

Evaluation of Strategies to Combine Multiple Biomarkers in Diagnostic Testing

By

Muna Balla Elshareef Mohammed

mimielshareef@gmail.com

Supervisor : Professor Henry G. Mwambi

mwambih@ukzn.ac.za

Co-Supervisor : Dr. Lori E. Dodd

doddl@niaid.nih.gov



School of Mathematics, Statistics and Computer Sciences

University of KwaZulu-Natal

Pietermaritzburg, South Africa

A thesis submitted for the fulfillment of the requirements for
Masters of Science at the
School of Mathematics, Statistics and Computer Sciences, University of
KwaZulu-Natal, Pietermaritzburg

December 2012



Abstract

A challenge in clinical medicine is that of correct diagnosis of disease. Medical researchers invest considerable time and effort to enhance accurate disease diagnosis. Diagnostic tests are important components in modern medical practice. The *receiver operating characteristic* (ROC) is a commonly used statistical tool for describing the discriminatory accuracy and performance of a diagnostic test. A popular summary index of discriminatory accuracy is the *area under ROC curve* (AUC). In the era of high-dimensional data, scientists are evaluating hundreds to multiple thousands of biomarkers simultaneously. A critical challenge is the combination of these markers into models that give insight into disease. In infectious disease, markers are often evaluated in the host as well as in the microorganism or virus causing infection, adding more complexity to the analysis. In addition to providing an improved understanding of factors associated with infection and disease development, combinations of relevant markers is important to diagnose and treat disease. Taken together, this presents many novel and major challenges to, and extends the role of, the statistical analyst.

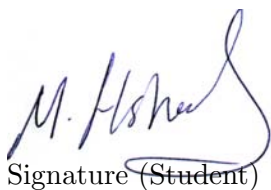
In this thesis, we will address the problem of how to select from multiple markers using existing methods. Logistic regression models offer a simple method for combining markers. We applied resampling methods (e.g., Cross-Validation and bootstrap) to adjust for overfitting associated with model selection. We simulated several multivariate models to evaluate the performance of the resampling approaches in this setting. We applied the methods to data collected from a study of tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS) in Cape Town. Baseline levels of five biomarkers were evaluated and we used this dataset to evaluate whether a combination of these biomarkers could accurately discriminate between Tuberculosis Immune Reconstitution

Inflammatory Syndrome (TB-IRIS) and non TB-IRIS patients, applying AUC analysis and resampling methods.

Preface

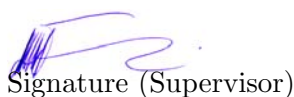
The work described in this thesis was carried out from February 2011 to December 2012, under the supervision and direction of Professor Henry G. Mwambi, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg and Dr Lori E. Dodd, Biostatistics Research Branch, Division of Clinical Research, National Institute of Allergy and Infectious Diseases (NIAID), USA.

The thesis represent original work of the author and has not been otherwise been submitted in any form for any degree or diploma to any University. Where use has been made of the work of others it is duly acknowledged in the text.



Signature (Student)

Date: 5th of December 2012



Signature (Supervisor)

Date: 5th of December 2012



Signature (Co-Supervisor)

Date: 5th of December 2012

Dedication

TO MY LOVELY PARENTS DR BALLA AND HANAN, MY DEAR HUSBAND AYOUB, MY LOVELY DAUGHTER FATIMA, MY BROTHERS MUJTABA, AHMED AND TO THE SOUL OF MY SISTER FATIMA (TOTA), I DEDICATE THIS WORK.

Acknowledgements

First of all, I thank ALLAH for his Grace and Mercy showered upon me. I heartily express my profound gratitude to my supervisors, Professor Henry G. Mwambi and Dr Lori E. Dodd, for their invaluable learned guidance, advises, encouragement, understanding and continued support they have provided me throughout the duration of my studies which led to the compilation of this thesis. I will be always indebted to them for introducing me to this fascinating area of application in health research and creating my interest in Biostatistics.

I lovingly thank my dear husband Ayoub, who supported me each step of the way and without his help and encouragement it simply never would have been possible to finish this work.

I also would like to thank my lovely parents Hanan and Balla for their continuous support and best wishes.

I am grateful for the facilities made available to me by the School of Mathematics, Statistics and Computer Science of the University of KwaZulu-Natal (UKZN), Pietermaritzburg. I am also grateful for the financial support that I have received from UKZN. My thanks extend to Professor Robert Wilkinson and Dr Suzaan Marais for supporting us with real dataset.

Finally I sincerely thank my entire extended family represented by Balla, Hanan, Mohammed Elmojutaba, Ahmed, Fatima (tota), Basheer, Suaad, Eihab, Adeeb, Nada, Tayseer and Samah.

Table of Contents

Abstract	ii
Preface	iv
Dedication	v
Acknowledgements	vi
Table of Contents	vii
List of Notations	x
1 Introduction	1
1.1 Motivation and Purposes	1
1.2 Background and Related Studies	3
1.3 Thesis Outlines	4
2 Receiver Operating Characteristic (ROC) Curves	7

2.1	Definitions and Basic Concepts	7
2.2	Introduction to Receiver Operator Characteristic Curve (ROC) for Continuous Tests	9
2.2.1	Definition of ROC Curves	9
2.2.2	Properties and Attributes of ROC Curves	10
2.2.3	Binormal ROC Curves	12
2.3	Some of ROC Curves Indices	14
2.3.1	Area Under ROC Curves (AUC)	14
2.3.2	The $ROC(t_0)$	16
2.3.3	Partial AUC	17
2.3.4	Kolmogorov-Smirnov (KS)	17
2.4	Motivation for Combining Multiple Biomarkers	18
2.4.1	Boolean Combinations	18
2.4.2	The Likelihood Ratio Method (LR)	19
2.4.3	Logistic Regression	20
3	Resampling Methods	21
3.1	Cross-Validation	21
3.2	Bootstrap Method	23
3.3	Permutation Test	25
3.4	Feature Selection	26
3.5	Resampling Methods in the Context of Combining Multiple Biomarkers	28

3.5.1	Algorithm to Obtain the AUC Through Cross-Validation	29
3.5.2	Algorithm to Estimate the Variance of AUC Through Bootstrapping	29
4	Simulation Studies and Application	31
4.1	Simulation Studies	31
4.1.1	Introduction to Simulated Data	31
4.1.2	Simulations Methods and Results	34
4.2	Application to Real Dataset	39
4.2.1	Tuberculosis Immune Reconstitution Inflammatory Syndrome (TB-IRIS) Dataset	39
4.2.2	Data Results	40
5	Conclusion	46
6	Appendix	48
	Bibliography	54

List of Notations

TP	true positive
TN	true negative
FP	false positive
FN	false negative
TPF	true positive fraction
FPF	false positive fraction
TNF	true negative fraction
FNF	false negative fraction
PPV	positive predictive value
NPV	negative predictive value
LR	likelihood ratio
LR ⁺	positive likelihood ratio
LR ⁻	negative likelihood ratio
ROC	receiver operator characteristic
AUC	area under ROC curve
TAUC	true AUC
pAUC	partial area under the curve
KS	Kolmogorov-Smirnov

BP	believe the positive
BN	believe the negative
RS	risk score
LOOCV	leave one out cross-validation
LPOCV	leave-pair-out cross-validation
LDA	linear discriminant analysis
AUC_{cv}	cross-validation estimation of the AUC
AUC_{bcv}	bootstrap cross-validation estimation of the AUC
AUC_{TLD}	true AUC from LDA
GLM	generalized linear model
CI	confidence interval
SE	standard error
SE_b	bootstrap standard error
SE_e	empirical standard error
$SE_{e,b}$	bootstrap empirical standard error
$SE_{e,cv}$	cross-validation empirical standard error
SE_{cv}	cross-validation standard error
PE	prediction error
TPE	true prediction error
TB-IRIS	tuberculosis immune reconstitution inflammatory syndrome

Introduction

1.1. Motivation and Purposes

A challenge in clinical medicine is that of correct diagnosis of disease. It is undesirable to declare someone as infected with a serious disease when in fact the individual is disease free and likewise undesirable not to declare someone as diseased when in fact the individual is diseased. Both errors have serious implications to the individual. Medical researchers invest considerable time and efforts to improve accurate disease diagnosis. The receiver operating characteristic (ROC) is a commonly used statistical tool for describing the discriminatory accuracy and performance of a diagnostic test (Pepe [47]). The ROC curve was first used in signal detection theory (Egan [21] and Green and Swets [27]). In the late 1980's, researchers began applying ROC curves methodology to medical diagnostic test evaluation (Hanley [31], Shapiro [51]). However the use of ROC curves in Radiology was earlier reported in the 1980s in a paper by Swets and Pickett [57]. In general the ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems (Swets [59]). A receiver operating characteristic (ROC) graph is a technique for visualizing and ranking classifiers based on their performance. It is a commonly used statistical tool for describing the discriminatory accuracy of a diagnostic test. In order to appropriately define the ROC curve in relation to disease diagnosis one needs to understand the difference between sensitivity and specificity of a test. Sensitivity is the probability that the test result is positive given the individual is truly diseased. Specificity is the probability that the test result is negative given the individual is truly disease free. Suppose the classification of a sample from an individual into diseased or disease free depends on a set threshold or cut-off value of a continuous biomarker. At each of these

cut-offs an estimate of the sensitivity and specificity of the test can be found. The ROC is a plot of sensitivity versus 1-specificity for different values of the cut-off points of the continuous biomarker.

Combining multiple biomarkers to estimate the AUC is of interest in the era of multiple assessments. When we have several biomarkers we can combine them to obtain better diagnostic accuracy and improve the AUC from all possible combinations (Fang et al., [22]). In this thesis we are interested in combining biomarkers to estimate the AUC. For this purpose we use Logistic regression as it is commonly used when the outcome or response is the presence is binary. In order to obtain the best (maximum) AUC we applied *feature selection*, also known as *variable selection*. This is a technique of selecting a subset of relevant features for building models and it improves model performance.

Resampling procedures are non-parametric inference methods based on generating repeated samples drawn from the original sample. They can be implemented computationally by simulating these new samples.

The *Cross-Validation* method is a standard tool for estimating prediction error and it is a specialized resampling procedure for application in model validation problems. It is mainly used in settings where the goal is prediction and one is interested to estimate how accurately a predictive model will perform in practice.

In 1979 Efron [18] introduced the *bootstrap* as a general method for estimating the sampling distribution of a statistic based on the observed data. This method is also used for assigning measures of accuracy to statistical estimates. Bootstrapping is accomplished by selecting with replacement n observations from among the original set of n observations (unlike in the cross-validation).

The main purpose of this thesis is to use the *resampling methods* (Cross-Validation and Bootstrap), discussed in Chapter 3, to estimate the AUC for procedures that select and combine biomarkers and also to make inference. We simulated several multivariate models to evaluate the performance of the resampling approaches in this setting. We applied the resampling methods to data collected from a study of tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS), as part of ongoing collaboration between my supervisors and Professor Robert Wilkinson from the Institute of Infectious Disease and Molecular Medicine (IIDMM), University of Cape Town. TB-IRIS occurs in 8 – 43% of HIV-infected patients receiving TB treatment after starting antiretroviral therapy (ART). Baseline levels of five biomarkers were evaluated and we used this dataset to investigate

whether a combination of these markers could accurately discriminate between IRIS and non-IRIS patients by applying the AUC analysis and resampling methods.

1.2. Background and Related Studies

The history of the ROC curves goes back to the second world war where it was firstly used in analyzing radar signals and later used in signal detection theory (see Fawcett [23] or Green and Swets [27]). Since then the usage and applications of ROC curves has spread to many other fields such as psychophysics, medicine (Hanley [31], Shapiro [51]), epidemiology (Aoki et al., [2]), radiology (Metz [44]), social sciences and evaluation of machine learning techniques (Spackman [55]). ROC analysis is a very rich area for research and a large number of articles have been published in the last two decades.

One of the earliest adopters of ROC graphs in machine learning was Spackman [55], who demonstrated the value of ROC curves in evaluating and comparing algorithms. Recent years have seen an increase in the use of ROC graphs in the machine learning community and for examining the effectiveness of diagnostic markers in distinguishing between diseased (D) and non-diseased (\bar{D}) individuals (Greiner et al., [28], Pepe [47], Shapiro [51] and Zhou et al., [67]). A diagnostic test result can be binary, ordinal or continuous. A binary test result simply provides the diagnosis as positive or negative. Ordinal and continuous tests provide measurements (on an ordinal or continuous scale). For instance, blood pressure, as an indicator of hypertension, serves as an example of a continuous marker. Ordinal markers are widely used in radiology for examining X-rays, where radiologists provide rankings corresponding to likelihood of disease.

The area under the ROC curve (AUC) is a popular measure to summarize the ROC curves in diagnostic testing. It is also used in non-diagnostic testing systems for example the use of *AUC* in clinical trials (Hauck [34]) and in toxicology (Bosch et. al., [24]).

Some experimental studies comparing different accuracy estimation methods have been previously done but most of them were on artificial or small datasets. We now describe some of these studies:

Dodd and Pepe [17] have proposed a new method for making inference about covariate effects on the performance of a classifier. Advantages of this approach are that “it can be simply applied by adapting standard binary regression methods, it requires fewer assumptions than existing ROC

regression methods”.

Zhang et.al., [66] considered clinical trials with two treatments and a non-normally distributed response variable.

The authors mentioned that the semi-parametric area under the ROC curve (AUC) regression model proposed by Dodd and Pepe [17] can be used. However, because a logistic regression procedure is used to obtain parameter estimates and a bootstrapping method is needed for computing parameter standard errors, their method may be cumbersome to implement. In [66] it is proposed to use a set of AUC estimates to obtain parameter estimates and combine DeLong’s method and the delta method for computing parameter standard errors. Their new method avoids heavy computation associated with the Dodd and Pepe’s method and hence is easy to implement.

It is of interest to estimate the AUC. The resampling methods, such as Cross-Validation and Bootstrap, can be used for this purpose.

Efron [19] conducted five sampling experiments and compared leave-one-out cross-validation, several variants of bootstrap, and several other methods. The purpose of the experiments was to investigate some related estimators, which seem to offer considerably improved estimation in small samples. The results indicate that leave-one-out cross-validation gives nearly unbiased estimates of the accuracy, but often with unacceptably high variability, particularly for small samples, and that the 632 bootstrap performed best.

Fang et. al., [22] considered the optimal linear combination that maximizes the AUC, this paper compared the estimating of the AUC associated with the estimated coefficients using Cross-Validation, Bootstrap and re-substitution methods. The authors recommended the Cross-Validation procedure, which works very well as an estimate for the AUC associated with the estimated coefficients.

1.3. Thesis Outlines

In this thesis we are mainly concerned with the ROC curves in the context of biomedical diagnostic testing and the computations of the area under the ROC curves (AUC). The thesis is divided into five chapters.

Chapter 1 is an introduction to the thesis, which is itself divided into three sections. In Sections 1.1 and 1.2 we introduce the purposes of the thesis, the ideas and background behind the ROC curves analysis. Section 1.3, which is the current one is to describe the structure of this thesis.

In Chapter 2 we discuss the concept of Receiver Operating Characteristic (ROC) Curves. This chapter is divided into four sections, where we first give some important definitions and basic concepts in Section 2.1. Section 2.2, which is mainly concerned with the ROC for continuous tests, is divided into three subsections, where in Subsection 2.2.1 we define the ROC curves and give an example for a ROC curve. In Subsection 2.2.2 we mention some properties and attributes of ROC curves. In Subsection 2.2.3 we introduce the *Binormal ROC Curve*, which is the classic form of ROC curve. In Section 2.3, we introduce four important indices of the ROC curves, where in Subsection 2.3.1, we discuss an important index, namely the *area under ROC curve*, denoted by AUC. In Subsection 2.3.2 we briefly review the ROC at a specific point ($ROC(t_0)$). In Subsection 2.3.3, we define the partial AUC, which is another important summary measure. Subsection 2.3.4 is devoted to the study of the Kolmogrov-Smirnov index. In Section 2.4, we give a motivation for combining multiple biomarkers as a better diagnostic test tool than a single biomarker on its own. We describe some methods for combining multiple biomarkers, namely the *Boolean combination* (Subsection 2.4.1) the *likelihood ratio* approach and risk score functions (Subsection 2.4.2) and the method based on *logistic regression* (Subsection 2.4.3).

In Chapter 3 we review resampling methods, namely *cross-validation*, *bootstrap* and *permutation test*. The chapter is composed of five sections. In Sections 3.1, 3.2 and 3.3 we discuss the methods of cross-validation, bootstrap and permutation test respectively. Section 3.4 is devoted to discussing the *Feature or Variable Selection* problem. Finally in Section 3.5 we discuss the resampling methods in the context of multiple biomarkers.

In Chapter 4 we are concerned with simulation studies and application to a real dataset. The chapter is divided into two main sections. Section 4.1 itself is divided into two subsections, where in Subsection 4.1.1 we explain how we generate different datasets from different settings. In Subsection 4.1.2 we used the free software R to obtain some results and we listed these results in Table 4.1. Section 4.2 is also divided into two main subsections where we apply the resampling methods discussed in Chapter 3 to a real dataset that has been collected from a study of tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS) at Cape Town.

Chapter 5 is a conclusion to the thesis and we suggest some of the future work that can be done as an extension to the current work.

Finally some R programs are supplied in the Appendix. We would also like to mention that 68 relevant references are listed under the Bibliography.

Receiver Operating Characteristic (ROC) Curves

We would like to mention that in most of the work on this chapter we follow mainly the book of Pepe [47] supplemented with our own understanding of the problem.

2.1. Definitions and Basic Concepts

In this section we quote some of the basic and important definitions and concepts that will be required throughout this thesis.

If a subject is classified as diseased or non-diseased and a test result as positive or negative, indicating the presence or absence of the disease, then there are four possible outcomes. These are

- when the test reports a positive result for a person who actually has the disease. We refer to this result as a *true positive* (TP),
- when the test reports a negative result for a person who actually is disease free. We refer to this result as a *true negative* (TN),
- when the test reports a positive result for a person who is disease free. We refer to this result as a *false positive* (FP),
- when the test reports a negative result for a person who actually has the disease. We refer to this result as a *false negative* (FN).

When a single test is performed, the person may in fact have the disease ($D = 1$) or the person may be disease free ($D = 0$). The test result may be positive ($Y = 1$), indicating the presence of disease, or the test result may be negative ($Y = 0$), indicating the absence of the disease. Using these actual disease status and test results variables, the previous four outcomes can be summarized in the following table.

	$D = 1$	$D = 0$
$Y = 1$	True Positives (TP)	False Positives (FP)
$Y = 0$	False Negatives (FN)	True Negatives (TN)

We define the *true positive and negative fractions* to be respectively $TPF = \frac{TP}{TP + FN}$ and $TNF = \frac{TN}{TN + FP}$. In Definition 2.1.1 we will refer to TPF and TNF as *sensitivity* and *specificity* respectively. These are original concepts in diagnostic testing literature and will be used throughout the development of this thesis.

Definition 2.1.1. *The **sensitivity** (true positive fraction TPF) is defined to be the probability that a test result will be positive when the disease is present in the individual, while the **specificity** (true negative fraction TNF) is defined to be the probability that a test result will be negative when the disease is not present.*

In probability notation the sensitivity and specificity are written respectively as

$$\begin{aligned}
 TPF &= P(Y = 1|D = 1) = TP/(TP + FN) \quad \text{and} \\
 TNF &= P(Y = 0|D = 0) = TN/(TN + FP).
 \end{aligned}$$

Sensitivity and specificity describe how well the test discriminates between patients with and without disease. In fact we are also interested to know the probability of disease, given a certain test result. This leads to the predictive values of the test.

Definition 2.1.2. *The **positive predictive value** PPV is defined as the probability that disease is present when the test is positive, while the **negative predictive value** NPV is defined as the probability that disease is not present when the test is negative.*

In probability notation, the PPV and NPV are written respectively as

$$\begin{aligned} PPV &= P(D = 1|Y = 1) = TP/(TP + FP) \quad \text{and} \\ NPV &= P(D = 0|Y = 0) = TN/(TN + FN). \end{aligned}$$

Definition 2.1.3. *The likelihood ratio LR is the probability of a given test result among people with a disease divided by the probability of that test result among people without the disease.*

In probability notation the LR is written as $P(Y = a|D = 1)/P(Y = a|D = 0)$, where $a = 0$ or 1 in the case of a binary test result.

Definition 2.1.4. *The positive likelihood ratio LR^+ is defined to be the ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease, while the negative likelihood ratio LR^- is defined to be the ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease.*

In probability notation the LR^+ and LR^- are written respectively as

$$\begin{aligned} LR^+ &= P(Y = 1|D = 1)/P(Y = 1|D = 0) \quad \text{and} \\ LR^- &= P(Y = 0|D = 1)/P(Y = 0|D = 0). \end{aligned}$$

Remark 2.1.1. Note that from Definitions 2.1.1 and 2.1.4 we obtain

$$LR^+ = \frac{\text{sensitivity}}{1 - \text{specificity}} \quad \text{and} \quad LR^- = \frac{1 - \text{sensitivity}}{\text{specificity}}.$$

2.2. Introduction to Receiver Operator Characteristic Curve (ROC) for Continuous Tests

2.2.1 Definition of ROC Curves

A *continuous test* means a test based on a continuous test variable or biomarker as a measure of presence of disease. For a threshold c , a binary test from the continuous test result Y is said to

be *positive* if $Y \geq c$ and *negative* if $Y < c$. The corresponding true and false positive fractions, at threshold c , are defined to be

$$TPF(c) = P[Y \geq c | D = 1], \quad (2.1)$$

$$FPF(c) = P[Y \geq c | D = 0], \quad (2.2)$$

respectively.

Definition 2.2.1. *The ROC curve is the set of all possible true and false positive fractions for Y for all c . That is to say*

$$ROC(.) = \{(FPF(c), TPF(c)) | c \in \mathbb{R}\}. \quad (2.3)$$

The *ROC* curve shows the trade off between specificity and sensitivity as the threshold for determining positivity varies.

Remark 2.2.1. Note that as c increases, both $FPF(c)$ and $TPF(c)$ decrease, while if c decrease, then both $FPF(c)$ and $TPF(c)$ increase. In the special cases if $c \rightarrow \infty$, then $\lim_{c \rightarrow \infty} FPF(c) = \lim_{c \rightarrow \infty} TPF(c) = 0$ and if $c \rightarrow -\infty$, then $\lim_{c \rightarrow -\infty} FPF(c) = \lim_{c \rightarrow -\infty} TPF(c) = 1$. Thus the ROC curve is a *monotone increasing* function in $(0, 1) \times (0, 1)$ (see Figure 4.1 of Pepe [47]).

The ROC curve can also be written in the form (see Pepe [47]):

$$ROC(.) = \{(t, ROC(t)) | t \in (0, 1)\}, \quad (2.4)$$

where $t \mapsto TPF(c)$, thus this defines the *ROC* function and c is the corresponding threshold to $FPF(c) = t$.

2.2.2 Properties and Attributes of ROC Curves

A test result is said to be *perfect* if $TPF(c) = 1$ and $FPF(c) = 0$ for some threshold c . Graphically, the diagnostic accuracy increases as its ROC curve approaches the left upper corner as shown in Figure 2.1.

On the other hand an *uniformative* test result is defined to be the test that does not separate between diseased and non-diseased subjects. That is $TPF(c) = FPF(c)$, $\forall c$. Graphically the

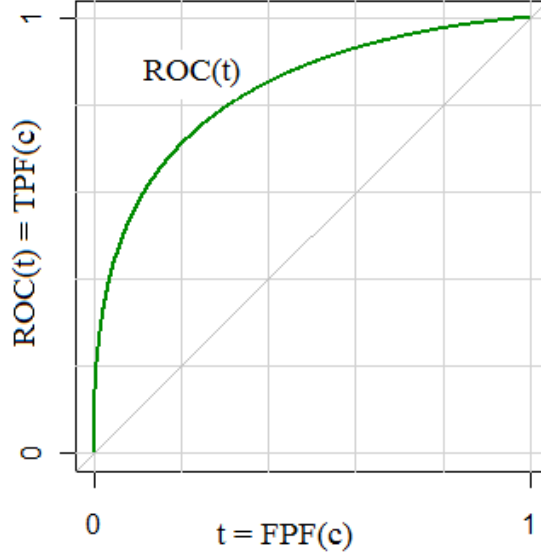


Figure 2.1: An example of a ROC curve

ROC curve of a uninformative test result is a straight line with slope 1 (i.e., the straight line joining the points $(0,0)$ and $(1,1)$).

In the following proposition we quote some important results from Pepe [47].

- Proposition 2.2.1.** (i) *The ROC curve is invariant to strictly increasing transformations of Y ,*
(ii) *if S_D and $S_{\bar{D}}$ denote the survivor function for Y in diseased and non-diseased populations, where $S_D(y) = P[Y \geq y | D = 1]$ and $S_{\bar{D}}(y) = P[Y \geq y | D = 0]$, then the ROC curve can be represented as follows:*

$$ROC(t) = S_D(S_{\bar{D}}^{-1}(t)), \quad t \in (0, 1), \quad (2.5)$$

- (iii) *with LR being the likelihood ratio, the optimal criterion based on Y for classifying subjects as positive for disease is $LR(Y) > c$, in the sense that it achieves the highest true positive fraction among all possible criteria based on Y with false positive fractions $t = P(LR(Y) > c | D = 0)$.*

PROOF. We only show (ii). For other statements see Results 4.1, and 4.4 of Pepe [47]. Now to show Equation (2.5), let $c = S_{\bar{D}}^{-1}(t)$, that is the corresponding $FPF = t$. Thus we have $P[Y \geq c | D = 0] = t$. The corresponding TPF is $P[Y \geq c | D = 1] = S_D(c)$. Therefore the TPF that corresponds to $FPF = t$ is $ROC(t) = TPF = S_D(c) = S_D(S_{\bar{D}}^{-1}(t))$. Hence the result. ■

We conclude this section by listing some of the important attributes of the ROC curves. These

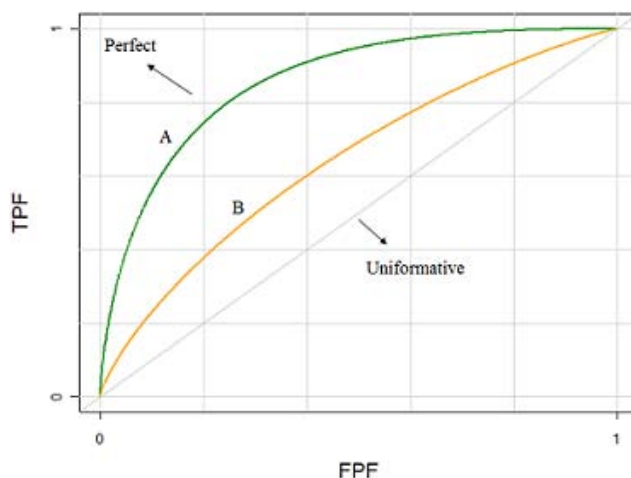


Figure 2.2: ROC curves for perfect, uninformative and two tests A and B . Test A is better than B attributes have been listed in Table 4.1 of Pepe [47] and in Fawcett [23]. In summary the ROC curve:

- Provides a tool for describing the test across a range of values and it is useful in early evaluation of tests when specific thresholds are unknown.
- Can be a useful guide for choosing thresholds in real applications.
- Is a useful mechanism for comparison between different non-binary tests, as it is scale invariant.

2.2.3 Binormal ROC Curves

The binormal ROC curve plays a major role in ROC analysis and it provides the classic model for ROC curves. Its form is derived from normal distributions for test results. To derive the functional form of binormal ROC curves, suppose that the test results are normally distributed in diseased and non-diseased populations.

Proposition 2.2.2. *Suppose that $Y_D \sim N(\mu_D, \sigma_D^2)$ and $Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}}^2)$. Then*

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)), \tag{2.6}$$

where $a = \frac{\mu_D - \mu_{\bar{D}}}{\sigma_D}$, $b = \frac{\sigma_{\bar{D}}}{\sigma_D}$ and Φ denotes the standard normal cumulative distribution function.

PROOF. Let c be any threshold. Then because of the symmetric nature of the normal distribution we have

$$\begin{aligned} FPF(c) &= P(Y_{\bar{D}} > c) = \Phi\left(\frac{\mu_{\bar{D}} - c}{\sigma_{\bar{D}}}\right), \\ TPF(c) &= P(Y_D > c) = \Phi\left(\frac{\mu_D - c}{\sigma_D}\right). \end{aligned}$$

For FPF we can see that $c = \mu_{\bar{D}} - \sigma_{\bar{D}}\Phi^{-1}(t)$. Thus

$$\begin{aligned} ROC(t) = TPF(c) &= \Phi\left(\frac{\mu_D - c}{\sigma_D}\right) \\ &= \Phi\left(\frac{\mu_D - \mu_{\bar{D}} + \sigma_{\bar{D}}\Phi^{-1}(t)}{\sigma_D}\right) \\ &= \Phi(a + b\Phi^{-1}(t)) \end{aligned}$$

completing the proof. ■

The *binormal ROC curve* is defined to be $ROC(t) = \Phi(a + b\Phi^{-1}(t))$. The coefficients a and b are referred to as the *intercept* and the *slope* of the binormal ROC curve respectively.

Remark 2.2.2. Note that the slope of the ROC curve at t is the likelihood ratio at the corresponding threshold c .

Now

- if $b = 1$, then the binormal ROC curve is concave everywhere,
- if $b > 1$, then the likelihood ratio decreases and then increases,
- if $b < 1$, then the likelihood ratio increases and then decreases, as $t \in (0, 1)$.

When $b \neq 1$, this leads to anomalies in the ROC curves. Thus the fact that the binormal ROC curve does not have the monotone likelihood ratio raises some concern about using it for approximation of real data. However Swets [58] and Hanley [32] and [33] showed that a binormal ROC curve is a good approximation in practice.

We have seen in Proposition 2.2.1 that the ROC is invariant to monotone increasing data transformations. Therefore if Y_D and $Y_{\bar{D}}$ have normal probability distributions and if we let $W_D = h(Y_D)$ and $W_{\bar{D}} = h(Y_{\bar{D}})$, where $h(\cdot)$ is a monotone strictly increasing function, then the ROC curve for W_D

and $W_{\overline{D}}$ is a binormal curve given by $ROC(t) = \Phi(a + b\Phi^{-1}(t))$. Conversely, to say that the ROC curve for Y_D and $Y_{\overline{D}}$ is binormal is simply to say that for some strictly increasing transformation $h(\cdot)$, the functions $h(Y_D)$ and $h(Y_{\overline{D}})$ have normal distributions (see Pepe [47]).

We conclude by mentioning that the binormal assumption states that some monotone transformation of the data exists to make Y_D and $Y_{\overline{D}}$ normally distributed and this can be taken as a weak assumption.

2.3. Some of ROC Curves Indices

In this section we briefly go over some of the ROC indices, which provide important information about the ROC curves. Many indices have been developed in the literature and are used in various applications, for example see Shapiro [51], Greiner et al., [28], Zhou et al., [67] and Pepe [47].

2.3.1 Area Under ROC Curves (AUC)

While the ROC curve contains most of the information about the accuracy of a continuous marker, we may want to reduce ROC performance to a single scalar value representing expected performance. The most commonly used global index is the *area under the ROC curve (AUC)*. It is a convenient way of comparing markers. For continuous markers the AUC is defined as

$$AUC = \int_0^1 ROC(t)dt. \tag{2.7}$$

We note from Equation (2.7) that the AUC is a portion of the area of the unit square. Hence its value will always be between 0 and 1. Values of AUC close to 1 indicate that the marker has high diagnostic accuracy and a test is called *perfect* if its $AUC = 1$, while a test is called an *uninformative* if its $AUC = 0.5$. AUCs less than 0.5 may suggest the scale needs transformation so that increasing values indicate increasing likelihood of diseased.

Definition 2.3.1. *Let A and B be two tests. We say that A is **better** than B if*

$$ROC_A(t) \geq ROC_B(t), \forall t \in (0, 1).$$

Proposition 2.3.1. *Let A and B be two tests such that A is better than B . Then*

$$AUC_A \geq AUC_B.$$

Remark 2.3.1. The converse of Proposition 2.3.1 is not necessarily true. For example it may be the case that for some number $k \in (0, 1)$, we have

$$ROC_A(t) \geq ROC_B(t), \forall t \in (0, k] \text{ and } ROC_B(t) \geq ROC_A(t), \forall t \in [k, 1).$$

Thus $\forall t \in (0, k]$ test A is better than B and $\forall t \in [k, 1)$ test B is better than A

The AUC has an interesting statistical interpretation (Bamber [6], Hanley and McNeil [31], Pepe [47]). It is equal to the probability that a test result chosen randomly from diseased subjects is greater than a test result chosen randomly from non-diseased subjects. In general

$$AUC = P(Y_D > Y_{\bar{D}}) + \frac{1}{2}P(Y_D = Y_{\bar{D}}).$$

For a continuous test we have $P(Y_D = Y_{\bar{D}}) = 0$. Thus the AUC for a continuous test will have the form

$$AUC = P(Y_D > Y_{\bar{D}}).$$

To show the above we have

$$\begin{aligned} AUC &= \int_0^1 ROC(t)dt = \int_0^1 S_D(S_{\bar{D}}^{-1}(t))dt \\ &= \int_{-\infty}^{-\infty} S_D(y)dS_{\bar{D}}(y) \\ &= \int_{-\infty}^{\infty} P(Y_D > y)f_{\bar{D}}(y)dy \\ &= \int_{-\infty}^{\infty} P(Y_D > y, Y_{\bar{D}} = y)dy \\ &= P(Y_D > Y_{\bar{D}}) \end{aligned}$$

by change of variable from t to $y = S_{\bar{D}}^{-1}(t)$, where $f_{\bar{D}}$ denotes the probability density function of $Y_{\bar{D}}$ and independence of Y_D and $Y_{\bar{D}}$, we can write the AUC in the form above.

The interpretation of AUC as probability of correctly ordering the diseased and non-diseased subjects is an interesting result but it does not provide the best interpretation of this important measure. We thus can regard the AUC as an average of TPF , averaged uniformly over the whole range of FPF in $(0, 1)$. Dodd [14] suggested the use of a weighted average approach, weighting certain parts of FPF domain more than others.

The AUC for the binormal ROC curve is given by

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

PROOF. Recall that $AUC = P(Y_D > Y_{\bar{D}}) = P(Y_D - Y_{\bar{D}} > 0)$. Let $W = Y_D - Y_{\bar{D}}$ then $W \sim N(\mu_D - \mu_{\bar{D}}, \sigma_D^2 + \sigma_{\bar{D}}^2)$ and

$$\begin{aligned} p(W > 0) &= 1 - \Phi\left(\frac{-\mu_D + \mu_{\bar{D}}}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}}\right) \\ &= \Phi\left(\frac{\mu_D - \mu_{\bar{D}}}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}}\right) \\ &= \Phi\left(\frac{\mu_D - \mu_{\bar{D}}}{(\sigma_D)} / \sqrt{1 + \frac{\sigma_{\bar{D}}^2}{\sigma_D^2}}\right) \\ &= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \end{aligned}$$

which completes the proof. ■

It is interesting to note that the area under the empirical ROC curve is the Mann-Witney U-statistic.

That is

$$AUC_e = \sum_{j=1}^{n_{\bar{D}}} \sum_{i=1}^{n_D} I[Y_{D_i} > Y_{\bar{D}_j}] + \frac{1}{2} I[Y_{D_i} = Y_{\bar{D}_j}] / n_D n_{\bar{D}}. \quad (2.8)$$

2.3.2 The $ROC(t_0)$

If we are interested in a specific FPF value say t_0 , then the corresponding TPF value $ROC(t_0)$ provides relevant summary index.

We can interpret $ROC(t_0)$ as a proportion of diseased subjects that have test results greater than $1 - t_0$ quantile for non diseased observations. If t_0 is small then the $ROC(t_0)$ is interpreted as the proportion of diseased subjects with test result values above the normal range.

One of the restrictions of $ROC(t_0)$ is that it does not give all the information as the $ROC(t)$. For two tests A and B such that $ROC_A(t_0) = ROC_B(t_0)$, if $ROC_A(t) \geq ROC_B(t)$ for any $t \in (0, t_0)$, then it is obviously that test A is better than test B with regard to the overall performance.

2.3.3 Partial AUC

The *partial area under the curve* $pAUC(t_0)$ is defined to be

$$pAUC(t_0) = \int_0^{t_0} ROC(t)dt. \quad (2.9)$$

It is a measure concerned with the values of $FPF \in (0, t_0)$ and it uses all points on $(ROC(0), ROC(t_0))$. A lower bound for $pAUC$ is $\frac{t_0^2}{2}$ and this happens when the test is uninformative ($TPF(c) = FPF(c)$ for all thresholds c). An upper bound for $pAUC$ is t_0 and this happens when the test is perfect.

The normalized value of $pAUC$ is defined to be $pAUC(t_0)/t_0$ and it is clearly that it ranges from $t_0/2$ to 1 for uninformative and perfect tests respectively. The normalized $pAUC$ can be interpreted as

$$\frac{pAUC(t_0)}{t_0} = P[Y_D > Y_{\bar{D}} | Y_{\bar{D}} > S_{\bar{D}}^{-1}(t_0)].$$

That is to say it is the probability of correctly ordering a diseased and non-diseased observation selected randomly given that the non-diseased observation is above $1 - t_0$ quantile of the non diseased distribution.

More general formula for Equation (2.9) has been given in Dodd and Pepe [16].

2.3.4 Kolmogorov-Smirnov (KS)

The maximum vertical distance between the ROC curve and $TPF = FPF$ is an index we refer to it as KS. We have

$$KS = \max_t |ROC(t) - t| = \max_t |S_D(S_{\bar{D}}^{-1}(t)) - t| = \sup_{c \in (-\infty, \infty)} |S_D(c) - S_{\bar{D}}(c)|.$$

We can see that this is exactly the Kolmogorov-Smirnov measure, which measures the distance between two distributions S_D and $S_{\bar{D}}$ of two tests Y_D and $Y_{\bar{D}}$ respectively (Gail and Green [26]). In fact we identify the index KS with Kolmogorov-Smirnov measure.

Another well-known measure is Youden index, which is a special case of Kolmogorov-Smirnov measure. For more information on this index, refer to Fluss [25].

2.4. Motivation for Combining Multiple Biomarkers

In this thesis we are mainly interested in combining multiple biomarkers since this combination may possess a better diagnostic accuracy than any single test on its own. For example, a single biomarker may not give sufficient sensitivity and specificity in the study of a population with ovarian cancer. However, combinations of biomarkers may.

Methods for combining multiple biomarkers can be found in Pepe [47]. Here we go briefly over some of these methods.

2.4.1 Boolean Combinations

This method is used when the predictors are binary. Suppose that there are P predictors $\{Y_1, Y_2, \dots, Y_P\}$. For example each of these predictors can be described as either present or absent. These predictors are combined by using the conjunctions “AND”, “OR” and “NOT”. Since each predictor Y_j for some $j \in \{1, 2, \dots, P\}$ has two possibilities, there will be 2^P combinations in total. Thus if $P = 2$, there are four possibilities namely $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$. The decision rules are then classified as the “Believe the positive” (BP) and “Believe the negative” (BN) rules (Marshall [43]) as defined below.

Definition 2.4.1. *For two binary tests Y_1 and Y_2 we define the “Believe the positive” (BP) to be that if Y_1 OR Y_2 is positive, then the subject is classified as diseased, while we define the “Believe the negative” (BN) to be that if Y_1 AND Y_2 are positive, then the subject is classified as diseased free.*

In general the BP test is more sensitive and less specific than all the component tests, while the BN test is more specific and less sensitive than each test.

We remark that as the number of predictors P get larger, the Boolean combination becomes more complex since the 2^P combinations become large.

2.4.2 The Likelihood Ratio Method (LR)

In Definition 2.1.3 we defined the likelihood ratio. The likelihood transformation is an interesting tool when we want to combine multiple test results $Y = \{Y_1, Y_2, \dots, Y_P\}$. First, we must provide some definitions. The following is a central result that plays an important role in finding the optimal combination, Pepe [47]. The $LR(Y)$ with $Y = \{Y_1, Y_2, \dots, Y_P\}$ gives all optimal decision rules based on P tests. That is $LR(Y) > c$ will

- maximize the sensitivity among all possible criteria based on Y with $FPF = t, \forall t \in (0, 1)$
- maximize the specificity among all possible criteria based on Y with $TPF = r, \forall r \in (0, 1)$
- minimize the overall misclassification probability

where the chosen c depends on an optimality criterion.

Although the LR is a useful tool for combining multiple tests, it is still not easy to apply in practice as the multivariate probability distribution for Y must be known in advance, which is not easy to calculate.

Next we introduce a principle that is equivalent to LR function. We refer to this principle as the *risk score* $RS(Y)$, which is commonly estimated with data.

Proposition 2.4.1. *The ROC curve based on $RS(Y)$ is the same as $LR(Y)$.*

PROOF. To show that $RS(Y)$ is a monotone increasing function of $LR(Y)$ it is sufficient to prove that the $RS(Y)$ and $LR(Y)$ have the same ROC curve. Now

$$\begin{aligned}
 RS(Y) &= P[D = 1|Y] \\
 &= \frac{P[Y|D = 1]P[D = 1]}{P[Y]} \\
 &= \frac{P[Y|D = 1]P[D = 1]}{P[Y|D = 1]P[D = 1] + P[Y|D = 0]P[D = 0]} \\
 &= \frac{LR(Y)P[D = 1]}{LR(Y)P[D = 1] + P[D = 0]},
 \end{aligned}$$

which is a monotone increasing function in $LR(Y)$. Hence the result ■

We conclude by mentioning that to obtain the optimal combination of a multiple test results, we should ascertain the RS function or alternatively use any monotone increasing function of the RS.

2.4.3 Logistic Regression

Logistic regression is commonly used when the outcome or response is the presence or absence of a condition, often a disease. In these cases, the explanatory variable is often a test or procedure used to detect this condition. Logistic regression allows us to convert these agreement proportions into probabilities of having the disease. In addition, these probabilities can be converted into sensitivity and specificity which can be used to determine the accuracy of a procedure or test in successfully predicting the absence or presence of a condition. A mathematical formula for the logistic regression is given by

$$\log(P(D = 1|Y)/1 - P(D = 1|Y)) = \beta_0 + \sum_{i=1}^p \beta_i Y_i.$$

Resampling Methods

Resampling procedures are statistical inference methods based on generating repeated samples drawn from the original sample. They can be implemented computationally by simulating these new samples.

In this chapter we review and discuss methods of resampling, namely *cross-validation*, *bootstrap* and *permutation test methods*. Much of this is described by the Efron and Tibshirani [20]. In our applications in Chapter 4 we use the cross-validation and bootstrap methods and their combinations.

3.1. Cross-Validation

Cross-Validation is a standard tool for estimating prediction error and it is a specialized resampling procedure that is designed specifically for application in model validation problems. It is mainly used in settings where the goal is prediction and one wants to estimate how accurately a predictive model will perform in practice. It can be used to estimate the error of a given model as a basis for model selection by choosing one of several models that has the smallest estimated prediction error. Cross-validation is important especially in cases where further samples are costly or impossible to collect.

Cross-validation is accomplished by implementing the following steps.

- leaving out a portion of the sample

- building the prediction rule on the remaining sample (training set)
- predicting the class labels of the left out (test set) sample.

Types of Cross-Validation

K -fold Cross-Validation

K -fold cross-validation can be summarized in the following steps:

- Split the full dataset into K randomly equal sized subsets. Keep one of them for testing the model and use the other $K - 1$ parts as training data.
- Fit the model for the $K - 1$ parts included and calculate the prediction error of the fitted model when predicting the k^{th} part of the data left out.
- repeat the above step for all $k = 1, 2, \dots, K$ and average the K results from the K -fold prediction.

The advantage of this method is that all observations are used for both training and testing and each observation is used for testing exactly once and used for training $K - 1$ times. Note that the variance of the resulting estimate is reduced as K increased. On the other hand the training algorithm has to be rerun from scratch K times.

The simplest case of K -fold cross-validation is when $K = 2$ (*2-fold Cross-Validation*). For each fold, we randomly assign data points to two sets, so that both sets are equal size. We then train on the first set and test on the second set, followed by training on the second set and testing on the first set. This has the advantage that each data point is used for both training and validation on each fold.

Note 3.1.1. Prediction error, PE , is a quantity that measures how well the model predict the response value of a future observation. It is often used for model selection since it is sensible to choose a model that has the lowest prediction error among a set of candidates [20]. In regression models it refers to the expected squared difference between a future response and its prediction from the model that is $PE = E(y - \hat{y})^2$.

Leave One Out Cross-Validation (LOOCV)

This is same as the K -fold cross-validation with K being equal to the number of observations in the original sample. We use a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. Leave one out Cross-Validation is a common choice for small sample sizes. It is accomplished through the following steps:

- The full dataset is divided into training and test (validation) sets. The test set contains a single observation.
- The prediction rule is built from scratch using the training set.
- The rule is applied to the observation in the test set for class prediction.
- The process is repeated until each observation has appeared once in the test set.

Repeated Random Subsampling Validation

This method randomly splits the data set into training and validation data. For each such split, the model is fit to the training data and predictive accuracy is assessed using the validation data. The results are then averaged over all the splits. The advantage of this method is that the proportion of the training/validation split is independent from the number of folds. On the other hand some observations may never be selected in the validation subsample, whereas others may be selected more than once. Cross-Validation can be used to provide an improved estimate of prediction error, but for variance estimation we use bootstrapping (with LOOCV as an internal loop).

3.2. Bootstrap Method

In 1979 Efron [18] introduced the *bootstrap* as a general method for estimating the sampling distribution of a statistic based on the observed data.

Bootstrapping can be used to estimate measures of accuracy to statistical estimates. Bootstrap estimation of the *true error rate* (See Simon [52]) is an alternative to cross validation. Bootstrapping is accomplished by selecting with replacement n observations from among the original set of n

observations (unlike in the cross-validation). With bootstrapping the original sample could be duplicated as many times as computing resources allow. Also every resample has the same number of observations as the original sample. Thus the bootstrap method has the advantage of modeling the impact of the actual sample size. It should be noted that a predictive model is developed from scratch, this includes the variable selection step with each bootstrap replicate. The model is then used to predict the class for each observation not in the bootstrap sample. Each prediction is recorded as correct or incorrect. This process is repeated for many bootstrap samples and the average number of misclassification per prediction is used as an estimate of the misclassification rate (Simon [52]).

We remark that bootstrap estimates have smaller variances, especially for small sample sizes. An estimate θ_B^* , $b = 1, 2, \dots, B$ of the parameter of interest θ is calculated from each pseudo sample. Then an estimate of the variance of the parameter of interest is calculated as follows for the bootstrap:

$$Var_{BS}(\theta) = \frac{1}{B-1} \sum_{b=1}^B (\theta_b^* - \theta^*)^2$$

where B is the number of replicate samples and $\theta^* = \frac{1}{B} \sum_{b=1}^B \theta_b^*$.

It has been suggested that the number of replicate samples needs to be large. Efron[20] stated that a large B would be 200 replicates, however if confidence intervals will be calculated then it has been suggested that B needs to be 1000. Generally variance decreases as sample size increases. For this reason we use 1000 bootstrap replicates in both simulation studies and application to real dataset.

In this thesis we calculated AUC for combined biomarkers by using bootstrap Cross-Validation method. Here we built a Leave One Out Cross-Validation (LOOCV) function and used this function inside bootstrap loop. That is to say we used an outer bootstrap loop with Cross-Validation as an inner loop. The LOOCV split the full dataset into *training* and a *test*, which contains a single observation. The training set consists of the other remaining observations. For the training set, we build our predictive models and we would like to mention that a variable selection procedure should be used within Cross-Validation loop. We then predicted the part of the data left out. This gives a value between 0 and 1. The process was repeated until all the possible sets are selected. Finally we used these values to estimate the AUC To obtain estimation of bootstrap Cross-Validated AUC, we drew a bootstrap sample and performed AUC Cross-Validation on the bootstrap sample. We used

1000 bootstrap replicates. The purpose of using bootstrap method is to get the variance estimates for Cross-Validated AUC.

3.3. Permutation Test

Permutation tests are computer intensive statistical methods that predate computers. They are a subset of non-parametric statistics methods and were first introduced by R. A. Fisher in 1930s. They are attractively simple and have close connection with bootstrap.

Permutation tests are called resampling tests and they are considered a special case of nonparametric tests. Nonparametric test statistics do not rely on a specific probability distribution that describes the underlying population. However, permutation tests are not quite distribution free. Some underlying assumptions are required with respect to the samples, namely (exchangeability) LaFleur [38].

Note 3.3.1. A collection of random variables X_i , $i \in \{1, 2, \dots\}$ is exchangeable if for every finite subset of random variables $X_{i_1}, X_{i_2}, \dots, X_{i_N}$ and every permutation operator $\rho(\cdot)$, which rearranges any set of integers, the joint probability distribution of the permuted sequence $X_{\rho(i_1)}, X_{\rho(i_2)}, \dots, X_{\rho(i_N)}$ is the same as joint probability distribution of the original sequence. We also remark that independent and identically distributed random variables are exchangeable.

The main application of permutation tests is the *two-sample problem*, which is defined next. Let F and G be the probability density functions of two independent random samples \mathbf{x} and \mathbf{y} respectively, where

$$\mathbf{x} = \{x_1, x_2, \dots, x_n\} \quad \text{and} \quad \mathbf{y} = \{y_1, y_2, \dots, y_m\}.$$

In other words \mathbf{x} is a random sample of size n from F and \mathbf{y} is a random sample of size m from G , where \mathbf{x} and \mathbf{y} are independent of each other. This setup is called the *two-sample problem*.

The permutations often are referred to as the physical act of permuting subject labels (eg. group 1 or 2 in the above example) and the permutation tests can be accomplished as follows:

- Data from an experiment are tested using some pre-specified test statistic.
- The test statistic is generated for the original sample of the data (observed permutations).

- The results are saved and data permutations are then generated.
- A test statistic is then calculated for each of the permutations and compared against the test statistic based on the original data.

We permute (rearrange) the data by shuffling their labels and then calculate the test statistic on each permutation. The collection of test statistic from the permuted data constructs the distribution under the null hypothesis, which states that there is no difference between the two density functions, that is $F = G$.

The permutations either can be all possible permutations or a random sample of all possible permutations. The permutation tests based on a random sample of permutations is still exact, in the sense that the significance level of the test is equal to the false rejection rate. However, random permutations can be less powerful than all possible permutations and increasing the number of permutations improves the approximation to the exact Type I error.

Permutation tests are therefore used when assumptions for parametric tests cannot be met or when an exact test is desired. They are often as powerful as the unbiased parametric test when sample sizes are small.

Permutation tests are restricted to testing under the null hypothesis and gives a simple way to compute the sampling distribution for test statistics under the null hypothesis. To estimate the sampling distribution of the test statistic we generate many samples under the strong null hypothesis. The null hypothesis was defined above.

A commonly used example of permutation test is the Fisher's exact test, which is used for evaluating the association between two dichotomous variables.

3.4. Feature Selection

Feature selection, also known as variable selection or feature reduction is the technique of selecting a subset of relevant features for building models. Variable and feature selection have become the focus of much research when tens or hundreds of thousands of variables are available. Such datasets are common in gene expression array studies. Feature selection is the common first step when developing a class predictor based on microarray data (Simon [52]). In fact it is reasonable

to assume that only some subset of many of measured genes contribute useful information for distinguishing the classes. By removing most redundant genes (variables) from the data, feature selection helps improve the performance of models. The objective of variable selection is to avoid overfitting, improve model performance and providing faster and more effective predictors.

One approach to feature selection, is to select variables based on their statistical significance in univariate tests of differences between the classes. For this purpose t -test or Wilcoxon rank-sum test can be used to assess univariate statistical significance (Simon [52]). Then those variables considered statistically significant are to be selected selected for inclusion in the multivariate model.

Various selection methods such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are available. The AIC is a measure of the goodness of fit of an estimated statistical model. The AIC is an operational way of assessing the trade off between the complexity of an estimated model against how well the model fits the data. The preferred model is the one with the lowest AIC value.

The most commonly used methods for variables selection are *backward elimination*, *forward selection*, and *stepwise selection*. We briefly summarize these methods.

Backward elimination begins with a full model consisting of all candidate predictor variables. Variables are sequentially eliminated from the model until a predefined stopping rule is satisfied. The variable whose elimination would result in the smallest decrease in a summary measure is eliminated. A common stopping rule is to stop when all variables that remain in the model are significant at a pre-specified significance level.

Forward selection begins with the empty model. Variables are added sequentially to a model until a pre-specified stopping rule is satisfied. At a given step of the selection process, the variable whose addition would result in the greatest increase in the summary measure is added to the model. A typical stopping rule is that if any added variable would not be significant at a pre-specified significance level, then no further variables are added to the model.

Stepwise regression is a standard procedure for variable selection, which is based on the procedure of sequentially introducing the predictors into the model one at a time. Stepwise selection is a variation of forward selection. At each step of the variable selection process, after a variable has

been added to the model, variables are allowed to be eliminated from the model. For instance, if the significance of a given predictor is above a specified threshold, it is eliminated from the model. The iterative process is ended when a pre-specified stopping rule is satisfied.

3.5. Resampling Methods in the Context of Combining Multiple Biomarkers

Diagnostic tests are important components in modern medical practice. Recall from Chapter 1 that the ROC curve is a graphical tool for evaluating the discriminatory accuracy of diagnostic tests and the AUC is the most popular summary index of discriminatory accuracy. When we have several biomarkers, we can combine them to obtain better diagnostic accuracy and maximum AUC overall possible combinations (Fang et al., [22]).

As pointed out in Fang et al., [22], the procedure of combining multiple test results has been well studied. For example, Su and Liu [56] discussed the optimal linear combination under the multiple-normal assumption; Pepe and Thompson [46]; Pepe, Cai, and Longton [48]; and Ma and Huang [30] discussed this procedure under the generalized linear model (GLM) assumption. Copas and Corbett [11] addressed the *overfitting problem* (using the same data both to fit the score and to calculate its ROC tends to give an over optimistic estimate of the performance of the score) when combining tests through logistic regression. In this thesis we use logistic regression, which is often used to find a linear combination of covariates that best discriminates between two populations.

The purpose of this section is to use the resampling methods that have been introduced in the previous sections (cross-validation, bootstrap and permutation test) to estimate the AUC in the context of variable selection.

Cross-Validation is the simplest and most widely used method for estimating the predictor error. As we have indicated earlier that if many diagnostic tests are available and some of them are redundant, then we want to seek an optimal subset of diagnostic tests that the combined test has the largest AUC. Thus for each subset of diagnostic tests we calculate the cross-validation estimation of the AUC and then we choose the subset of diagnostic tests, which gives the largest or maximizes the cross-validated AUC as the best one. We remark that including the redundant diagnostic tests in the combination will decrease the AUC. This gives rise to the *variable selection problem* (see Efron and Tibshirani [20]).

In this thesis we used Leave one out cross-validation (LOOCV) as it is nearly unbiased, easy to implement and to understand. In LOOCV the classifier is trained with all but one observation and test on the observation that left out. This process repeated until every observation is tested once. The tested observations are pooled together to estimate the AUC. Formally, the AUC is calculated with LOOCV as

$$\frac{1}{|X^+||X^-|} \sum_{x_i \in X^+} \sum_{x_j \in X^-} H(C_{\{i\}}(x_i) - C_{\{j\}}(x_j)),$$

where H is the Heaviside step function defined by

$$H(x) = \begin{cases} 1 & \text{if } x \geq 0.5, \\ 0 & \text{if } x < 0.5, \end{cases}$$

$C_{\{i\}}$ and $C_{\{j\}}$ denote classifiers trained without the i^{th} and j^{th} respectively and $X^+ \subset X$ and $X^- \subset X$ denote the positive and negative samples in the training set X respectively.

3.5.1 Algorithm to Obtain the AUC Through Cross-Validation

The following is a Leave One Out Cross-Validation algorithm.

- The full dataset is divided into training and test (validation) sets. The test set contains a single observation.
- For the training set, feature selection is performed from scratch to build a predictive model.
- Predicting the part of the data left out.
- This gives a value in (0,1) for each subjects.
- The process is repeated until all the possible sets are selected.
- These values can be used to estimate the AUC.

3.5.2 Algorithm to Estimate the Variance of AUC Through Bootstrapping

The Bootstrap is used to obtain variance estimates of the cross-validated AUC. The following is the procedure to do this.

- Draw a bootstrap sample, stratifying by disease status
- Perform cross-validation as described in previous algorithm on the bootstrap sample.
- Estimate the SE of the cross-validated AUC based on the bootstrap replicates.

Simulation Studies and Application

In this chapter the proposed methods introduced in Chapter 3 (Cross-Validation and Bootstrap) are illustrated by some simulation studies and application to real dataset. We first start by presenting simulation studies.

4.1. Simulation Studies

Simulation studies use computer intensive procedures to evaluate particular hypotheses and assess the appropriateness of a variety of statistical methods, under specific models where the truth is known. These techniques provide empirical estimation of the sampling distribution of the parameters of interest that could not be achieved from a single study. Simulation studies are increasingly being used in the era of increased computational resources. In addition, simulations can be used as instructional tools to help with the understanding of many statistical concepts.

4.1.1 Introduction to Simulated Data

In this section we are mainly concerned with examining the performances of different methods, with particular interest to Cross-Validation and Bootstrap Cross-Validation, as methods for the estimation of the AUC and its variance, in our context. We simulate datasets under the following group settings: Assume that there are k diagnostic tests (corresponding to k biomarkers) X_1, X_2, \dots, X_k . In our case, we let $k = 5$, that is five biomarkers X_1, X_2, X_3, X_4 and X_5 . Denote the mean vector of the k biomarkers in diseased and non-diseased by μ_k^D and $\mu_k^{\bar{D}}$ respectively.

With the above settings, the biomarker outcomes y_{ik}^D and $y_{jk}^{\bar{D}}$ for diseased and non-diseased populations respectively are given by

$$y_{ik}^D = \mu_k^D + a_i^D + \varepsilon_{ik}^D$$

and

$$y_{jk}^{\bar{D}} = \mu_k^{\bar{D}} + a_j^{\bar{D}} + \varepsilon_{jk}^{\bar{D}},$$

where the notation and assumptions in our case are

- n (resp. m) is the number of individuals from diseased (resp. non-diseased) population and i (resp. j) is in the set $\{1, 2, \dots, n\}$ (resp. $\{1, 2, \dots, m\}$),
- k is number of biomarkers,
- a_i^D (resp. $a_j^{\bar{D}}$) is the subject specific random effect with normal distribution $a_i^D \sim N(0, 0.5)$ (resp. $a_j^{\bar{D}} \sim N(0, 0.5)$),
- ε_{ik}^D (resp. $\varepsilon_{jk}^{\bar{D}}$) is random error effect with normal distribution $\varepsilon_{ik}^D \sim N(0, 0.25)$ (resp. $\varepsilon_{jk}^{\bar{D}} \sim N(0, 0.25)$).

The outcomes y_{ik}^D and $y_{jk}^{\bar{D}}$ are generated from three multivariate normal distributions with means being $\mu_k^D = (0.5, 0.25, 0, 0, 0)$ and $\mu_k^{\bar{D}} = (0, 0, 0, 0, 0)$ for three group settings defined by three different variance-covariance matrices. The three variance-covariance matrices are as follow:

- For the first setting we assume independence for the biomarkers and consequently we will have variances in the main diagonal and zeros elsewhere. In this case the resulting variance–covariance matrix will have the form:

$$\Sigma_1 = \begin{pmatrix} 0.75 & 0 & 0 & 0 & 0 \\ 0 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.75 & 0 & 0 \\ 0 & 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0 & 0 & 0.75 \end{pmatrix}$$

- For the second setting we add dependence for the biomarkers with mean shift for diseased (X_3, X_4 and X_5). In this case the resulting variance–covariance matrix will have the form:

$$\Sigma_2 = \begin{pmatrix} 0.75 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0.75 & 0 & 0 \\ 0 & 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0 & 0 & 0.75 \end{pmatrix}$$

- For the third setting we assume the same dependence across all the biomarkers. The resulting exchangeable or compound symmetry variance–covariance matrix is:

$$\Sigma_3 = \begin{pmatrix} 0.75 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.75 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.75 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.75 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.75 \end{pmatrix}.$$

We briefly explain how to derive the third matrix. Suppose that $y_{ik} = \mu_k + a_i + \varepsilon_{ik}$. If we assume that $Cov(\varepsilon_{ik}, \varepsilon_{il}) = 0$ for any different biomarkers k and l and that $Cov(\varepsilon_{ij}, a_i) = 0$ then

$$\begin{aligned} Var(y_{ik}) &= var(a_i) + var(\varepsilon_{ik}) \\ Var(y_{ik}) &= \sigma_a^2 + \sigma_\varepsilon^2 \end{aligned}$$

and for $\ell \neq k$ the covariance between y_{ik} and $y_{i\ell}$ is

$$\begin{aligned} Cov(y_{ik}, y_{i\ell}) &= Cov(a_i + \varepsilon_{ik}, a_i + \varepsilon_{i\ell}) \\ &= E[(a_i + \varepsilon_{ik})(a_i + \varepsilon_{i\ell})] \\ &= E[a_i^2 + a_i\varepsilon_{ik} + a_i\varepsilon_{i\ell} + \varepsilon_{ik}\varepsilon_{i\ell}] \\ &= E[a_i^2] \\ &= \sigma_a^2. \end{aligned}$$

We remark that the same variance-covariance matrix was used for diseased and non-diseased subjects in all previous settings.

4.1.2 Simulations Methods and Results

We are interested in combining biomarkers to obtain better diagnostic accuracy and AUC estimation. For this purpose we used Logistic Regression as it is commonly used when the outcome or response is the presence or absence of a condition, often a disease.

Biomarkers from simulated datasets were used to evaluate whether a combination of these biomarkers can accurately discriminate between two groups of diseased and non-diseased after applying logistic regression for biomarkers combination and resampling methods.

An original R program was created to carry out this process for the simulated data. This program requires the user to specify

- the number of simulations or repetitions,
- the total sample size N ($N/2$ diseased and $N/2$ disease free),
- mean vector of the five biomarkers for diseased subjects,
- mean vector of the five biomarkers for non-diseased subjects,
- variance-covariance matrix or matrices and
- the number of bootstrap replicates B .

We calculated AUC after combining the biomarkers using bootstrap Cross-Validation denoted by AUC_{bcv} . To do that we built a Leave One Out Cross-Validation (LOOCV) function and used this function inside bootstrap loop. Estimates of variances are necessary to evaluate the significance of the AUC statistic. The bootstrap is a replication method that can be used for variance estimation.

In the LOOCV function, we split the full dataset into *training* and a *test*, which contains a single observation. The training set consists of the other remaining observations. For the training set, we used logistic regression to build our predictive models using the “glm” function, in the library “stats” together with stepwise variable selection method (“stepAIC” function from “MASS” library for stepwise model selection). We then predicted the part of the data left out. This gives a value between 0 and 1. The process was repeated until all the possible sets are selected. Finally we used

these values to estimate the AUC and to achieve this we used “auc” function from “pROC” library. For estimation of bootstrap Cross-Validated AUC, we drew a bootstrap sample and performed AUC Cross-Validation on the bootstrap sample. We used 1000 bootstrap replicates and obtained 1000 AUC estimates, then we calculated the mean for these 1000 AUCs in order to obtain a single AUC. Also the standard error of the Cross-Validated AUC based on the bootstrap replicates, denoted by SE_b , was obtained.

Computing coverage probability is complex in this setting, therefore we evaluated the true performance of the fixed predictor (model) on a large dataset (10 000 total sample size) using AUC and PE, denoted by TAUC and TPE. This method considers the performance of fixed predictor (based on the smaller dataset) by computing its performance in a large sample.

We used normal linear discriminant approach *LDA* to estimate the true value of the AUC. As it has been mentioned before that the simulated outcomes y^D and $y^{\bar{D}}$ are distributed as a multivariate normal with means μ^D and $\mu^{\bar{D}}$ for the diseased and non-diseased population respectively, and variance-covariance matrices Σ^D and $\Sigma^{\bar{D}}$ from the diseased and non-diseased population respectively. With these notations the true AUC is given by:

$$AUC_{LD} = \Phi \left[\sqrt{(\mu^D - \mu^{\bar{D}})'(\Sigma^D + \Sigma^{\bar{D}})(\mu^D - \mu^{\bar{D}})} \right]. \quad (4.1)$$

Our analysis was based on the estimation of the following quantities: The true AUC based on LDA (AUC_{TLD}), the true AUC (TAUC) obtained from a large dataset, the mean of Bootstrap Cross-Validation AUC (AUC_{bcv}) across 1000 simulations, the mean of Cross-Validated AUC (AUC_{cv}) across 1000 simulations, the confidence interval (CI) for AUC_{cv} using standard error from Hanley and McNiel [31], the confidence interval of AUC_{bcv} based on asymptotic normality using bootstrap standard error, proportion of times lower confidence limits of AUC_{cv} CIs that excludes 0.5, proportion of times lower confidence limits of AUC_{bcv} CIs that excludes 0.5, coverage probability of AUC_{cv} CIs that include the TAUC, coverage probability of AUC_{bcv} CIs that include the TAUC, the empirical standard error for AUC_{bcv} ($SE_{b,e}$), bootstrap standard error SE_b , the empirical standard error for AUC_{cv} ($SE_{e,cv}$), standard error for AUC_{cv} (SE_{cv}) using Hanley and McNeil equation [31] and finally our last two quantities are the prediction error (see Note 3.1.1) and true prediction error denoted by PE and TPE respectively. The disease prevalence (this is the proportion of cases in the total populations) in our simulation was 50%. We summarize our results in Table 4.1.

Table 4.1: Summary of Simulation Studies

	True AUC Based on LDA	CV-AUC Mean across 1000 simulations	BS-CV AUC Mean across 1000 simulations	CI from Hanley & McNeil based on CV-AUC: Mean of lower bound, Mean of upper bound (where mean is computed from 1000 replicates)	CI based on asymptotic normality using BS-SE estimate	Proportion of times lower of CV-based CI bound excludes 0.5	Proportion of times lower of BS-SE-based CI bound excludes 0.5	Coverage probability of CV-AUC based CI (how often does true value fall in CI)	Coverage probability of BS-CV based CI (how often does true value fall in CI)
Model 1: (the no correlation case)	0.6772	0.6492	0.6656	(0.5734, 0.7250)	(0.5628, 0.7684)	0.903	0.875	0.867	0.963
Model 2: (with correlation only between the first two markers)	0.6628	0.6385	0.655	(0.5621, 0.7149)	(0.5494, 0.7607)	0.873	0.793	0.856	0.972
Model 3: (the correlation between all the biomarkers)	0.7422	0.7123	0.7295	(0.6411, 0.7834)	(0.6444, 0.8146)	0.982	1	0.896	0.960

Continued on next page

Table 4.1 (continued)

	SE _{e.b}	SE _b	SE _{e.cv}	SE _{cv}	PE	TPE	TAUC
Model 1: (the no correlation case)	0.039	0.0525	0.0515	0.0387	0.389	0.384	0.662
Model 2: (with correlation only between the first two markers)	0.032	0.0539	0.0516	0.039	0.399	0.393	0.649
Model 3: (the correlation between all the biomarkers)	0.033	0.0434	0.048	0.036	0.342	0.339	0.722

With a total sample size $N = 200$, bootstrap replicates $B = 1000$ and number of simulations $nsim = 1000$, we used the previous three variance-covariance matrices (Σ_1 , Σ_2 and Σ_3) which are used to obtain Models 1, 2 and 3 respectively in Table 4.1) together with the mean vectors of diseased and non-diseased μ^D and $\mu^{\bar{D}}$ to perform our simulation. We obtained 1000 AUC estimates.

From Table 4.1, we can see that:

- For Model 1, the AUC_{TLD} equals to 0.6772 while the TAUC equals to 0.662 and that means using large dataset gives AUC values nearly close to true AUC from LDA. The AUC_{cv} and AUC_{bcv} equal to 0.6492 and 0.6656 respectively. This result indicates that values of AUC_{cv} and AUC_{bcv} are very close to each other and they are nearly unbiased as their values are very close to the true AUC values. Based on CI's for both AUC_{cv} and AUC_{bcv} we deduce that the two methods yield a significant discriminatory probability (the CI's do not include 0.5). We also investigated the the level of discrimination of the two methods by looking at how often the lower limits exclude an $AUC = 0.5$. The proportion of times lower of CI's for AUC_{cv} and AUC_{bcv} that exclude 0.5 are 0.903 and 0.875 respectively. These can be interpreted as the power of the methods. However, both methods (Cross-Validation and Bootstrap Cross-Validation) perform well. The coverage probabilities (proportion of time the CI's include true AUC values) for AUC_{cv} and AUC_{bcv} are 0.875 and 0.963 respectively, indicating that CI's for the cross-validation method are too low. This shows that 875 out of 1000 CI's of AUC_{cv} include the true AUC values, while 963 out of 1000 CI's of AUC_{bcv} include the true AUC values, and indicates that the bootstrap cross-validated AUC estimation perform better than

just cross-validated AUC estimation. We also found that the $SE_{e,b}$ and SE_b are equal to 0.039 and 0.0525 respectively. The $SE_{e,cv}$ and SE_{cv} are 0.0515 and 0.0387 respectively. This shows that the bootstrap affords larger variances and using Hanley & McNeil for standard error gives smaller standard error. This is due to the fact that the sample size is small (especially for cross-validated AUC). Finally we found that both PE and TPE values are similar to each other (0.389 and 0.384 respectively).

- For Model 2, the AUC_{TLD} equals to 0.6628 while the TAUC equals to 0.649. The AUC_{cv} and AUC_{bcv} equal to 0.6385 and 0.655 respectively. This result indicates that values of AUC_{cv} and AUC_{bcv} are close to each other and they are nearly unbiased as their values are close to true AUC values. Based on CI's for both AUC_{cv} and AUC_{bcv} we deduce that the two values of AUC_{cv} and AUC_{bcv} are statistically significant. However the AUC values tend to be lower here than under Model 1. The proportion of times lower of CI's for both AUC_{cv} and AUC_{bcv} that exclude 0.5 are 0.873 and 0.793 respectively. The coverage probabilities of CI's (that include true AUC values) for both AUC_{cv} and AUC_{bcv} are 0.856 and 0.972 respectively. This shows that 972 out of 1000 CI's of AUC_{bcv} include the true AUC values, indicating that the bootstrapping is better than just cross-validation in order to estimate the coverage probability and it gives AUC values close to true AUC. We also found that the $SE_{e,b}$ and SE_b are equal to 0.036 and 0.0539 respectively. The $SE_{e,cv}$ and SE_{cv} are 0.053 and 0.039 respectively. This shows that the bootstrap affords larger standard errors. Finally we found that both PE and TPE values are similar to each other (0.399 and 0.393 respectively).
- For Model 3, the AUC_{TLD} equals to 0.7422 while the TAUC equals to 0.722. The AUC_{cv} and AUC_{bcv} equal to 0.7123 and 0.7295 respectively. We can see that the values of AUC_{cv} and AUC_{bcv} are nearly unbiased as their values are close to true AUC values. Based on CI's for both AUC_{cv} and AUC_{bcv} we deduce that the two values of AUC_{cv} and AUC_{bcv} are statistically significant (the CI's do not include 0.5). The proportion of times lower limits of CI's for both AUC_{cv} and AUC_{bcv} that exclude 0.5 are 0.982 and 1 respectively. This shows high power for both methods. The coverage probabilities of CI's (include true AUC values) for both AUC_{cv} and AUC_{bcv} are 0.896 and 0.960 respectively, indicating that the bootstrap is affords confidence intervals that most of them include the true AUC. We also found that the $SE_{e,b}$ and SE_b are equal to 0.033 and 0.0434 respectively. The $SE_{e,cv}$ and SE_{cv} are 0.048 and 0.036 respectively. Finally we found that both PE and TPE values are 0.342 and 0.339 respectively.

respectively.

From the above results we can see that LOOCV is nearly unbiased in order to estimate the AUC for the three models. It appears that the bootstrap cross-validated AUC values are larger than the cross-validated AUC values, and coverage is better with the bootstrap cross-validation approach. Most of the confidence intervals of bootstrap cross-validated AUC's contained the true values of AUC. We conclude by remarking that using Model 3 (with correlation) is preferable since we obtained the highest AUC and smallest variance compared to other two models. Furthermore the bootstrap obtain larger variances compared to empirical variances. We considered a relatively large sample size. Future study of smaller sample sizes are worth consideration.

4.2. Application to Real Dataset

In Section 4.1 we evaluated the performance of resampling methods through simulation studies. In this section we evaluate the performance of resampling methods through application to a real dataset to see if the results of the simulation studies are consistent when applied to real data. We first introduce the dataset that was used in our application.

4.2.1 Tuberculosis Immune Reconstitution Inflammatory Syndrome (TB-IRIS) Dataset

The dataset that we are using here has been collected from a study of tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS). According to a recent paper by Marais et. al., [40], paradoxical tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS) occurs in 8 – 43% of HIV-infected patients receiving TB treatment after starting antiretroviral therapy (ART) in South Africa. It is found that TB-IRIS results from rapid restoration of Mycobacterium tuberculosis (*M. tuberculosis*)-specific immune responses.

Neurological TB-IRIS occurs in a substantial proportion, specifically in 12% of TB-HIV patients on concurrent treatment cases, and is the commonest cause of central nervous system (CNS) deterioration during the first year of ART in settings of high TB/HIV prevalence. Mortality is high (up to 30%) in those affected. Manifestations of neurological TB-IRIS include: *meningitis, intracranial*

tuberculomata, *brain abscesses*, *radiculomyelitis* and *spinal epidural abscesses*. For more details about these biomarkers, we refer the reader to [3, 12, 39, 49, 42, 61, 64]. There are not many studies describing tuberculous meningitis (TBM)-IRIS that have been published. The authors of [40] investigated clinical and laboratory findings in ART-naive HIV-infected patients who presented with TBM. They undertook serial cerebrospinal fluid (CSF) sampling in patients who did and did not develop TBM-IRIS.

A prospective, observational study targeting 34 adults (≥ 18 years) ART-naive HIV-infected patients presenting with meningitis was carried out at GF Jooste Hospital, a public sector referral hospital in Cape Town. The hospital serves a low-income, high-density population in which the TB notification rate exceeds 1.5% per year with 70% of TB cases co-infected with HIV. This study was carried from March 2009 through October 2010.

4.2.2 Data Results

First, we consider AUC estimation for every biomarker to evaluate the performance of each biomarker in distinguishing between IRIS and Non-IRIS groups. But our main purpose is to calculate AUC after combining biomarkers using resampling methods. Baseline levels of five biomarkers were used to evaluate whether a combination of these biomarkers could accurately discriminate between IRIS and non-IRIS patients. This was accomplished by applying AUC analysis and resampling methods. We denote the five biomarkers by *il12p40*, *tnfa*, *infg*, *il10* and *il6* respectively.

We report the coefficients of the biomarkers from fitting logistic regression for full sample size as in Table 4.2.

Table 4.3 contains the AUC values for each biomarker, their 95% confidence intervals (CI) and standard errors SE. The CI and SE were estimated using the bootstrap method with 1000 replicates. We used the R libraries “pROC” and “pevsuite” for this purpose.

We recall that a larger AUC value would suggest that the test is more accurate in distinguishing between IRIS and non-IRIS subjects and it is expected that the higher the AUC, the less variability there would be. As can be seen in Table 4.3, based on logistic regressions for each biomarker, the smallest AUC values were for *il12p40* and *il10* equal to 0.67 and 0.66 respectively. The bootstrap

Table 4.2: Coefficients of the Biomarkers

	Estimate	Std. Error	z-value	$Pr(> z)$
(Intercept)	-1.668e+02	4.543e+04	-0.004	0.997
<i>il12p40</i>	-7.169e-01	6.442e+02	-0.001	0.999
<i>tnfa</i>	4.168e+00	1.108e+03	0.004	0.997
<i>infq</i>	-1.073e-02	1.750e+01	-0.001	1.000
<i>il10</i>	-3.376e+00	9.303e+02	-0.004	0.997
<i>il6</i>	-9.921e-03	8.698e+00	-0.001	0.999

Table 4.3: AUC Values for Biomarkers

Biomarkers	AUC value	CI	SE
<i>il12p40</i>	0.67	(0.48, 0.84)	0.09
<i>tnfa</i>	0.87	(0.74, 0.96)	0.06
<i>infq</i>	0.77	(0.58, 0.94)	0.09
<i>il10</i>	0.66	(0.45, 0.84)	0.1
<i>il6</i>	0.86	(0.74, 0.96)	0.06

technique provided 95% confidence interval between (0.48 , 0.84) and (0.45 , 0.84) respectively and their respective standard errors equal to 0.09 and 0.1. Based on null hypothesis $H_0 = 0.5$ and from confidence intervals for the *il12p40* and *il10*, we conclude that *il12p40* and *il10* are not statistically significant and do not have high ability to distinguish between two groups of IRIS and non-IRIS, in addition to their variances which are the largest variances compared to other biomarkers. For *tnfa* and *il6*, the AUC estimate equals to 0.87 and 0.86 respectively, while the 95% confidence interval for both of them is approximately (0.74, 0.96), indicating that these two biomarkers are statistically significant. The standard errors equal to 0.06 for both of them and that means they have smallest variances. These results suggest that *tnfa* and *il6* have high ability to distinguish between two groups of IRIS and non-IRIS. For *infq* a logistic regression, yields an AUC estimate equals to 0.77. The bootstrap technique provides 95% confidence interval from 0.58 to 0.94 and standard error equals to 0.09. The CI for *infq* indicates that this biomarker is statistically significant, however

the AUC value suggests that this biomarker does not have high ability to discriminate between two groups of IRIS and non-IRIS.

From Table 4.3 we conclude that the two biomarkers *tnfa* and *il6* have more ability to distinguish between two groups of IRIS and non-IRIS compared to other biomarkers as they have the highest AUCs values and less variability. Each of Figures 4.1, 4.2, 4.3, 4.4 and 4.5 contains the ROC curve, AUC and CI for each biomarker. Figure 4.6 which combines all the biomarkers together supports our results.

Figure 4.1: ROC Curve For *il12p40*

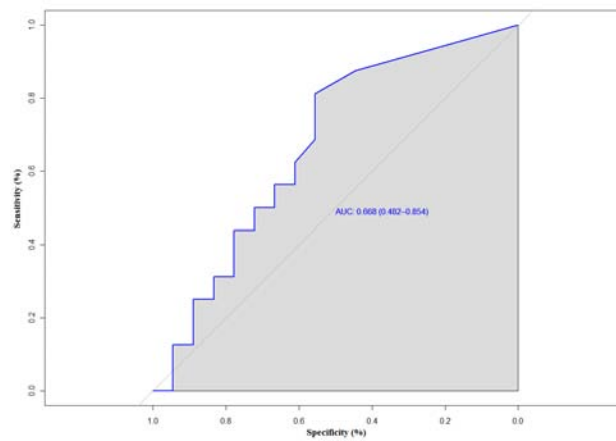


Figure 4.2: ROC Curve For *tnfa*

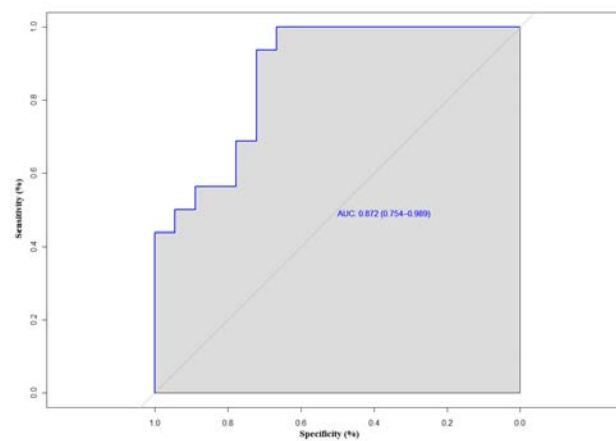


Figure 4.3: ROC Curve For *infg*

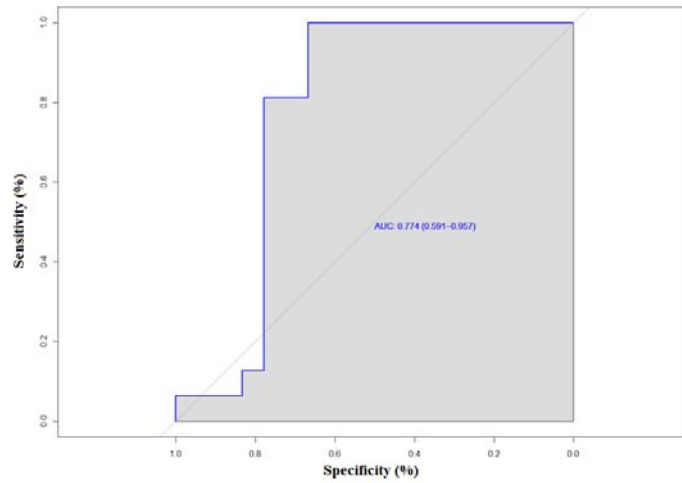
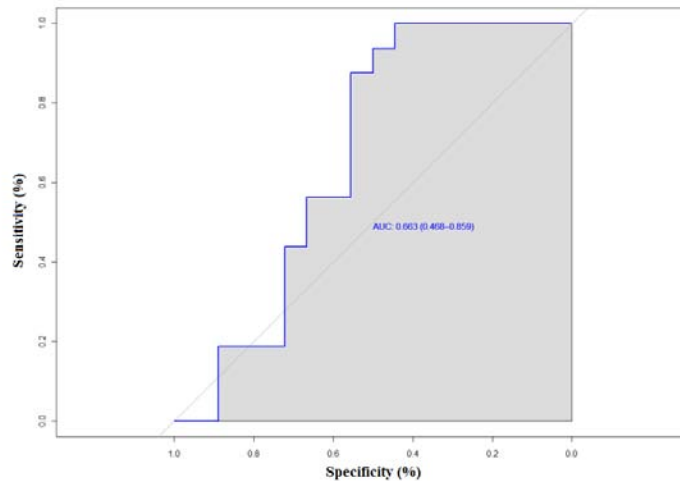


Figure 4.4: ROC Curve For *il10*



Note that we are interested in combining biomarkers for the purpose of estimating the AUC. For this purpose we use Logistic Regression as we have a binary response (IRIS states).

Baseline levels of five biomarkers were evaluated and we used them to assess whether a combination of these biomarkers could distinguish between IRIS and non-IRIS patients with high probability. This is accomplished by applying AUC analysis and resampling methods.

We calculated AUC after combining the *il12p40*, *tnfa*, *infg*, *il10* and *il6* using bootstrap Cross-Validation. To do this we built a Leave One Out Cross-Validation (LOOCV) function and used this function inside bootstrap loop as before in simulated data. We also calculated the variance for bootstrap cross validated AUC the same way for simulated data.

Figure 4.5: ROC Curve For *il6*

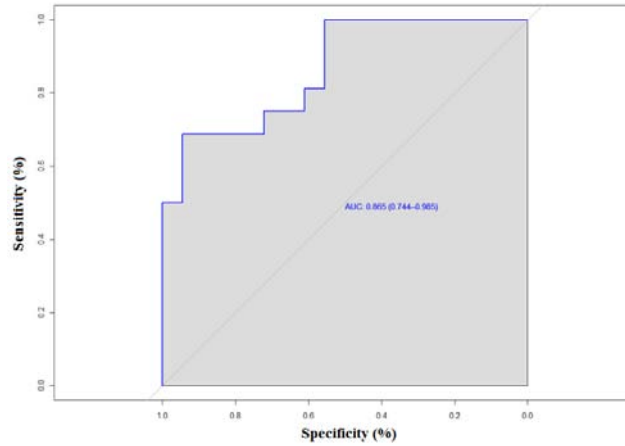
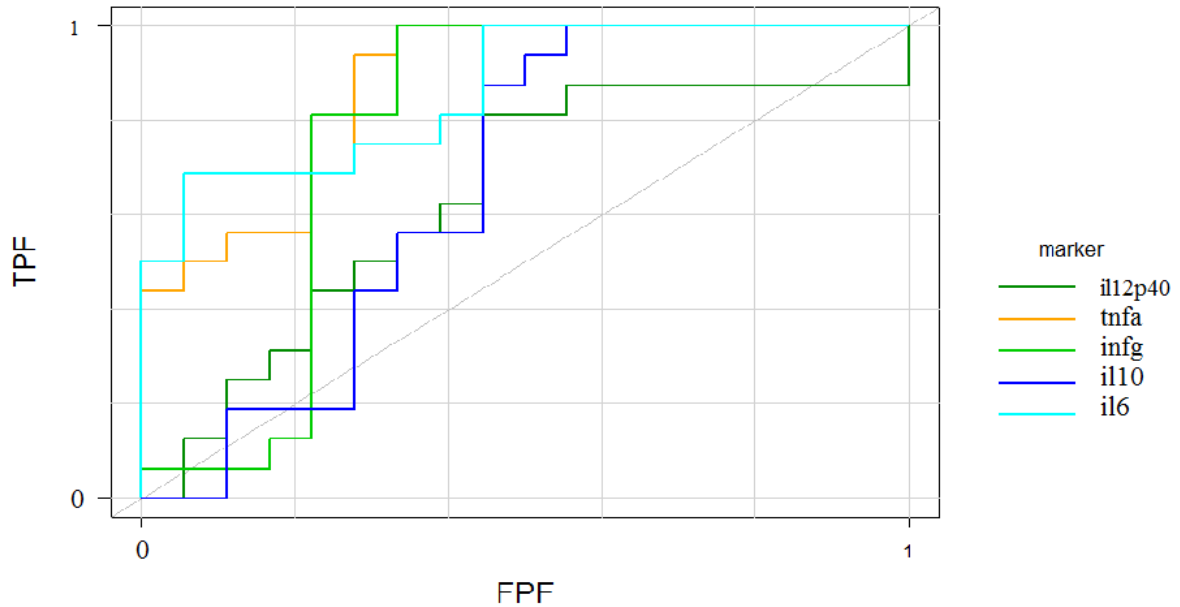


Figure 4.6: ROC Curves For *il12p40*, *tnfa*, *infg*, *il10* and *il6*



An R program similar to the one used in simulation studies, was created to apply the bootstrap cross-validation technique of AUC and variance estimation to the IRIS data. In this R program the user is required to specify an input data set and the number of bootstrap replicates. In our case we used 1000 bootstrap replicates. Table 4.4 shows the results obtained from the application of the bootstrap Cross-Validation technique.

After applying bootstrap Cross-Validated technique we obtained AUC_{bcv} equals to 0.956. The estimated Standard error for this AUC based on the bootstrap replicates SE_b is 0.036. The confidence

Table 4.4: Bootstrap Cross-Validation of AUC

AUC_{bcv}	SE_b	CI
0.956	0.036	(0.8153, 0.9907)

interval is found to be (0.8153, 0.9907), which does not contain 0.5 and thus the estimated AUC_{bcv} is statistically significant. The values of AUC_{bcv} , SE_b and CI suggests that the combination of biomarkers yield a high AUC value and small variability. This shows that the combination of biomarkers has high ability to distinguish between diseased states. Applying the Cross Validation method (without bootstrapping), we found that the Cross-Validated AUC (AUC_{cv}) is given by 0.967, which is larger than the AUC from bootstrap AUC_{bcv} , although the difference is small. We also estimated the PE from Cross-Validation applied to the logistic regression and found that the $PE = 0.059$, indicating that the model performance is good. An important point that we would like to mention is that the AUC obtained by combining multiple biomarkers using both Cross-Validation and Bootstrap Cross-Validation methods have high distinguishing accuracy between IRIS and non-IRIS subjects compared to AUC based on involving single biomarker in the model.

We conclude this chapter by remarking that resampling methods (e.g., cross-validation and bootstrap) for the purpose of estimating the area under the receiver operating characteristic curve (AUC) for a model that combine biomarkers from simulated and real datasets (TB-IRIS) gives deep insight in understanding the disease and provide a more accurate diagnosis of the disease.

Conclusion

Diagnostic tests are important components in modern medical practice. In clinical medicine, correct diagnosis of disease is of great interest and hence researchers invest considerable time and developing methods to enhance accurate disease diagnosis. The receiver operating characteristic (ROC) is a commonly used statistical tool for describing the discriminatory accuracy and performance of a diagnostic test. The area under receiver operating characteristic (AUC) is a popular summary index of discriminatory accuracy. In this thesis we evaluated some of the resampling methods namely Cross-Validation (more specifically Leave One Out Cross-Validation) together with Bootstrap for the purpose of estimating the (AUC) for models that combine multiple biomarkers. According to our simulation studies, using large number of bootstrap replicates would give better AUC estimates and small variability. We found that Model 3 gives higher AUC values compared to the other models. This tells us that the assumptions under which data is simulated will affect the AUC values. We also deduced that Leave One Out Cross-Validation (LOOCV) gives nearly unbiased AUC estimation, but that it alone can not be used to obtain confidence intervals with the correct coverage. Using bootstrap cross-validated AUC gives high coverage probability. We would like to indicate that the bootstrap affords larger variances compared to empirical variances, and coverage probability is at or above the nominal level. Also AUC estimates based on bootstrap cross-validation are larger than those based on cross-validation. An interesting point is that using large dataset gives AUC estimates nearly equal to those found using LDA.

An application to IRIS dataset reveals that Cross-Validation method gives higher AUC estimates than the Bootstrap method. However both cross-validation and bootstrap cross-validation yield AUC estimations closely to each other.

We conclude by remarking that resampling methods (e.g., cross-validation and bootstrap) for the purpose of estimating the area under the receiver operating characteristic curve (AUC) for a model that combine biomarkers from simulated and real datasets (TB-IRIS) gives insight in understanding the disease and provide a more accurate diagnosis of the disease.

In my future work I plan to continue developing techniques considered in the this project. Simulation studies evaluating performance with smaller sample sizes are warranted. The extension of this work is to the setting of combining markers measured over time using a time to event outcome. The introduction of a longitudinal component and censored observations adds new complexities. An important goal of this analysis is to develop methods that can be iteratively explored with our medical collaborators. Methods modeling the joint survival distribution have been previously developed by Heagerty et al [29], but an imputation-based approach might perform similarly. Such an approach would certainly be easier to implement. Specifically, imputation methods will be considered to address censoring of survival times. Then longitudinal models can be fit. We will apply multiple imputation for censored data and obtain time dependent prediction curves. After that we will compare the results to existing methods for ROC analysis.

Appendix

```
### Simulation R Code
```

```
cat(NULL,file="my.csv")
```

```
library(MASS)
```

```
library(pROC)
```

```
Sim<-function(nsim,N,mu1,mu2,Sigma, nboot,seed){
```

```
  set.seed(seed)
```

```
  for ( i in 1:nsim){
```

```
    X1=mvrnorm(N/2,mu1,Sigma)
```

```
    X2=mvrnorm(N/2,mu2,Sigma)
```

```
    X=rbind(X1,X2)
```

```
    y1=rep(1,N/2)
```

```
    y2=rep(0,N/2)
```

```
    Y=c(y1,y2)
```

```
  Dataset = data.frame(Y,X)
```

```
X1.BIG=mvrnorm(5000,mu1,Sigma)
X2.BIG=mvrnorm(5000,mu2,Sigma)
X.BIG=rbind(X1.BIG,X2.BIG)
y1.BIG=rep(1,5000)
y2.BIG=rep(0,5000)
Y.BIG=c(y1.BIG,y2.BIG)

Dataset.BIG= data.frame(Y.BIG,X.BIG)

#####

LOOCV.AUC<-function(Dataset.b){
  pred.score<-c()
  for (i in 1:rows){

    model<-glm(Y~X1+X2+X3+X4+X5,Dataset.b[-i,],family=binomial)
    MS<-stepAIC(model,direction="both")

    pred.score[i]<-predict(MS,newdata=Dataset.b[i,], type="response")
  }
  P = ifelse(pred.score[i]>0.5,1,0)
  PE1= mean((Y - P )^2)

  Auc=auc(Dataset.b$Y, pred.score)
  Tab=data.frame(Auc=Auc,PE1=PE1)
  return(Tab)
}
auc.b<-c()

for (j in 1:nboot){
  rows<- nrow(Dataset)
```

```
s<- sample(1:rows,rows,replace=TRUE)
  Dataset.b<-Dataset[s,]
  auc.b[j]<-LOOCV.AUC(Dataset.b)$Auc
}
auc.b
AUC=mean(auc.b)
SE=sqrt(var(auc.b))

#####

PE=LOOCV.AUC(Dataset)$PE1

AUC.cv=LOOCV.AUC(Dataset)$Auc

pred.new<-c()
model<-glm(Y~X1+X2+X3+X4+X5,Dataset,family=binomial)
MS<-stepAIC(model,direction="both", trace=FALSE)
pred.new<-predict(MS,newdata=Dataset.BIG, type="response")
P=ifelse(pred.new>0.5,1,0)
PE.new= mean((Y.BIG - P)^2)
Auc.new=auc(Dataset.BIG$Y, pred.new)

out <- data.frame(AUC,SE,PE,AUC.cv,Auc.new=Auc.new,PE.new)
write.table(out, file = "my.csv", append=(i>1))
}
}
Sim(nsim,N,mu1,mu2,Sigma, nboot,seed)
file.show("my.csv")
```

```
### R Code For IRIS Data
```

```
roccurve(dataset="TBM", d="Y", markers=c("V1","V2","V3","V4","V5"),
level=95,nsamp=1000,rocmeth="nonparametric")
```

```
#####
```

```
ROC1<-roc(TBM $Y, TBM $V1)
```

```
ROC2<-roc(TBM $Y, TBM $V2)
```

```
ROC3<-roc(TBM $Y, TBM $V3)
```

```
ROC4<-roc(TBM $Y, TBM $V4)
```

```
ROC5<-roc(TBM $Y, TBM $V5)
```

```
ROC1 ROC2 ROC3 ROC4 ROC5
```

```
sqrt(var(ROC1,method="bootstrap", boot.n=1000))
```

```
sqrt(var(ROC2,method="bootstrap", boot.n=1000))
```

```
sqrt(var(ROC3,method="bootstrap", boot.n=1000))
```

```
sqrt(var(ROC5,method="bootstrap", boot.n=1000))
```

```
sqrt(var(ROC5,method="bootstrap", boot.n=1000))
```

```
#####
```

```
ci.auc(TBM $Y,TBM$V1,boot.n=1000,conf.level=0.95,method="bootstrap")

ci.auc(TBM $Y,TBM$V2,boot.n=1000,conf.level=0.95,method="bootstrap")

ci.auc(TBM $Y,TBM$V3,boot.n=1000,conf.level=0.95,method="bootstrap")

ci.auc(TBM $Y,TBM$V4,boot.n=1000,conf.level=0.95,method="bootstrap")

ci.auc(TBM $Y,TBM$V5,boot.n=1000,conf.level=0.95,method="bootstrap")

#####

LOOCV.AUC<-function(Dataset.b){ pred.score<-c()
  for (i in 1:rows){
    model<-glm(Y~V1+V2+V3+V4+V5,Dataset.b[-i,],family=binomial)
    MS<-stepAIC(model,direction="both")
    pred.score[i]<-predict(MS,newdata=Dataset.b[i,], type="response")
  }
  P = ifelse(pred.score[i]>0.5,1,0)

  PE1= mean((Y - P )^2) Auc=auc(Dataset.b$Y, pred.score)
  Tab=data.frame(Auc=Auc,PE1=PE1) return(Tab)
}

nboot=1000
auc.b<-c()

for (j in 1:nboot){
  rows<-nrow(TBM)
  s<-sample(1:rows,rows,replace=TRUE)
```

```
Dataset.b<-TBM[s,]  
auc.b[j]<-LOOCV.AUC(Dataset.b)$Auc  
}  
auc.b  
AUC=mean(auc.b)  
AUC  
SE=sqrt(var(auc.b))  
SE  
PE=LOOCV.AUC(TBM)$PE1  
PE
```

Bibliography

- [1] A. Airola, T. Pahikkala, W. Waegeman, B. Baets and T. Salakoski, *A comparison of AUC estimators in small-sample studies*, Machine Learning in Systems Biology, **8** (2010), 3 - 13.
- [2] K. Aoki, J. Misumi, T. Kimura, W. Zhao and T. Xie, *Evaluation of cuto levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogens I, II and of PG I/PG II ratios in a gastric cancer case-control study*, Journal of Epidemiology, **7** (1997), 143 - 151.
- [3] V. Asselman, F. Thienemann, D. J. Pepper, et al., *Central nervous system disorders after starting antiretroviral therapy in South Africa*, AIDS **24** (2010) 2871 - 2876.
- [4] P. C. Austin and J. V. Tu, *Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality*, Journal of Clinical Epidemiology, **57** (2004), 1138 - 1146.
- [5] T. L. Baily, C. Elkan, *Estimating the accuracy of learned concepts*, Proceedings of International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, (1993), 895 - 900.
- [6] D. Bamber, *The area above the ordinal dominance graph and the area below the receiver operating characteristic graph*, Journal of Mathematical Psychology, **12** (1975), 387 - 415.
- [7] A. I. Bandos, H. E. Rockette and D. Gur, *A permutation test sensitive to differences in areas for comparing ROC curves from a paired design*, Statistics in Medicine, **24** (2005), 2873 - 2893.
- [8] T. M. Braun and T. A. Alonzo, *A modified sign test for comparing paired ROC curves*, Biostatistics, **9** (2008), 364 - 372.

-
- [9] L. Breiman, P. Spector, *Submodel selection and evaluation in regression the x random case* *International Statistical Review*, *International Statistical Review*, **60** (1992), 291–319.
- [10] G. Campbell, *General methodology I: advances in statistical methodology for the evaluation of diagnostic and laboratory tests*, *Statistics in Medicine*, **13** (1994), 499–508.
- [11] J. B. Copas and P. Corbett, *Overestimation of the receiver operating characteristic curve for logistic regression*, *Biometrika*, **89** (2002), 315–331.
- [12] J. A. Crump, M. J. Tyrer, S. J. Lloyed-Owen, L. Y. Han, M. C. Lipman and M. A. Johnson, *Miliary tuberculosis with paradoxical expansion of intracranial tuberculomas complicating human immunodeficiency virus infection in a patient receiving highly active antiretroviral therapy*, *Clin Infect Dis* **26** (1998) 1008–1009.
- [13] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach*, *Biometrics* **44** (1988), 837–845.
- [14] L. E. Dodd, *Regression Methods for Areas and Partial Areas Under the ROC Curves*, PhD Thesis, University of Washington, 2001.
- [15] D. D. Dorfman, K. S. Berbaum and C. E. Metz, *Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method*, *Investigative Radiology*, **27** (1992), 723–731.
- [16] L. E. Dodd and M. S. Pepe, *Partial AUC estimation and regression*, *Biometrics*, **59** (2003), 614–623.
- [17] L. E. Dodd and M. S. Pepe, *Semiparametric Regression for the Area under the Receiver Operating Characteristic Curve*, *Journal of the American Statistical Association*, **98** (2003), 409–417.
- [18] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, *The Annals of Statistics*, **7** (1979), 1–26.
- [19] B. Efron, *Estimating the error rate of prediction rule: improvement on cross-validation*, *Journal of the American Statistical Association*, **78** (1983), 316–330.
- [20] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, *Monographs on Statistics and Applied Probability*, **57**, Chapman and Hall, 1993.
-

-
- [21] J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, 1975.
- [22] Y. Fang, Q. Gengsheng and X. Huang, *Optimal combinations of diagnostic tests based on AUC*, *Biometrics*, **67** (2011), 568 - 576.
- [23] T. Fawcett, *An introduction to ROC analysis*, *Pattern Recognition Letters Journal*, **27** (2006), 861 - 874.
- [24] J. P. Fine and R. J. Bosch, *Risk assessment via a robust probit model, with application to toxicology*, *Journal of American Statistical Association*, **95** (2000), 375 - 382.
- [25] R. Fluss, *Estimation of the ROC Curves and its Associated Indices Under Verification Bias*, PhD Thesis, University of Haifa, 2007.
- [26] M. H. Gail and S. B. Green, *A generalization of one-sided two sample Kolmogorov-Smirnov statistics for evaluating diagnostic tests*, *Biometrics*, **32** (1976), 561 - 570.
- [27] D. M. Green and J. A Swets, *Signal Detection Theory and Psychophysics*, Wiley, New York, 1966.
- [28] M. Greiner, D. Pfeiffer and R. D Smith, *Principals and practical application of the receiver operating characteristic analysis for diagnostic tests*, *Preventive Veteranary Medicine*, **45** (2000), 23 - 41.
- [29] P. J. Heagerty, T. Lumley and M. S. Pepe, *Time-dependent ROC curves for censored survival data and a diagnostic marker*, *Biometrics*, **56** (2000), 337 - 344.
- [30] S. Ma and J. Huang *Regularized ROC method for disease classification and biomarker selection with microarray data*, *Bioinformatics*, **21** (2005), 4356 - 4362.
- [31] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under an ROC curve*, *Radiology*, **143** (1982), 29 - 36.
- [32] J. A. Hanley, *The robustness of the 'binormal' assumptions used in fitting ROC curves*, *Medical Decision Making*, **8** (1988), 197 - 203.
- [33] J. A. Hanley, *The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests*, *Statistitcs in Medicine*, **15** (1996), 1575 - 1585.

-
- [34] W. W. hauck, T. Hyslop and S. Anderson, *Generalized treatment effects for clinical trials*, *Statistics in Medicine*, **15** (2000), 887 - 899.
- [35] W. Hoeffding, *A class of statistics with asymptotically normal distribution*, *Annals of Mathematical Statistics*, **19** (1948), 293 - 325.
- [36] A. K. Jam, R. C. Dubes and C. Chen, *Bootstrap techniques for error estimation*, *IEEE transactions on pattern analysis and machine intelligence PAMI*, **9** (1987), 628 - 633.
- [37] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Morgan Kaufmann, (1995), 1137 - 1143.
- [38] B. J. LaFleur and R. A. Greevy, *Introduction to permutation and resampling-based hypothesis tests*, *Journal of Clinical Child and Adolescent Psychology*, **38** (2009), 286 - 294.
- [39] C. H. Lee, C. C. Lui and J. W. Liu, *Immune reconstitution syndrome in a patient with AIDS with paradoxically deteriorating brain tuberculoma*, *AIDS Patient Care STDS* **21** (2007) 234 - 239.
- [40] S. Marais, G. Meintjes, D. J. Pepper, L. E. Dodd, C. Schutz1, Z. Ismail, K. A. Wilkinson, R. J. Wilkinson, *Frequency, severity and prediction of tuberculous meningitis immune reconstitution inflammatory syndrome*, *Clinical Infectious Diseases*, to appear, 2012.
- [41] C. E. Metz, *Some practical issues of experimental design and data analysis in radiological ROC studies*, *Investigative Radiology*, **24** (1989), 234 - 245.
- [42] W. Manosuthi, S. Kiertiburanakul, T. Phoorisri and S. Sungkanuparph, *Immune reconstitution inflammatory syndrome of tuberculosis among HIV-infected patients receiving antituberculous and antiretroviral therapy*, *J Infect* **53** (2006) 357 - 363.
- [43] R. J. Marshal, *The predictive value of simple rules for combining two diagnostic tests*, *Biometrics*, **45** (1989), 1213 - 1222.
- [44] C. E. Metz, *Some practical issues of experimental design and data analysis in radiological ROC studies*, *Investigative Radiology*, **24** (1989), 234 - 245.
- [45] A. Moise, B. Cliement and M. Raissis, *A test for crossing receiver operating characteristic (ROC) curves*, *Communications in Statistics Theory and Methods*, **17** (1988), 1985 - 2003.

-
- [46] M. S. Pepe and M. L. Thompson, *Combining diagnostic test results to increase accuracy*, *Biostatistics*, **1** 2000, 123–140.
- [47] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Sciences Series, **31**, Cambridge University Press, 2003.
- [48] M. S. Pepe, T. Cai and G. Longton, *Combining predictors for classification using the area under the receiver operating characteristic curve*, *Biometrics*, **62** (2006), 221–229.
- [49] D. J. Pepper, S. Marais, G. Maartens, et al, *Neurologic manifestations of paradoxical tuberculosis associated immune reconstitution inflammatory syndrome: a case series*, *Clin Infect Dis* **48** (2009) 96–107.
- [50] F. Provost and T. Fawcett, *Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions*, Proc. Third Internat. Conf. on Knowledge Discovery and Data Mining (KDD-97). AAAI Press, Menlo Park, CA, (1997) 43–48
- [51] , D. E. Shapiro, *The interpretation of diagnostic tests*, *Statistical Methods in Medical Research*, **8** (1999), 113–134.
- [52] R. Simon, E. Korn, L. McShane, M. Radmacher, G. Wright G and Y. Zhao, *Design and Analysis of DNA Microarray Investigations*, Springer-Verlag, New York, 2003.
- [53] H. Skalská and V. Freylich, *Web-Bootstrap estimate of area under ROC curve*, *Austrian Journal of Statistics*, **25** (2006), 325–330.
- [54] H. H. Song, *Analysis of correlated ROC areas in diagnostic testing*, *Biometrics* **53** (1997), 370–382.
- [55] K. A. Spackman, *Signal detection theory: Valuable tools for evaluating inductive learning*, In: Proc. Sixth Internat. Workshop on Machine Learning. Morgan Kaufman, San Mateo, CA, 1989, 160–163.
- [56] J. Q. Su and J. S. Liu, *Linear combination of multiple diagnostic markers*, *Journal of the American Statistical Association*, **88** (1993), 1350–1355.
- [57] J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press, 1982.

- [58] J. A. Swet, *Indices of discrimination or diagnostic accuracy: Their ROCs and implied models*, Psychological Bulletin, **99** (1986), 100 - 117.
- [59] J. A. Swet, *Measuring the accuracy of diagnostic systems*, Science, **240** (1988), 1285 - 1293.
- [60] J. A. Swet, R. M. Dawes, J. Monahan, *Better decisions through science*, Scientific American, **283** (2000), 82 - 87.
- [61] F. F. Tuon, G. C. Mulatti, W. P. Pinto, de Siqueira Franca FO and R. C. Gryscek *Immune reconstitution inflammatory syndrome associated with disseminated mycobacterial infection in patients with AIDS*, AIDS Patient Care STDS **48** (2007) 527 - 532.
- [62] E. S. Venkatraman and C. B. Begg, *A distributionfree procedure for comparing receiver operating characteristic curves from a paired experiment*, Biometrika **83** (1996), 835 - 848.
- [63] E. S. Venkatraman, *A permutation test to compare receiver operating characteristic curves*, Biometrics **56** (2000), 1134 - 1138.
- [64] J. E. Vidal, S. Cimerman, R. Schiavon Nogueira and et al., *Paradoxical reaction during treatment of tuberculous brain abscess in a patient with AIDS*, Rev Inst Med Trop Sao Paulo **45** (2003) 177 - 178.
- [65] S. Wieand, M. H. Gail, B. R. James and K. L. James, *A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data*, Biometrika **76** (1989), 585 - 592.
- [66] L. Zhang, Y. D. Zhao and J. D. Tubbs, *Inference for Semiparametric AUC Regression Models with Discrete Covariates*, Journal of Data Science **9** (2011), 625 - 637.
- [67] X. H. Zhou, N. A. Obuchowski and D. K. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley NY, 2002.
- [68] K. H. Zou, *Receiver operating characteristic (ROC) literature research*, 2002. On-line bibliography available from: <http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html>