the app can be achieved with R by an experienced user however the time to do so would be substantially longer than using the app.

6

CONCLUSION

ROC curves have aided statistical analysis and decision statistics for over half a century, their simple classification model and straightforward summary statistics have been paramount in the virtually limitless two-class prediction problem.

The curve's history is filled with new coefficients, derivations, summary statistics and interpretations for existing theory to better fit such an ubiquitous tool to specific applications.

Covariate specific and covariate adjusted ROC curves are a relatively newer addition to the existing theory hence the lack of specific tools mentioned in section 3.1. This addition however widens the scope for this tool even further with works like the ones from Janes and Pepe (2008) and Rodríguez-Álvarez et al. (2018) only serving to expand its potential.

The goal for this dissertation was to cement these new found applications for the ROC curve via the development of a specialize tool. RealROC was developed as a way to introduce users to the AROC concept and apply it to their database and hopefully lead them to more in depth conclusions about their data and the relations between variables.

While this dissertation focused on the bioinformatics and medical applications for this tool, RealROC is built to allow the same diversity of data the classic ROC analysis is able to compute.

The case study presented in the previous chapter demonstrated an intuitive software that simplifies the AROC analysis and allows users to perform their intended studies in a shorter amount of time and effort, and while undeniable that a pure R code can provide greater freedom to an experienced user in customization and specific calculations the app greatly reduces the know-how entry barrier.

The application is already released and available at https://frmachadoecosta.shinyapps.io/RealROC/ however new improvements are expected such as introducing several covariates to the confounding study as well as specifying the relation between covariates and between each covariate and the marker, introducing the ability to compare markers in the Comparison module, new additions to the Advanced tab to allow smoother data manipulation inside the app and any and all performance and bug fixes reported by users through github.

BIBLIOGRAPHY

Bastos, G., Gomes, A., Oliveira, P., and Da Silva, A. T. (1997). Comparação de quatro escalas de avaliação da gravidade clínic a (CRIB, SNAP, SNAP-PE, NTISS) em recém nascidos prematuros. *Acta Medica Portuguesa*, 10(2-3):161–165.

Beeley, C. (2013). *Web application development with R using Shiny*. Packt Publishing.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Braga, A. C., Costa, L., and Oliveira, P. (2013). An alternative method for global and partial comparison of two diagnostic systems based on roc curves. *Journal of Statistical Computation and Simulation*, 83(2):307–325.

Braga, A. C., Frade, H., Carvalho, S., and Santiago, A. M. (2016). *Comp2ROC: Compare Two ROC Curves that Intersect*. R package version 1.1.4.

Brito, A. S., Matsuo, T., Gonzalez, M. R. C., de Carvalho, A. B. R., and Ferrari, L. S. L. (2003). CRIB score, birth weight and gestational age in neonatal mortality risk evaluation. *Revista de saúde pública*, 37(5):597–602.

Choi, Y.-K., Johnson, W. O., Collins, M. T., and Gardner, I. A. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(2):210–229.

Dodd, L. E. and Pepe, M. S. (2003). Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*, 98(462):409–417.

Dorling, J. S., Field, D. J., and Manktelow, B. (2005). Neonatal diseases severity scoring systems. *Archives of Disease in Childhood: Fetal and Neonatal Edition*, 90(1):11–16.

Egan, J. P. (1975). *Signal detection theory and ROC-analysis*, volume 1. ISBN 0122328507.

Erkanli, A., Sung, M., Jane Costello, E., and Angold, A. (2006). Bayesian semi-parametric roc analysis. *Statistics in Medicine*, 25(22):3905–3928.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):179–192.

Felice, C. d., del Vecchio, A., and Latini, G. (2005). Evaluating illness severity for very low birth weight infants: Crib or crib-ii? *The Journal of Maternal-Fetal & Neonatal Medicine*, 17(4):257–260.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 182(2):389–402.

Gagliardi, L., Cavazza, A., Brunelli, A., Battaglioli, M., Merazzi, D., Tandoi, F., Cella, D., Perotti, G. F., Pelti, M., Stucchi, I., Frisone, F., Avanzini, A., and Bellù, R. (2004). Assessing mortality risk in very low birthweight infants: a comparison of crib, crib-ii, and snappe-ii. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 89(5):F419–F422.

Gu, J., Ghosal, S., and Roy, A. (2008). Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine*, 27(26):5407–5420.

Harahap, N. C., Handayani, P. W., and Hidayanto, A. N. (2019). Informatics in Medicine Unlocked Barriers and technologies of maternal and neonatal referral system in developing countries : A narrative review. *Informatics in Medicine Unlocked*, 15(January):100184.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.*, 24(1):25–40.

Hug, L., Alexander, M., You, D., and Alkema, L. (2019). National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *The Lancet Global Health*, 7(6):e710–e720.

Janes, H. and Pepe, M. S. (2008). Adjusting for Covariates in Studies of Diagnostic, Screening, or Prognostic Markers: An Old Concept in a New Setting. *American Journal of Epidemiology*, 168(1):89–97.

Janes, H. and Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96(2):371–382.

Krzanowski, W. and Hand, D. (2009). *ROC Curves for Continuous Data*, volume 111. ISBN 978-1-4398-0021-8.

Machado E Costa, F. (2019). Neonatalportugal2018. https:/data.mendeley.com/datasets/br8tnh3h47/1.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Metz, C. E., Herman, B. A., and Shen, J.-H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, 17(9):1033–1053.

Mourão, M. F., Braga, A. C., and Oliveira, P. N. (2014). CRIB conditional on gender: Nonparametric ROC curve. *International Journal of Health Care Quality Assurance*, 27(8):656–663.

Obuchowski, N. A. (2003). Receiver Operating Characteristic Curves and Their Use in Radiology. *Radiology*, 229(1):3–8.

Obuchowski, N. A. and Bullen, J. A. (2018). Receiver operating characteristic (ROC) curves: Review of methods with applications in diagnostic medicine. *Physics in Medicine and Biology*, 63(7).

Park, S. H., Goo, J. M., and Jo, C.-H. (2004). Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean Journal of Radiology*, 5(1):11.

Parry, G., Tucker, J., Tarnow-Mordi, W. O., and UK Neonatal Staffing Study Collaborative (2003). CRIB II : an update of the clinical risk index for babies score For personal use . Only reproduce with permission from The Lancet Publishing Group . *Lancet*, 361:1789–1791.

Pepe, M. S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84(3):595–608.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54(1):124–35.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Sciences Series. ISBN 0198565828.

Rodriguez-Alvarez, M. X. and Inacio de Carvalho, V. (2018). *AROC: Covariate-Adjusted Receiver Operating Characteristic Curve Inference*. R package version 1.0.

Rodriguez-Alvarez, M. X. and Javier Roca-Pardinas (2017). *npROCRegression: Kernel-Based Nonparametric ROC Regression Modelling*. R package version 1.0-5.

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(4):483–499.

Rodríguez-Álvarez, M. X., Roca-Pardiñas, J., Cadarso-Suárez, C., and Tahoces, P. G. (2018). Bootstrap-based procedures for inference in nonparametric receiver-operating characteristic curve regression analysis. *Statistical Methods in Medical Research*, 27(3):740–764.

Smith, P. J. and Thompson, T. J. (1996). Correcting for Confounding in Analyzing Receiver Operating Characteristic Curves. *Biometrical Journal*, 38(7):857–863.

StataCorp (2013). *Stata: Release 13*. Stata Press. ISBN 1-59718-115-3.

Terzic, S. and Heljić, S. (2012). Assessing mortality risk in very low birth weight infants. *Medicinski arhiv*, 66:76–9.

Wang, C., Turnbull, B., Gröhn, Y., and Nielsen, 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(1):128–146.