

**Statistical methods to evaluate disease outcome diagnostic  
accuracy of multiple biomarkers with application to HIV and  
TB research**

By

Muna Balla Elshareef Mohammed

mimielshareef@gmail.com

Supervisor : Professor Henry G. Mwambi

mwambih@ukzn.ac.za



School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

Pietermaritzburg, South Africa

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

October 2015



# Abstract

One challenge in clinical medicine is that of the correct diagnosis of disease. Medical researchers invest considerable time and effort to improving accurate disease diagnosis and following from this diagnostic tests are important components in modern medical practice. The *receiver operating characteristic* (ROC) is a statistical tool commonly used 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 medical research data, scientists are simultaneously evaluating hundreds of biomarkers. A critical challenge is the combination of biomarkers into models that give insight into disease. In infectious disease, biomarkers are often evaluated as well as in the micro organism 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 are important to the diagnosis and treatment of disease. Taken together, this extends the role of, the statistical analyst and presents many novel and major challenges. This thesis discusses some of the various strategies and issues in using statistical data analysis to address the diagnosis problem, of selecting and combining multiple markers to estimate the predictive accuracy of test results. We also consider different methodologies to address missing data and to improve

---

the predictive accuracy in the presence of incomplete data.

The thesis is divided into five parts. The first part is an introduction to the theory behind the methods that we used in this work. The second part places emphasis on the so called *classic ROC analysis*, which is applied to cross sectional data. The main aim of this chapter is to address the problem of how to select and combine multiple markers and evaluate the appropriateness of certain techniques used in estimating the area under the ROC curve (AUC). Logistic regression models offer a simple method for combining markers. We applied resampling methods to adjust for over-fitting associated with model selection. We simulated several multivariate models to evaluate the performance of the resampling approaches in this setting. We applied these methods to data collected from a study of tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS) in Cape Town, South Africa. Baseline levels of five biomarkers were evaluated and we used this dataset to evaluate whether a combination of these biomarkers could accurately discriminate between TB-IRIS and non TB-IRIS patients, by applying AUC analysis and resampling methods.

The third part is concerned with a time dependent ROC analysis with event-time outcome and comparative analysis of the techniques applied to incomplete covariates. Three different methods are assessed and investigated, namely *mean imputation*, *nearest neighbor hot deck imputation* and *multivariate imputation by chain equations (MICE)*. These methods were used together with bootstrap and cross-validation to estimate the time dependent AUC using a non-parametric approach and a Cox model. We simulated several models to evaluate the performance of the resampling approaches and imputation methods. We applied the above methods to a real data set.

The fourth part is concerned with applying more advanced variable selection methods to predict the survival of patients using time dependent ROC analysis. The least absolute shrinkage and

---

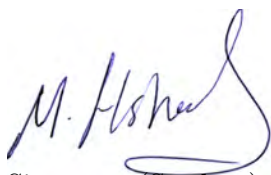
selection operator (LASSO) Cox model is applied to estimate the bootstrap cross-validated, 632 and 632+ bootstrap AUCs for TBM/HIV data set from KwaZulu-Natal in South Africa. We also suggest the use of ridge-Cox regression to estimate the AUC and two level bootstrapping to estimate the variances for AUC, in addition to evaluating these suggested methods.

The last part of the research is an application study using genetic HIV data from rural KwaZulu-Natal to evaluate the sequence of ambiguities as a biomarker to predict recent infection in HIV patients.

# Preface


The work described in this thesis was carried out from March 2013 to October 2015, under the supervision and direction of Professor Henry G. Mwambi, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg.

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: 31st of March 2016



Signature (Supervisor)

Date: 31st of March 2016

# Dedication

TO MY LOVELY PARENTS DR BALLA AND HANAN, MY DEAR HUSBAND AYOUB, MY LOVELY  
DAUGHTER FATIMA, MY BROTHERS MUGTABA, 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 supervisor, Professor Henry G. Mwambi, for his invaluable learned guidance, advises, encouragement, understanding and continued support he has provided me throughout the duration of my studies which led to the compilation of this thesis. I will be always indebted to him 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 Dr Balla for their continuous support and best wishes.

Also I would like to thank Mr Rob Ettershank, for his kindness and valuable corrections, comments and suggestions through the editing and proofreading process.

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 and the South African



---

Center for Epidemiological Modelling and Analysis (SACEMA). My thanks extend to Professor Robert Wilkinson, Dr Suzaan Marais and Professor Tulio de Oliveira for supporting us with real datasets.

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

# Table of contents

<b>Abstract</b>	<b>ii</b>
<b>Preface</b>	<b>v</b>
<b>Dedication</b>	<b>vi</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Table of contents</b>	<b>ix</b>
<b>List of notations</b>	<b>xvi</b>
<b>List of figures</b>	<b>xix</b>
<b>List of tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>

---

1.1	Motivation, purposes and objectives . . . . .	1
1.2	Background and related studies on ROC analysis . . . . .	5
1.3	Thesis outlines . . . . .	8
<b>2</b>	<b>Receiver operating characteristic (ROC) curves</b>	<b>12</b>
2.1	Definitions and basic concepts . . . . .	12
2.2	Introduction to receiver operator characteristic curve (ROC) for continuous tests	15
2.2.1	Properties and attributes of ROC curves . . . . .	16
2.3	Some of ROC curves indices . . . . .	19
2.3.1	Area under ROC curves (AUC) . . . . .	19
2.3.2	The $ROC(t_0)$ . . . . .	21
2.3.3	Partial AUC . . . . .	21
2.3.4	Kolmogorov-Smirnov (KS) index . . . . .	22
2.4	Binormal ROC curves . . . . .	23
2.5	The ROC for ordinal tests . . . . .	26
2.5.1	Ordered discrete tests results . . . . .	26
2.5.2	The latent decision variable model . . . . .	27
2.5.3	The discrete ROC curve . . . . .	27

---

2.6	The ROC curve estimation . . . . .	29
2.6.1	Non-parametric empirical estimator . . . . .	30
2.6.2	Modeling test results . . . . .	34
2.6.3	Parametric methods . . . . .	36
2.7	Modeling covariate effects on test results . . . . .	38
2.8	Modeling covariate effects on ROC curve . . . . .	39
<b>3</b>	<b>Time dependent receiver operating characteristic curves</b>	<b>41</b>
3.1	Extensions of sensitivity and specificity . . . . .	44
3.2	Time dependent true positive rate ( $TPR(t)$ ) and false positive rate ( $FPR(t)$ ) .	46
3.2.1	Time dependent true positive rate . . . . .	47
3.2.2	Time dependent false positive rate . . . . .	48
3.3	Combinations of time dependent $TPR$ and $FPR$ . . . . .	49
3.3.1	Incident-Static combination . . . . .	50
3.3.2	Incident-Dynamic combination . . . . .	51
3.3.3	Cumulative-Dynamic AUC $AUC^{C/D}(t)$ . . . . .	52
<b>4</b>	<b>Missing data and imputation methods</b>	<b>55</b>
4.1	Missing reasons . . . . .	57

---

4.1.1	Missing completely at random (MCAR) . . . . .	57
4.1.2	Missing at random (MAR) . . . . .	58
4.1.3	Missing not at random (MNAR) . . . . .	58
4.1.4	Ignorable and non-ignorable missingness . . . . .	59
4.2	Imputation strategies . . . . .	60
4.2.1	Mean imputation . . . . .	63
4.2.2	Hot deck imputation . . . . .	64
4.2.3	Multiple imputation . . . . .	67
4.2.4	Multiple imputation via chained equations . . . . .	71
<b>5</b>	<b>Combining multiple biomarkers in diagnostic testing for cross-sectional data</b>	<b>76</b>
5.1	Introduction . . . . .	76
5.2	Variable selection . . . . .	79
5.2.1	Backward elimination . . . . .	80
5.2.2	Forward selection . . . . .	81
5.2.3	Stepwise regression . . . . .	81
5.3	Resampling methods in the context of combining multiple biomarkers and estimation of the AUC . . . . .	83

---

5.3.1	Over-fitting . . . . .	84
5.3.2	Cross-validation . . . . .	86
5.3.3	$K$ -fold cross-validation . . . . .	88
5.3.4	Leave one out cross-validation (LOOCV) . . . . .	89
5.3.5	Bootstrap method . . . . .	89
5.3.6	Bootstrap standard errors and confidence intervals . . . . .	91
5.3.7	Bootstrap cross-validation . . . . .	93
5.3.8	Leave-one-out bootstrap . . . . .	93
5.3.9	Algorithm to obtain the AUC through cross-validation . . . . .	95
5.3.10	Algorithm to estimate the variance of AUC through bootstrapping . . . . .	95
5.4	Logistic regression . . . . .	96
5.5	Linear discriminant analysis . . . . .	99
5.6	Algorithm . . . . .	99
5.7	Simulation studies . . . . .	101
5.8	Application to TB-IRIS . . . . .	108
5.8.1	Conclusion . . . . .	113

## 6 Predictive accuracy of multiple time dependent biomarkers with missing

---

<b>values in diagnostic testing</b>	<b>116</b>
6.1 Introduction . . . . .	116
6.2 Missing data and imputation methods . . . . .	117
6.3 Methods for estimation of $AUC(t)$ . . . . .	118
6.4 Models for predictive scores . . . . .	122
6.5 Algorithm . . . . .	126
6.6 Simulation studies . . . . .	127
6.7 Application to primary biliary cirrhosis (PBC) . . . . .	131
6.8 Conclusion . . . . .	138
 <b>7 Use of resampling methods to predict the outcome of tuberculous meningitis in high HIV prevalence patients in South Africa</b>	 <b>140</b>
7.1 Introduction . . . . .	140
7.2 Penalized Cox methods . . . . .	142
7.2.1 LASSO-Cox regression . . . . .	143
7.2.2 Ridge-Cox regression . . . . .	146
7.3 Resampling methods . . . . .	147
7.3.1 632 bootstrap . . . . .	148

---

---

7.3.2	632+ bootstrap . . . . .	149
7.3.3	Estimation of standard errors . . . . .	150
7.4	Algorithm . . . . .	150
7.5	Simulation studies . . . . .	152
7.6	Application of tuberculous meningitis (TBM) in high HIV prevalence . . . . .	157
7.6.1	The TBM/HIV description . . . . .	157
7.6.2	Statistical analysis, results and discussion . . . . .	158
7.7	Conclusion . . . . .	175
<b>8</b>	<b>The role of ambiguous nucleotides as biomarkers of recent HIV infection in rural KwaZulu-Natal, South Africa</b>	<b>176</b>
8.1	Introduction . . . . .	176
8.2	Methods . . . . .	178
8.2.1	Data description . . . . .	178
8.2.2	Statistical methods . . . . .	180
8.3	Results and discussion . . . . .	181
<b>9</b>	<b>Conclusion and future work</b>	<b>186</b>
	<b>Bibliography</b>	<b>190</b>

---



## 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	generalised 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
TBM	tuberculosis meningitis

BCV	bootstrap cross-validation
KM	Kaplan-Meier estimator
NNE	nearest neighbor estimator
CPH	Cox proportional hazard
MCAR	missing completely at random
MAR	missing at random
MNAR	missing not at random
$NNI_{HD}$	nearest neighbor hot deck imputation
STD	standard deviation
PBC	primary biliary cirrhosis
DBS	dried blood spots
MI	multiple imputation
MICE	multivariate imputation by chain equations
LASSO	least absolute selection and shrinkage operator
OLS	ordinary least squares
logvl	log viral load
CSF	cerebrospinal fluid
lymp	lymphocytes
poly	polymerase
wcc	white cells count
632	632 bootstrap method
632+	632+ bootstrap method

# List of figures

An example of a ROC curve	2.1
ROC curves for perfect and uninformative of two tests	2.2
Predictiveness curves	5.1
ROC curves for <i>tnfa</i> and <i>il10</i>	5.2
ROC curve for <i>tnfa</i>	5.3
ROC curves for Cox model and logistic regression using PBC dataset	6.1
AUC estimates up to 10 years for PBC dataset	6.2
ROC curves at different times for PBC dataset	6.3
Survival curves for patients on ART and non ART for TBM/HIV dataset	7.1
Survival curves for TBM stages for TBM/HIV dataset	7.2
Survival curves for TBM diagnosis for TBM/HIV dataset	7.3
Survival curves for HIV status for TBM/HIV dataset	7.4

---

Observed values of multivariate signature for TBM/HIV dataset	7.5
AUC according to the prognostic times and the different estimators using the ridge-Cox model regression for TBM/HIV dataset	7.6
Ambiguous AUC estimation using optimal cut-off for HIV genetic dataset	8.1
Proportion of ambiguities for individuals HIV genetic dataset	8.2

# List of tables

Combinations of time dependent TPR and FPR	3.1
Mean of the AUC from the 1000 simulated samples for each scenario, $B = 1000$ bootstrap replications are performed for computing the cross-validation and the bootstrap cross-validation for the AUCs	5.1
AUC values for TB-IRIS biomarkers	5.2
Bootstrap cross-validation of AUC for composite marker from TB-IRIS	5.3
Mean of the time dependent AUC at 120 days obtained from 250 simulated samples under MCAR and MAR for each combination of censoring rate, predictive model and imputation method. $B = 1000$ bootstrap replications are performed for computing the $AUC_{cv}$ , $AUC_{bcv}$ , $SE$ and $CI$ s for the AUCs	6.1
Missing rate in PBC dataset	6.2
Cox regression estimates for PBC dataset	6.3

The $AUC_{cv}$ and $AUC_{bcv}$ estimations using mean imputation, nearest neighbor imputation and multiple imputation	6.4
Pooled estimates from MI for some variables in PBC dataset	6.5
Mean of the time dependent AUC at 6 months obtained from 250 simulated samples for each scenario $B = 200$ , bootstrap replications are performed for computing the apparent, BCV, 632, 632+ and 2-level bootstrap standard error(SE) for the AUCs	7.1
Mean of the time dependent AUC at 6 months obtained from 250 simulated samples for each combination of over-fitting level, censoring rate and penalized model, $B = 200$ bootstrap replications are performed for computing the apparent, BCV, 632 and 632+ estimations	7.2
The demographic, clinical and investigative findings for patients with definite, probable and possible TBM	7.3
Management and outcome in patients with TBM	7.4
Univariate and multivariate analyses of association between variables and in hospital mortality in all patients	7.5
Univariate and multivariate analyses for association with inpatient mortality for HIV-infected patients	7.6
Univariate and multivariate analyses for association with survival for all patients	7.7
Univariate and multivariate analyses for association with survival for HIV-infected patients	7.8

AUC values from different resampling methods for composite biomarker from TBM/HIV dataset	7.9
AUC estimations using LASSO and ridge methods for TBM/HIV dataset in different time points	7.10
Summary of the HIV-1 variables in recent and chronic patients	8.1
Univariate and multivariate analysis for genetic HIV data	8.2
AUC estimations for some variable in HIV genetic data	8.3



# Introduction

---

## 1.1. Motivation, purposes and objectives

---

One challenge in clinical medicine is that of the correct diagnosis of disease. It is patently undesirable to declare someone as being infected with a serious disease when in fact the individual is disease free and likewise undesirable to declare someone as disease-free when in fact the individual is diseased. Both errors have serious implications to the individual and the community at large. 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 diagnostic tests (Pepe [104]). The ROC curve was first used in signal detection theory (Egan [40]; and Green and Swets [59]). In the late 1980's, researchers started applying ROC curves methodology to medical diagnostic test evaluation (Hanley [61], Shapiro [119]). However the use of ROC curves in Radiology was earlier reported in the 1980s in a paper by Swets and Pickett [132]. In general the ROC analysis has been extended for use in visualizing and analysing the behavior of diagnostic systems (Swets [134]). 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 *area under the ROC curve* or the AUC is of interest in this era of multiple assessments. When we have several biomarkers we can combine them to obtain better diagnostic accuracy and improve the AUC by maximizing its value over all possible combinations of the biomarkers (Fang et al. [44]). We are interested in combining, selecting and evaluating biomarkers to estimate the AUC and predict specified diseases. For this purpose we use Logistic regression as it is commonly used when the outcome or response of the presence is binary. A Cox model is also used in case of a time to event outcome or response. In order to obtain a better AUC we applied *feature selection*, also known as *variable selection*, which is desirable in order to obtain an interpretable prediction rule. This is a technique of selecting a subset of relevant features for building models and it improves model performance. Two methods are used, namely stepwise selection and LASSO, the latter being very attractive as it simultaneously performs variable selection and shrinkage.

In our work, we use resampling procedures which 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 in estimating how accurately a predictive model will perform in practice.

In 1979 Efron [37] 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 drawing with replacement  $n$  observations from among the original set of  $n$  observations (unlike in the cross-validation). In addition to that we also use the 632+ bootstrap method which was proposed by Efron [37] and Efron and Tibshirani [39] in order to reduce the upward bias of the parameter of the *leave-one-out bootstrap* method.

Some methodologists have described the problem of missing data as one of the most important statistical and design problems in research. This is of greater concern when decisions are to be made about the appropriateness of the care a patient should receive and also when one is interested in using the predictive model to discriminate subjects as likely to have a certain characteristic from those who do not. Missing values can severely affect the results if there is dependence between the outcome and the missing data process, therefore dealing with missingness in the data becomes necessary. Current available methods in analysing *ROC* curves are limited to complete data sets and classical ROC analysis. In the development of prognostic models the presence of missing data is a frequently encountered problem. Thus we use the time dependent area under ROC curves to compare different imputation methods.

The main purpose of this thesis is to examine the performance of different resampling methods with a particular interest to cross-validation and Bootstrap methods to estimate the AUC for procedures that select and combine biomarkers and also to make inferences. We simulated

several multivariate models to evaluate the performance of the resampling approaches in this setting. We applied the resampling methods to data collected to study TB-IRIS by the Institute of Infectious Disease and Molecular Medicine (IIDMM), University of Cape Town, South Africa. The author was given permission to use the data through Professor Robert Wilkinson the lead PI in the project. TB-IRIS occurs in 8%–43% of HIV-infected patients receiving TB treatment after starting antiretroviral therapy (ART) [90, 95]. Baseline levels of five biomarkers were evaluated and this dataset was used 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.

In addition, we applied time dependent ROC analysis to data collected at GF Jooste Hospital in Cape Town - a secondary-level hospital. This was done to predict the survival of patients having meningitis in a high TB/HIV prevalence setting in this era of increasing availability of ART. The hospital serves high density low income patients: it is a 200-bed public sector referral hospital that serves adult patients from a community of approximately 1.3 million people. We use this data set to explain how well the predictor index of combined variables in TBM/HIV patients can accurately discriminate between the patients that may die during the first six months and those who may be still alive beyond that time. Moreover we have extended our discussion to genetic data on HIV drug resistance collected by the Genomics centre of the University of KwaZulu-Natal.

The thesis objectives can be expressed as follow:

- To provide a solid understanding of the diagnosis of diseases and the use of ROC curves for this purpose.
- To evaluate whether a combination of biomarkers can accurately discriminate between

two groups of patients - namely diseased and non-diseased subjects and examining the performances of different methods, with a particular interest in cross-validation and bootstrap cross-validation, as methods for the estimation of the AUC and its variance.

- To handle missing values in the data and to, compare the different imputation methods in evaluating different resampling methods and to estimate the time dependent area under ROC curves.
- To compare resampling methods with respect to predictive power. These methods were used together with the penalized Cox model using the LASSO method. We proposed ridge Cox model to estimate time dependent AUC using bootstrap methods. We also proposed two level bootstrapping techniques to estimate variances and evaluated these techniques through simulation studies.
- To apply ROC analysis to a genetic data set and to evaluate the effect of genetic ambiguities in biomarker detection of recent HIV infection. We then examine which variables are correlated with recently infected patients.

---

## **1.2. Background and related studies on ROC analysis**

---

The history of ROC curves goes back to the Second World War where the methodology was firstly used in analysing radar signals and later used in signal-detection theory (see Fawcett [47] or Green and Swets [59]). Since then the usage and applications of ROC curves has spread to many other fields such as psychophysics, medicine (Hanley [61], Shapiro [119]), epidemiology (Aoki et al. [8]), radiology (Metz [97]), social sciences and evaluation of machine learning techniques (Spackman [126]). ROC analysis is a very rich area for research and a large number of articles have been published in the last two decades.

The medical decision-making community has extensive literature on the use of ROC graphs for diagnostic testing (Zou [157]). Swets et al. [135] brought ROC curves to the attention of the wider public with their Scientific American article.

One of the earliest adopters of ROC graphs in machine learning was Spackman [126], 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., [60], Pepe [104], Shapiro [119] and Zhou et al. [156]). 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 curve in diagnostic testing. It is also used in non-diagnostic testing systems, for example the use of AUC in clinical trials (Hauck [64]) and in toxicology (Bosch et. al. [48]).

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

Dodd and Pepe [32] proposed a new method for making inferences about covariate effects on the performance of a classifier. The advantage of this approach is that “it can be simply applied by adapting standard binary regression methods as it requires fewer assumptions than

existing ROC regression methods”.

Zhang et al., [155] 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 [32] 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 [155] it is proposed that a set of AUC estimates be used to obtain parameter estimates and combine DeLong’s method [29] and the delta method for computing parameter standard errors. Their new method avoids the heavy computation associated with the method of Dodd and Pepe and hence is easy to implement.

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

Efron [38] 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 indicated that a 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. [44] considered the optimal linear combination that maximises the AUC and compared the estimation 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.

The authors of [152] proposed two easily-implemented algorithms, to find the best linear combination of multiple biomarkers that optimise the partial AUC (pAUC), for given a range of specificity values. Analysis of synthesized and real datasets shows that the proposed algorithms achieve larger predictive pAUC values on future observations than existing methods, such as Su and Lius method [130], logistic regression and others.

---

### **1.3. Thesis outlines**

---

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

Chapter 1, which is the current chapter 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 - the current section - describes the structure of this thesis.

In Chapter 2 we discuss the concept of Receiver Operating Characteristic (ROC) Curves. This chapter is divided into eight sections. We first give some important definitions and basic concepts in Section 2.1. Section 2.2 is mainly concerned with using ROC curves for continuous tests. In Section 2.3, we introduce four important indices of ROC curves; each is discussed in a separate subsection. Section 2.4 discusses binormal ROC curves. In Section 2.5, we discuss ROC curves for ordinal data. The ROC estimation is discussed in Section 2.6 and this is divided into three subsections. This chapter also briefly discusses the modeling of covariates effects on test results (Section 2.7) and on ROC curves (Section 2.8).

In Chapter 3, we discuss time dependent ROC curves. Section 3.1 is an extension of classical



sensitivity and specificity analysis in the context of time dependent ROCs. Section 3.2 explains time dependent true and false positive fractions. Section 3.3 illustrates the combination of time dependent true and false positive fractions.

In Chapter 4 we discuss missing values in data sets. Section 4.1 explains the reasons for missing values and this section is divided into four subsections. In Section 4.2 the imputation strategies are discussed; these are the mean imputation, nearest neighbor hot deck imputation and multivariate imputation via chain equation.

In Chapter 5 we start applying the proposed methods using a cross sectional data set. We give an introduction in the beginning of this chapter (Section 5.1). Section 5.2 discusses variable selection methods for biomarkers. This section is divided into three subsections. In Section 5.3, we discuss some of the resampling methods in the context of multiple biomarkers. We divided this section into ten subsections discussing over-fitting and various resampling methods, while the last two subsections contain two algorithms to obtain and estimate the AUC through cross-validation and bootstrapping respectively. The next two sections (5.4 and 5.5) discuss logistic regression and linear discriminant models respectively. Section 5.6 presents our proposed algorithm to obtain the AUC. In Section 5.7 we are concerned with simulation studies, while in Section 5.8 we apply resampling methods to a real dataset that has been collected from a study of tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS) in Cape Town.

Chapter 6 discusses the problem of combining multiple variables to estimate predictive accuracy; however the response variable in this chapter is time dependent. We also introduce the problem of estimation accuracy in presence of missing values in some variables. This chapter includes an introduction (Section 6.1), methods to address the missing data problem (Section 6.2), methods for estimating time dependent AUC (Section 6.3) and Section 6.4 discusses the models that have been used in the analysis. Section 6.5 is an algorithm. Section 6.6 presents

the simulation studies, followed by an application using real data (Section 6.7) and specific conclusions to the chapter (Section 6.8).

In Chapter 7 we are mainly interested in predicting the outcome for TBM/HIV dataset from Cape Town. Section 7.1 is an introduction. Section 7.2 discusses estimation methods for predictive accuracy using the penalized Cox model, followed by important resampling methods in Section 7.3. Section 7.4 is an algorithm. Section 7.5 presents the simulation studies, followed by an application to TBM in Section 7.6 and a conclusion (Section 7.7).

Chapter 8 evaluates proportion of ambiguities as a biomarker to predict recent HIV infection in rural KwaZulu- Natal, South Africa. This chapter includes three sections, where Section 8.1 is an introduction, Section 8.2 introduces the HIV data with genetic information and discusses the methods used and in Section 8.3 we give key results and a discussion.

Finally, Chapter 9 is a conclusion to the thesis where also suggest some future work that can be done as an extension to the current work.

We would like to mention that key publications out of this thesis are under review and preparation. These are:

1. M. B. Elshareef, L. Dodd and H. G. Mwambi, *Combining multiple biomarkers in diagnostic testing with an application to TB disease data from Cape Town*, submitted to African Health Sciences.
2. M. B. Elshareef and H. G. Mwambi, *Predictive accuracy of multiple time dependent biomarkers with missing values in diagnostic testing*, submitted to Pakistan Journal of Statistics.
3. M. B. Elshareef and H. G. Mwambi, *The role of ambiguous nucleotides as biomarkers of*

*recent HIV infection in rural KwaZulu-Natal, South Africa*, to be submitted.

4. M. B. Elshareef and H. G. Mwambi, *Use of resampling methods to predict the outcome in tuberculous meningitis in a high HIV prevalence patients in South Africa*, to be submitted.
5. M. B. Elshareef and H. G. Mwambi, *Classic and time dependent AUC estimations: A survey study*, in preparation.

# 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 [104] supplemented with our own understanding of the problem.

---

## 2.1. Definitions and basic concepts

---

In this section we present 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 test result-true status outcome combinations. 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 test result-true status combinations 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  $TNF = \frac{TN}{TN + FP}$ .

**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

$$TPF = P(Y = 1|D = 1) = TP/(TP + FN) \quad \text{and} \quad (2.1)$$

$$TNF = P(Y = 0|D = 0) = TN/(TN + FP). \quad (2.2)$$

Sensitivity and specificity describe how well the test discriminates between patients with and without disease. In fact we are also interested in the probability of disease, given a positive test result and likewise the probability of no disease given a negative test result. This leads to two predictive values of the test formally defined below.

**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 the disease divided by the probability of that the 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

$$LR^+ = P(Y = 1|D = 1)/P(Y = 1|D = 0) \quad \text{and}$$

$$LR^- = P(Y = 0|D = 1)/P(Y = 0|D = 0).$$

**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

---

### 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

$$TPF(c) = P[Y \geq c|D = 1], \tag{2.3}$$

$$FPF(c) = P[Y \geq c|D = 0], \tag{2.4}$$

respectively.

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

$$ROC(.) = \{(FPF(c), TPF(c)) | c \in \mathbb{R}\}. \tag{2.5}$$

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 of  $c \rightarrow \infty$ , 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 [104] and Figure 2.1 below).

The ROC curve can also be written in the form (see Pepe [104]):

$$ROC(.) = \{(t, ROC(t)) \mid t \in (0, 1)\}, \quad (2.6)$$

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

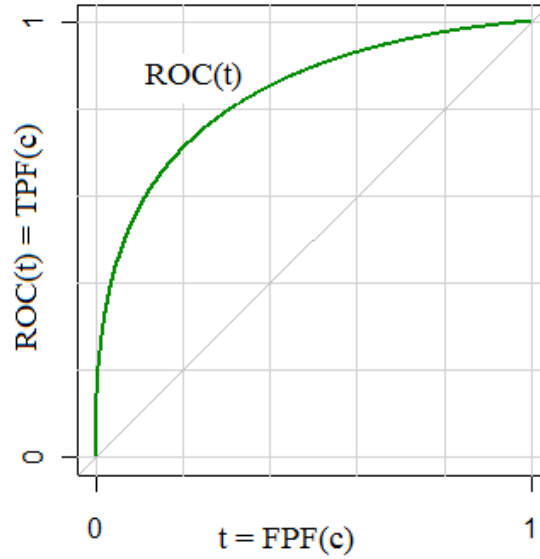


Figure 2.1: A sketch of a ROC curve

### 2.2.1 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 2.2.

On the other hand an *uninformative* 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 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) and the area under such curve is 0.5).

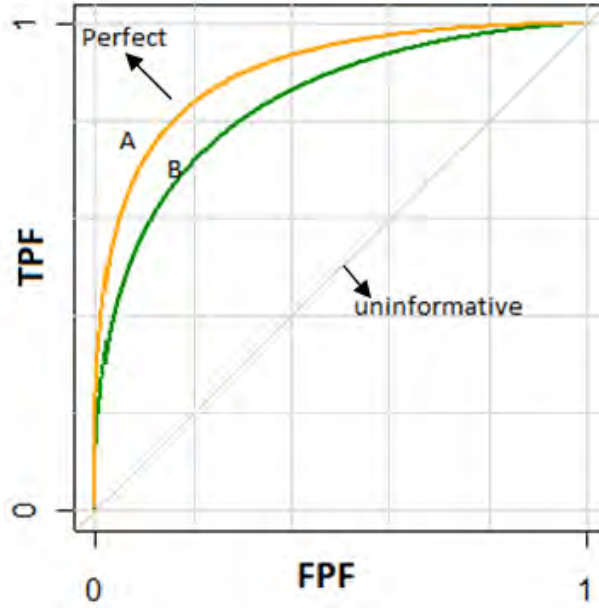


Figure 2.2: ROC curves for perfect, uninformative and two tests  $A$  and  $B$ . Test  $A$  is better than  $B$

In the following proposition we quote some important results from Pepe [104].

**Proposition 2.2.1.** (i) *The ROC curve is invariant to strictly increasing transformations of  $Y$ ,*

(ii) *if  $S_D$  and  $S_{\overline{D}}$  denote the survivor function for  $Y$  in diseased and non-diseased populations, where  $S_D(y) = P[Y \geq y | D = 1]$  and  $S_{\overline{D}}(y) = P[Y \geq y | D = 0]$ , then the ROC curve*

can be represented as follows:

$$ROC(t) = S_D(S_D^{-1}(t)), \quad t \in (0, 1), \quad (2.7)$$

(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 [104]. Now to show Equation (2.7), let  $c = S_D^{-1}(t)$ , 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  $ROC(t) = TPF = S_D(c) = S_D(S_D^{-1}(t))$ . Hence the result. ■

We conclude this section by listing some of the important attributes of the ROC curves. These attributes have been listed in Table 4.1 of Pepe [104] and in Fawcett [47]. 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.

---

**2.3. 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 [119], Greiner et al., [60], Zhou et al., [156] and Pepe [104].

**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 statistic 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

$$AUC = \int_0^1 ROC(t)dt. \quad (2.8)$$

We note from Equation (2.8) that the AUC is a portion of the area of the unit square. Hence its value is always bounded 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 disease.

**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 [13], Hanley and McNeil [61], Pepe [104]). 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 [30] suggested the use of a weighted average approach, weighting certain parts of  $FPF$  domain more than others.

### 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 a 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 normalised value of  $pAUC$  is defined to be  $pAUC(t_0)/t_0$  and it clearly ranges from  $t_0/2$  to 1 for uninformative and perfect tests respectively. The normalised  $pAUC$  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.

A more general formula for Equation (2.9) has been given in Dodd and Pepe [31].

#### 2.3.4 Kolmogorov-Smirnov (KS) index

The maximum vertical distance between the  $ROC$  curve and  $TPF = FPF$  is an index we refer to as the KS index. 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 with survival functions  $S_D$  and  $S_{\bar{D}}$  for two tests  $Y_D$  and  $Y_{\bar{D}}$  respectively (Gail and Green [52]). In fact we identify the index KS with Kolmogorov-Smirnov measure.

Another well-known measure is the Youden index, which is a special case of Kolmogorov-Smirnov measure. For more information on this index, refer to Fluss [49].

## 2.4. 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 2.4.1.** *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*

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)), \quad (2.10)$$

where  $a = \frac{\mu_D - \mu_{\bar{D}}}{\sigma_D}$ ,  $b = \frac{\sigma_{\bar{D}}}{\sigma_D}$  and  $\Phi$  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

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

and this completes the proof. ■

Thus 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 2.4.1.** 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, for  $t \in (0, 1)$ .

Thus  $b \neq 1$  leads to ill behaved ROC curve. Therefore the fact that the binormal ROC curve may not have the monotone likelihood ratio raises some concern about using it for approximation of real data. However Swets [133] and Hanley [62] and [63] showed that a binormal ROC curve is a good approximation in practice.

We have seen in Proposition 2.2.1 that the ROC curve is invariant to monotone increasing data transformations. Therefore if  $Y_D$  and  $Y_{\overline{D}}$  have normal probability distributions and if we let  $W_D = h(Y_D)$  and  $W_{\overline{D}} = h(Y_{\overline{D}})$ , 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 simple means that for some strictly increasing transformation  $h(\cdot)$ , the functions  $h(Y_D)$  and  $h(Y_{\overline{D}})$  have normal distributions (see Pepe [104]).

Although the binormal form is the classic parametric form for ROC curves, other parametric forms can be adopted. Any parametric form adopted can be fitted using ordinal data likelihood methods. Usually the AUC summary indices are used as the basis for comparing binormal ROC curves. The standard error of the difference is calculated using the delta method and alternative



summary indices can likewise be used. Metz [97] suggested that instead of comparing summary indices, the fitted ROC parameters can instead be compared.

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

**The binormal AUC:** The AUC for the binormal ROC curve is given by (Pepe [104])

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

**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. ■

### Robustness of the binormal estimator

The choice of the binormal estimator to fit a ROC curve is usually justified by theoretical considerations, mathematical tractability, familiarity with the normal model or just by convenience.

The word robust can have many different meanings. Here it is used in the sense of robust statistics, i.e. meaning that in the presence of a certain amount of observations coming from a non-normal distribution the binormal estimator will yield reliable results. Robustness, in Swets [135] and Hanley [63], is understood as the ability of the binormal estimator to fit a ROC curve that looks right in comparison either with the theoretical ROC curve or with the observed rating method.

---

## **2.5. The ROC for ordinal tests**

---

### **2.5.1 Ordered discrete tests results**

Although many of test results are continuous, some tests yield discrete results. For example the minimal inhibitory concentration of an antibiotic as standard measure of bacterial residence is measured on a continuous scale, but some questionnaire reporting systems may also yield discrete numeric results. However it should be noted that many tests are not numeric at all, for example in the assessment of an image by a radiologist. In this case the radiologist assessment that the disease is present is classified on an ordinal scale.

The key difference between qualitative assessments that are measured on ordinal scales and quantitative assessments made on numeric scales is the recognition that different assessors can use the ordinal scale differently. For example an image considered as highly suspicious for cancer by one radiologist may be considered as possibly malignant by another even though both have the same perception of the image. ROC analysis has been very popular for use with rating data as it helps to disentangle the inherent discriminatory capacity of the test or imaging device from the particular use of the scale by the assessor [104].

### 2.5.2 The latent decision variable model

Suppose that there is an unobserved latent continuous variable  $L$  corresponding to the assessor's perception of the image. The reader or assessor has some thresholds values that correspond to his/her classification or rating of the image. If  $Y$  denotes the reported classification, then:  $c_{y-1} < L < c_y, y = 1, \dots, P$ , where  $c_0 = -\infty$  and  $c_P = \infty$  which yields a  $P$ -level ordinal variable. The reader classifies the image in the  $y^{th}$  category if  $L$  falls within the interval corresponding to his/her implicit definition for the  $y^{th}$  category in terms of the latent variable  $L$ . Different raters might perceive an image in the same way but classify it differently because their implicit decision thresholds may be different. The ROC curve for  $L$  can be defined as follows: If  $Y \geq y$  and  $L > c_{y-1}$  then we can denote the true and false positive fractions as  $TPF(c_{y-1})$  and  $FPF(c_{y-1})$  respectively. Then the ROC curve in terms of  $L$  can be identified for the  $P + 1$  points and represented as follows:  $(FPF(c_{y-1}), TPF(c_{y-1})), y = 1, \dots, P + 1$ .

The latent variable framework with decision thresholds that give rise to observed test results, is an appealing conceptual model. However strictly speaking the latent variable does not have an explicit clinical meaning and thus the interpretations of the ROC curve for  $L$  are somewhat dubious. Also, the available set of points for the ROC curve are only the set of discrete observable points and thus the curve is not fully identifiable.

### 2.5.3 The discrete ROC curve

One popular approach to ROC analysis for ordinal data is to simply define the ROC curve as a discrete function. This alternative curve is defined as

$$ROC = (t_y, ROC(t_y)), y = 1, 2, \dots, P + 1. \quad (2.11)$$

In Equation (2.11),  $t_y = P[Y \geq y|D = 0]$  and  $ROC(t_y) = P[Y \geq y|D = 1]$ . The first cutoff point or threshold at  $y = 1$  is the corner point  $(0,0)$  while the threshold at  $y = P + 1$  is the  $(1,1)$  corner point. This difference between this definition and the one for continuous results is that the set of possible false positive fractions (domain) is finite. One of the binormal form with discrete domain  $T = \{t_y | y = 1, 2, \dots, P + 1\}$  that can be used for discrete ROC function is given by

$$ROC(t_y) = \Phi(a + b\Phi^{-1}(t_y)), y = 1, 2, \dots, P + 1. \quad (2.12)$$

The ROC curve for a discrete decision variable is not like the one for continuous decision variable. It serves more as a visual aid to depict an ROC function associated with the discrete observed decision variable  $Y$ . Moreover the discrete domain is required together with the parameters  $a, b$  to completely characterise the discrete ROC function. A very important point about discrete ROC functions is that two ROC functions differ if their domains differ, even if their points lie on the same smooth curve. The discrete ROC analysis requires, in addition to the trade-off between true and false positive fractions, consideration of the *FPFs* that are attainable with the test. This is different from continuous tests as all *FPFs* in the range  $(0, 1)$  are attainable.

For the discrete ROC functions the summary measure cannot be applied directly. A summary of the discrete ROC curve can be calculated by joining the points and then calculating the area relative to the resulting curve, but they are difficult to interpret. When the linearity is used for joining points, the area under that curve has an interesting interpretation of the probability of correctly ordering diseased and non-diseased observation. Thus formally the area under the ROC function based on linear interpolation between points (AUC) is:

$$AUC = P[Y_D > Y_{\bar{D}}] + \frac{1}{2}P[Y_D = Y_{\bar{D}}]. \quad (2.13)$$

When case and control values are tied, the ROC curve has simultaneous horizontal and vertical

jumps and thus the AUC can be calculated as an averages of tie-corrected percentile values. The second part of the RHS of the above equation represents a situation of indistinguishable of diseased and non-diseased cases as a result of discretisation.

Finally we would like to conclude this section by mentioning that the ROC curve is the most popular tool for describing the accuracy of a continuous or ordinal valued tests. The ROC curve has been popular for a long time in radiology research, in addition to the fact that it provides a description of separation between distributions and is still very useful in clinical trials research. The binormal ROC curve described before in Section 2.4 is the classic parametric model, but one of the weakness of the model is that in some applications it may not be concave.

---

## 2.6. The ROC curve estimation

---

This section introduces the statistical methodology for making inference about the ROC curve from data. We describe three approaches for estimating the ROC curve and its summary indices [104]. The first method is based on applying non-parametric empirical methods to the data to obtain the empirical ROC curve from which the empirical summary measures can be calculated especially for continuous test results. The second approach is by modeling the distributions of  $Y_D$  and  $Y_{\overline{D}}$ , after which the parameters in these distribution are estimated and then the induced ROC curve is calculated. However this approach requires strong assumptions about the form of the distributions of test results which make it less popular. The ROC curve is concerned only with the relationship between the distributions of  $Y_D$  and  $Y_{\overline{D}}$ . The third approach is to use a smooth parametric function of the ROC curve rather than modeling the distributions (second approach). The parameters of the third approach are estimated from the rankings of the test results for diseased and non-diseased subjects.

We assume that the data can be presented as test results for  $n_D$  cases and  $n_{\overline{D}}$  controls as

follows:  $Y_{D_i}, i = 1, \dots, n_D$  and  $Y_{\overline{D}_j}, j = 1, \dots, n_{\overline{D}}$ . We assume that  $Y_{D_i}$  and  $Y_{\overline{D}_j}$  are selected randomly from the populations of test results associated with diseased and non-diseased outcomes, respectively.

### 2.6.1 Non-parametric empirical estimator

The aim is to apply non-parametric empirical methods to the data to obtain an empirical ROC curve, which is a popular choice for continuous tests results settings. The empirical estimator of the ROC curve simply applies the definition of the ROC curve to the observed data. For each  $c$  the empirical true and false positive fractions are calculated as follows:

$$\widehat{TPF}(c) = \sum_{i=1}^{n_D} I[Y_{D_i} \geq c] / n_D, \quad (2.14)$$

$$\widehat{FPF}(c) = \sum_{j=1}^{n_{\overline{D}}} I[Y_{\overline{D}_j} \geq c] / n_{\overline{D}}, \quad (2.15)$$

where  $I(A)$  is a function which takes value 1 when  $A$  is true and 0 otherwise. Then it follows we can write the empirical ROC denoted by  $\widehat{ROC}_e$  as:

$$\widehat{ROC}_e(t) = \hat{S}_D(\hat{S}_{\overline{D}}^{-1}(t)), \quad (2.16)$$

where  $\hat{S}_D$  and  $\hat{S}_{\overline{D}}$  are the empirical survivor functions for  $Y_D$  and  $Y_{\overline{D}}$ , respectively. The empirical ROC curve is a function based only on the rank of the data. That is it depends on the relative ordering of the test results and their status as being from diseased and non-diseased individuals. Therefore the empirical ROC curve is invariant to strictly increasing transformations of the data. That is if  $Y_D > Y_{\overline{D}}$ , then  $h(Y_D) > h(Y_{\overline{D}})$ , where  $h(Y)$  is the increasing transformation of  $Y$ .

The empirical AUC given by  $\int_0^1 \widehat{ROC}_e(t) dt$ , can be considered as a Mann-Whitney U-statistic

calculated as a double summation given below

$$\sum_{j=1}^{n_{\overline{D}}} \sum_{i=1}^{n_D} [I[Y_{D_i} > Y_{\overline{D}_j}] + \frac{1}{2} I[Y_{D_i} = Y_{\overline{D}_j}]] / n_D n_{\overline{D}}. \quad (2.17)$$

Following Pepe and Cai [105], the ROC curve can be interpreted as a cumulative distribution function for the discriminatory measure  $Y$  in the affected population ( $D = 1$ ) after  $Y$  has been standardised to the distribution in the reference population ( $D = 0$ ). The standardised values are called *placement* values. Using the distribution of  $Y_D$  as the reference distribution the placement for  $y$  in the diseased population is defined as  $P[Y_D \geq y] = S_D(y)$  and then the empirical placement value is given by  $\hat{S}_D(y)$ . We can also write  $\widehat{AUC}_e$  as the sample average of empirical disease placement values for non-disease observation:

$$\widehat{AUC}_e = \sum_{j=1}^{n_{\overline{D}}} \frac{\hat{S}_D(Y_{\overline{D}_j})}{n_{\overline{D}}}.$$

The variance of a summary measure such as  $\widehat{ROC}_e$  is often complicated and in practice bootstrapping is used to calculate confidence intervals.

DeLong et al. [29] proposed an expression for  $var(\widehat{AUC}_e)$  in terms of the variability of placement values. This method provides a nice computational algorithm for estimating the variance of  $\widehat{AUC}_e$  in a large sample given by:

$$var(\widehat{AUC}_e) = \frac{var(S_D(Y_D))}{n_D} + \frac{var(S_D(Y_{\overline{D}}))}{n_{\overline{D}}} \quad (2.18)$$

Which is estimated using the sample variances for the empirical standardized values and given by:

$$var(\widehat{AUC}_e) = \frac{\widehat{var}(S_D(Y_{D_i}))}{n_D} + \frac{\widehat{var}(S_D(Y_{\overline{D}_j}))}{n_{\overline{D}}} \quad (2.19)$$

The variance is directly a function of the variability in the placement values of diseased observations within the non-diseased reference and of non-diseased observations within the reference distribution diseased. The confidence interval for the AUC can be given by:

$$\widehat{AUC}_e \pm \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \sqrt{\widehat{var}(\widehat{AUC}_e)}. \quad (2.20)$$

An asymmetric confidence interval that guarantees an interval in  $(0, 1)$  is preferred. Thus we can use a logistic transformation to compute the confidence interval for logit AUC ( $\log(\text{AUC}/(1 - \text{AUC}))$ ) which is given by:

$$\log \left( \frac{\widehat{AUC}_e}{1 - \widehat{AUC}_e} \right) \pm \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \frac{\sqrt{\text{var}(\widehat{AUC}_e)}}{\widehat{AUC}_e(1 - \widehat{AUC}_e)} \quad (2.21)$$

An alternative representation of the asymptotic variance was derived by Hanley and McNeil [61]. Assume that there are no ties between diseased and non-diseased observations, so that formula (2.17) simplifies to  $\widehat{AUC}_e = \sum_{j=1}^{n_{\overline{D}}} \sum_{i=1}^{n_D} I[Y_{Di} \geq Y_{\overline{D}_j}] / n_D n_{\overline{D}}$ . Then the variance of the empirical ROC can be defined as:

$$\text{var}(\widehat{AUC}_e) = \frac{1}{(n_D n_{\overline{D}})} \text{AUC}(1 - \text{AUC}) + (n_D - 1)(Q1 - \text{AUC}^2) + (n_{\overline{D}} - 1)(Q2 - \text{AUC}^2) \quad (2.22)$$

where

$$Q1 = P[Y_{Di} \geq Y_{\overline{D}_j}, Y_{D_{i'}} \geq Y_{\overline{D}_j}],$$

$$Q2 = P[Y_{Di} \geq Y_{\overline{D}_j}, Y_{Di} \geq Y_{\overline{D}_{j'}}],$$

$(Y_{Di}, Y_{D_{i'}})$  and  $(Y_{\overline{D}_j}, Y_{\overline{D}_{j'}})$  denote random pairs of observations from the diseased and non-diseased populations, respectively. Observe that empirical estimates of each component are easily calculated to yield a variance estimator.

The empirical methods can be used for continuous or ordinal test results data. The methods that rely on the ROC curve being defined as curves with domain on the interval  $(0, 1)$  and therefore apply only to ROC curves for continuous tests.

The most commonly used statistic for comparing two ROC curves when test results are continuous is based on the difference in empirical AUC estimates. We denote the two curves by  $ROC_A$  and  $ROC_B$ , then the estimated difference is given by,

$$\Delta \widehat{AUC}_e = \widehat{AUC}_{Ae} - \widehat{AUC}_{Be}.$$



The null hypothesis  $H_0 : ROC_A = ROC_B$  is tested by comparing the value of  $\frac{\widehat{\Delta AUC_e}}{\sqrt{\widehat{var}(\Delta AUC_e)}}$  with standard normal distribution tails values. If data for the two ROC curve estimates are derived from independent samples, then

$$var(\widehat{\Delta AUC_e}) = var(\widehat{AUC_{Ae}}) + var(\widehat{AUC_{Be}}).$$

Other summary indices, estimated empirically can likewise be used as the basis for a non-parametric comparison of ROC curves, with resampling methods employed for formal statistical inference.

We can also use empirical methods to estimate discrete ROC functions for tests with ordinal results. The  $\widehat{ROCe}$  is defined as in previous section except that linear interpolation between estimated ROC points  $(F\hat{P}F(y), T\hat{P}F(y))$ ,  $y = 1, \dots, P$  is not performed. In finite samples one will not know what false positive fractions are attainable with a discrete test. Hence fixing  $t$  and making inferences about  $ROC(t)$  is not feasible in the same way that it is for continuous test.

For discrete data, the empirical AUC index is usually calculated using a linear interpolation between ROC points and the trapezoidal rule. Although, it is not interpreted as an area under the curve because the ROC for an ordinal test is a discrete function not a curve. The AUC as a summary index for the discrete ROC function is identical to the Mann-Whitney U-statistic and its interpretation as the probability  $P[Y_D \geq Y_{\overline{D}}]$ .

Similarly, comparisons between AUCs can be made as before, however it is not always sensible to compare AUCs with two discrete ROC functions. Differences between two AUCs may exist that are simply caused by the fact that their domains are different. Empirical methods can be used for continuous or ordinal test result data.

### 2.6.2 Modeling test results

#### Fully parametric modeling

The  $ROC(t)$  can be estimated through constituent distribution functions parametrically and to calculate the induced ROC curve estimate. The fully parametric method makes strong assumptions about the forms of the distributions,  $S_D$  and  $S_{\bar{D}}$ . Suppose that we model each distribution as a parametric distribution with parameters  $\alpha$  and  $\gamma$  for non-diseased and diseased populations, respectively:

$$S_{\bar{D}(y)} = S_{\alpha, \bar{D}}(y)$$

and

$$S_{D(y)} = S_{\gamma, D}(y).$$

Then the resultant ROC estimate is

$$\widehat{ROC}_{\hat{\alpha}, \hat{\gamma}}(t) = S_{\hat{\gamma}, D}(S_{\alpha, \bar{D}(t)}^{-1}). \quad (2.23)$$

The standard error for  $\widehat{ROC}(t)$  can be calculated using the variance of  $(\hat{\alpha}, \hat{\gamma})$  and delta method.

The ROC estimate will be fully efficient assuming that the models are correctly specified and the parameters are estimated with maximum likelihood methods.

Comparing ROC curves in this framework is not easy because one cannot simply compare parameters. The same ROC curve can result from different pairs of constituent test result distributions. Comparing parameters of the distributions  $\alpha_A$  and  $\alpha_B$  and  $\gamma_A$  and  $\gamma_B$  for two curves indexed by  $A$  and  $B$  does not achieve a comparison of ROC curves. Wieand [150] evaluated the comparison of two ROC curves, with the difference in AUC indices estimated with fully parametric normal models and this approach was compared with that based on the non-parametric  $\widehat{\Delta AUC_e}$ . The parametric model is more efficient as expected.

### Semiparametric models

For the test results  $Y_{D_i}$  and  $Y_{\bar{D}_j}$ , the semiparametric location scale model for independent errors can be given by:

$$Y_{D_i} = \mu_D + \sigma_D \varepsilon_i,$$

$$Y_{\bar{D}_j} = \mu_{\bar{D}} + \sigma_{\bar{D}} \varepsilon_j,$$

where  $\varepsilon$  are independent errors, mean = 0 and variance = 1 random variables with survivor function  $S_0$ . Pepe [104] suggested to leave  $S_0$  unspecified.

The ROC curve can be written using the location-scale model as:

$$ROC(t) = S_0\left(\frac{\mu_{\bar{D}} - \mu_D}{\sigma_D}\right) + \left(\frac{\sigma_{\bar{D}}}{\sigma_D}\right) S_0^{-1}(t). \quad (2.24)$$

The empirical survivor function is a consistent estimator and is given by:

$$\hat{S}_0(y) = \frac{1}{n_D + n_{\bar{D}}} \left[ \sum_i^{n_D} I\left[\frac{Y_{D_i} - \hat{\mu}_D}{\hat{\sigma}_D} \geq y\right] + \sum_j^{n_{\bar{D}}} I\left[\frac{Y_{\bar{D}_j} - \hat{\mu}_{\bar{D}}}{\hat{\sigma}_{\bar{D}}} \geq y\right] \right]. \quad (2.25)$$

Then  $R\hat{O}C(t)$  estimator is given by:

$$R\hat{O}C(t) = \hat{S}_0\left(\frac{\hat{\mu}_{\bar{D}} - \hat{\mu}_D}{\hat{\sigma}_D}\right) + \left(\frac{\hat{\sigma}_{\bar{D}}}{\hat{\sigma}_D}\right) \hat{S}_0^{-1}(t). \quad (2.26)$$

The form of the function  $S_0(\cdot)$  is not specified, thus this model is semiparametric.

In [104], Pepe mentioned that the idea of modeling test results in order to estimate the ROC curve is somewhat unnatural. The ROC curve is invariant to monotone increasing transformations of the test results measurement. However the parametric and semi parametric methods that model the test results in order to estimate the ROC are not invariant to such data transformations. They are not distribution free in the sense that the ROC curve relies on the distributional forms for both  $S_D$  and  $S_{\bar{D}}$ , not just on their relationship or separation.

Modeling the test results can be restated as modeling the quantiles of  $S_D$  and  $S_{\overline{D}}$ . The main advantages of this approach of modeling test results over the empirical methods are firstly, the ROC curves are smoother and secondly, there is potential for increased statistical efficiency.

### 2.6.3 Parametric methods

We discussed non-parametric methods for making statistical inferences (see Subsection 2.6.1) and then discussed an approach that modeled the distributions of test results in order to estimate ROC curve in Subsection 2.6.2. We now discuss strategies that are intermediate between these two methods [104].

The first approach is suggested by Metz et al. [97] in which the authors described a parametric distribution free which is an extension of the one for ordinal data. One way to deal with continuous data is to categorize them into a finite number of predefined categories and to apply methods for fitting parametric ROC curves to ordinal data. Note that the ordinal data methods only make assumptions about the parametric form for the ROC curve. No assumptions are made about the survivor function  $S_{\overline{D}}$  for the discrete test result  $Y_{\overline{D}}$ .

Another rank based estimator is suggested by Pepe [104]. She parameterised the form of the ROC curve without making additional assumptions about the distributions of test results. This approach produces smooth parametric ROC curves but does not require that the test result distributions be modeled, rather they are based only on the ranks of the data. We previously defined the ROC as:

$$\begin{aligned} ROC(t) &= P[Y_D > S_{\overline{D}}^{-1}(t)] \\ &= P[S_{\overline{D}}(Y_D) \leq t]. \end{aligned}$$

Writing  $U_{it} = I[S_{\overline{D}}(Y_{D_i}) \leq t]$ , the binary variable denoting whether or not the placement value

exceeds  $t$  we see that:

$$\begin{aligned} E(U_{it}) &= P[S_{\overline{D}}(Y_{\overline{D}}) \leq t] \\ &= ROC(t) \end{aligned}$$

Now a parametric form for the ROC curve can be expressed as:

$$g(ROC(t)) = \sum_s \alpha_s h_s(t) \quad (2.27)$$

where  $g$  is a link function and  $h = h_1, \dots, h_S$  are specified functions. As a special case the binormal model is specified when  $g = \Phi^{-1}$ ,  $h_1(t) = 1$  and  $h_2(t) = \Phi^{-1}(t)$ . The ROC model in Equation (2.27) defines a generalised linear model for  $U_{it}$  with link function  $g$  and covariates  $h_s(t), s = 1, \dots, S$ . The ROC-GLM approach is designed to use procedures for fitting generalised linear models to binary data in order to estimate the parameters  $\alpha_s, s = 1, 2, \dots, S$ . For a set  $T = t \in T$  over which the model is to be fitted the empirical placement values  $\hat{S}_{\overline{D}(Y_{D_i})}, i = 1, \dots, n_D$  are calculated. For each  $t \in T$  the binary indicators based on the empirical placement values:

$$\hat{U}_{it} = I[\hat{S}_{\overline{D}}(Y_{D_i}) \leq t],$$

for  $i = 1, \dots, n_D$ . Binary regression methods with link function  $g$  and covariates  $h(t) = h_1(t), \dots, h_S(t)$  provide estimates of  $\alpha_1, \dots, \alpha_S$  from the data arranged as  $n_D \times n_T$  where  $n_T$  is the number of points in  $T$

$$\hat{U}_{it}, h_1(t), \dots, h_S(t), t \in T, i = 1, \dots, n_D.$$

The ROC-GLM procedure is based only on the ranks of the data and requires a model only for the ROC curve, not for the distributions of the test results. Hence it is a parametric rank based distribution free method. Pepe investigated the efficiency of ROC-GLM for estimating a binormal ROC curve with ordinal data. She found that its performance was close to the

method proposed by Dorfman [33]. The main advantage of ROC-GLM over the Dorfman [33] method is that it is easier and faster computationally when data are continuous.

The ROC-GLM method provides estimates of parameters and an estimate of variance covariance matrix via resampling. Pointwise confidence intervals for  $\text{ROC}(t)$  can therefore be constructed. Models can be fitted simultaneously for multiple curves and hypothesis tests for parameter estimates can also be done.

Now we conclude this section by mentioning that [104]:

- Three approaches are described: empirical methods, distribution free parametric methods and distributional modeling methods.
- The empirical and distribution free methods are based only on the ranks of data while the latter is not.
- The distinction between the empirical and distribution free methods is that the the former places no structure on the ROC curve while the later assume a parametric form for it.
- To compare two ROC curve when parametric form is assumed then the comparison can be based on the estimated parameters and their standard errors.
- To traditionally compare two ROC curves, differences in AUCs are typically calculated using the empirical or parametric distribution free methods.

---

## **2.7. Modeling covariate effects on test results**

---

The first approach for evaluating covariate effects on the ROC curve was proposed by Tosteson and Begg [141]. Their development was geared specifically towards ordinal data but it actually

applies more generally. The basic idea is to model  $S_{D,Z}$  in addition to  $S_{\bar{D},Z}$  and then calculate the induced covariate specific ROC curve,  $ROC_Z(t) = S_{D,Z}(S_{\bar{D},Z}^{-1}(t))$  for any particular covariate values  $Z$  of interest. A comprehensive model that includes both  $Y_D$  and  $Y_{\bar{D}}$  in one model by incorporating disease status as a covariate can be fitted.

The strategy of modeling the test results distributions and calculating induced ROC curves is the longest established approach for evaluating covariate effects on the ROC. It is popular in part because distributional modelling is a familiar task for statisticians. However it does not yield simple ways of summarizing covariate effects on the ROC curve.

---

### 2.8. Modeling covariate effects on ROC curve

---

We can model the covariate effects on ROC curves directly by modeling the ROC curve itself. There are several advantages to direct modeling of the ROC approach.

- The interpretation of model parameters pertains directly to the ROC curves.
- Multiple tests can be accommodated simultaneously.
- It provides a mechanism for comparing two tests even when their results are quantified in different units.

A ROC regression model to quantify covariate effects on the ROC curves has two components. These are the covariables  $X$  and secondly a formulation for the ROC curve as a function of  $t$ . Let  $h_0(\cdot)$  and  $g(\cdot)$  denote monotone increasing functions on  $(0, 1)$  then the equation

$$g(ROC_Z(t)) = h_0(t) + \beta X,$$

with  $t \in T_Z \subset (0, 1)$  is an ROC-GLM regression model. The link function  $g$  is specified as part of the model and the baseline function  $h_0(t)$  is unknown. A parametric form for it can be

specified or it can remain completely unspecified.

The restriction on  $h_0$  and on  $g$  are meant to ensure that  $ROC_Z(t)$  is an ROC curve in the sense that the domain and range are in  $(0, 1)$  and that it is increasing in  $t$ . However the model need not be defined for the entire interval  $t \in (0, 1)$  but possibly only on a proper or “concave” subset [40] and that the subset can vary with  $Z$ . Thus these models are applicable to ordinal tests where  $T_Z$  denotes the attainable false positive fractions for the test operating at covariate value  $Z$ .



# Time dependent receiver operating characteristic curves

In this chapter we will discuss time dependent ROC curves. In event-time analysis both time to event and the binary outcome are observed. Assuming the event occurs, ( $Y = 1$ ) then the time to event is uncensored whereas if an individual is followed up and the event did not occur ( $Y = 0$ ) then the time to the event is censored. Event-time analysis is used to describe data that correspond to the time from a time origin until the occurrence of the specified event or end point. For example the time origin in medical research will often correspond to the recruitment of an individual into an experimental study, the end point may be the death of the patient, relief of pain and so on.

Event-time often refer to the development of a particular symptom or to relapse after remission of a disease, as well as to the time to death. A significant and important feature of event-time analysis studies is that the event of interest is very rarely observed in all subjects. Such event-times are termed censored, to indicate that the period of observation was cut off before the event of interest occurred. As an example suppose that in clinical trial, a patient moves to another part of the country and can no longer be traced. In this case the time when the

individual experiences the event will not be known hence the time to event will be censored for such an individual.

The purpose of the ROC analysis is to characterise the prognostic potential of a marker (or model) by focusing on the correct classification rates. The ROC can be extended to time dependent ROC curves when we have time to event data.

The classical ROC curve deals with dichotomous diagnostic tasks (presence or absence of disease at a given time), we call this a cross sectional data type. In the real world we often deal with disease outcomes that depend on time and in this case the ROC curve will be a function of time. In time to event studies, the end point is subject to censoring and ignoring censoring will introduce bias. There are many existing proposed methods that accommodate censored data, which we will discuss later in this chapter (Section 3.3).

Time dependent ROC curves entail extending the concepts of sensitivity and specificity to time dependent binary variables such as vital status, allowing characterisation of diagnostic accuracy for censored event-time outcomes.

For test results defined on continuous scales, the ROC curves are standard summaries of accuracy. As described in Section 2.2, suppose  $Y$  denotes the diagnostic test or marker, with higher values more indicative of disease and  $D$  is a binary indicator of disease status, then the ROC curve for  $Y$  is a plot of the sensitivity associated with the dichotomized test  $Y > c$  versus  $(1 - \text{specificity})$  for all possible threshold values  $c$ .

Diagnostic tests are often developed to detect or predict the occurrence of an event, such as the onset of cancer, infection and so forth. In this context  $D$  is a time dependent variable. The time at which the diagnostic test is performed relative to the incidence of the event - or outcome in general - has a big influence on its operating characteristics.

**Example: Mother to infant transmission of HIV-1 trial [127]**

A biomarker evaluation for predicting the after birth HIV-1 infection events for babies delivered by HIV-1 infected mothers, is a common example for the applications of time dependent ROC curves. The study was carried out between November 1997 and January 2001 by the HIV prevention trial network (HPTN) as mentioned in Hu [100]. The primary endpoints of this trial were:

- HIV-1 infected rate of the infants and
- The proportion of those infants who were alive and free of HIV at 18 months of age.

The main purpose of the study was to compare the efficacy of two treatment regimens:

- Nevirapine (200 mg at labor onset and 2 mg/Kg for babies within 72 hours of birth) and
- Zidovudine (600 mg orally at labor onset then 300 mg every 3 hours until delivery and 4 mg/Kg orally twice daily for babies for 7 days).

One interesting question to be answered in this randomized trial is the evaluation of the capacities of two baseline biomarkers, the maternity HIV-1 RNA level and *CD4* for identifying who would be infected by their mothers after birth at various points in time or at various time intervals. There are various factors that may affect the biomarker distribution and performance in predicting a disease event. These factors are those covariates that need to be adjusted in time dependent ROC models. It is necessary to adjust for treatment regimens (Nevirapine and Zidovudine) when constructing the time dependent ROC curve of the biomarkers as it has an important impact on the prognostic capacity of HIV-1 RNA level and CD4.

---

**3.1. Extensions of sensitivity and specificity**

---

It has been previously mentioned that ROC curves are commonly used in the analysis of diagnostic test results  $Y$  for a binary disease outcome  $D$ . However in practice many disease outcomes depend on time  $t$ , which we denote by  $D(t)$ , and hence ROC curves that vary as a function of time may be more appropriate and various definitions and estimators have been proposed. Subjects are initially non-diseased but can succumb to disease during the course of the study [67]. A common example of a time-dependent variable is vital status, where  $D(t) = 1$  if a patient has died prior to time  $t$  and zero otherwise. It is clear that this type of data is most appropriately handled using time to event or survival analysis.

Let  $T_i$  be the survival time for subject  $i$  and assume that we only observe the minimum of  $T_i$  and  $C_i$ , where  $C_i$  represents an independent censoring time. Here survival times is interpreted to mean the time until an individual experiences an event of interest and  $C_i \leq T_i$  if the time is censored. Define the follow-up time as  $X_i = \min(T_i, C_i)$ , and let  $\Delta_i = I(T_i, C_i)$  denote the censoring indicator. The advantage of survival analysis approaches as opposed to a snapshot cross-sectional analysis of the binary outcome at a given time is that the time accrual until the event is taken into account. We then show that a certain choice of time dependent true positive rate (TPR(t)) and false positive rate (FPR(t)) definitions leads to time dependent ROC curves and time dependent AUC summaries. As we mentioned before with survival data we need to take the time into account since the accuracy may be higher when the markers are measured closer to the onset of disease.

To extend the notion of diagnostic accuracy to incorporate the time domain, the outcome is the time elapsed until an event takes place. This can be viewed as a binary outcome of function

of time. Equations (2.1) and (2.2) are now replaced by:

$$\text{sensitivity}(c, t) = P(Y > c | D(t) = 1) \quad \text{and} \quad (3.1)$$

$$\text{specificity}(c, t) = P(Y \leq c | D(t) = 0), \quad (3.2)$$

which signify that the sensitivity and specificity are functions of time in the context of time to event data. Using Equations (3.1) and (3.2), we can estimate the sensitivity and specificity for each  $c$  and plot these estimates to get the ROC curve at a specific time point  $t$ . These estimates can be obtained using the following relations, which follow from definitions of conditional probability as well as the application of Bayes' Theorem:

$$\begin{aligned} P(Y > c | D(t) = 1) &= \frac{1 - S(t|Y > c)P(Y > c)}{1 - S(t)} \quad \text{and} \\ P(Y \leq c | D(t) = 0) &= \frac{S(t|Y \leq c)P(Y \leq c)}{S(t)}, \end{aligned}$$

where  $S(t)$  denotes the survival function, i.e.  $S(t) = P(T > t)$  and  $S(t|Y > c)$  is the conditional survival function for the subset defined by  $Y > c$ .

The definition of the time dependent ROC curves follows from definitions of the usual ROC curves and relies on first defining time dependent sensitivity and specificity. Then simple plots of TPR vs FPR for different values of the threshold  $c$  will yield the ROC at time  $t$ . The time dependent AUC at time  $t$  is then defined as the area under this curve,

$$AUC(t) = \int_{-\infty}^{\infty} TPR(c, t) \left| \frac{\partial FPR(c, t)}{\partial c} \right| dc. \quad (3.3)$$

There are several definitions of cases and controls in the survival outcome setting. It is necessary to mention that the definitions of sensitivity and specificity are given in terms of the actual survival time  $T_i$ . In addition censoring needs to be addressed for valid estimation. A certain choice of time dependent true positive and false positive definitions leads to time dependent ROC curves and time dependent AUC summaries. We remark that time dependent

AUC summary is directly related to concordance summary for survival data. Concordance probability measures how often predictions and outcomes are concordant. The probability of concordance is defined as the number of concordant pairs plus the number of tied pairs divided by the number of all informative pairs [55]. For a binary outcome, the area under the empirical ROC curve is equivalent to the concordance probability, which is defined on a pair of subjects where one of the pair has the outcome and the other does not. The probability that the subject with the outcome has a greater marker value than the subject without outcome is called the concordance probability. Define

$$\psi(Y_i, Y_j) = \begin{cases} 1 & \text{if } Y_i > Y_j, \\ 0.5 & \text{if } Y_i = Y_j, \\ 0 & \text{if } Y_i < Y_j. \end{cases}$$

Hence  $\psi$  indicates which member of the pair has the higher value, with ties indicated by 0.5.

The concordance probability can be written as follows:

$$P(\text{concordance}) = \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m \psi(Y_i, Y_j),$$

where  $n$  and  $m$  are the numbers of patients with and without outcome respectively. This summation represents the number of pairs that have  $Y_i > Y_j$ , so the entire expression is the fraction of patient pairs where the one with the higher marker value had the outcome.

The idea of concordance can be extended to time to event setting. Let  $T_1$  and  $T_2$  be the event times in a given pair of patients with marker values  $Y_1$  and  $Y_2$ . The concordance between a marker  $Y$  and the time to event outcome  $T$  is defined as  $CP(Y, T) = P(T_1 > T_2 | Y_1 > Y_2)$ .

---

### **3.2. Time dependent true positive rate ( $TPR(t)$ ) and false positive rate ( $FPR(t)$ )**

---

This section discusses different definitions for time dependent true positive and false positive rates [67].

### 3.2.1 Time dependent true positive rate

True positive rate ( $TPR$ ) or sensitivity in classic settings with binary outcomes is defined to be the probability that a test result will be positive when the disease is present in the individual. However in practical settings the outcome may be failure time, where in this case, there are various definitions of  $TPR$ . In the following sections, we list two definitions of time dependent  $TPR$ .

#### Incident true positive rate $TPR^I$

The incident  $TPR$  for a biomarker value  $Y$  at event time  $t$  for any threshold  $c$ , denoted by  $TPR^I(c, t)$  is given by:

$$TPR^I(c, t) = P(Y > c | T = t). \quad (3.4)$$

Using this definition the cases are stratified according to the time at which events occur and in this case we are more interested in the disease incidence at a fixed time. There are many advantages of this definition of  $TPR^I$ , given it is based on diseased cases occurring at a given time. This feature is helpful when the total sample size is small. In addition this definition does not contain redundant information on disease cases.

#### Cumulative true positive rate $TPR^C$

The cumulative  $TPR$  for a biomarker  $Y$  at event time  $t$  for any threshold  $c$ , denoted by  $TPR^C(c, t)$  is given by:

$$TPR^C(c, t) = P(Y > c | T \leq t). \quad (3.5)$$

$TPR^C$  evaluates the sensitivity of the biomarkers for detecting events occurring throughout the follow up time up to  $t$ . Using this definition we are more interested in predicting the disease prevalence of the study at a given time. Thus the definition of  $TPR^C$  is useful when

disease prevalence is of interest. The disease cases are calculated based on cumulative time interval, which results in a large number of disease cases that are used for estimating the sensitivity. However ( $TPR^C$ ) contains redundant information, unlike  $TPR^I$ . In addition, it cannot distinguish the sensitivity for early events from that for late events.

### 3.2.2 Time dependent false positive rate

False positive rate ( $FPR$ ) with binary outcomes is defined to be the probability that a test result will be positive when the disease is absent in the individual. Time dependent FPR are of various types according to the definition of the controls. In the following sections we list two definitions of time dependent  $FPR$ .

#### Static false positive rate $FPR^S$

The static time dependent false positive rate denoted by  $FPR^S$  for a biomarker value  $Y$  and any threshold  $c$ , is defined to be:

$$FPR^S(c, t^*) = P(Y > c | T > t^*). \quad (3.6)$$

Using this definition controls are subjects who are event free through a fixed follow up time  $(0, t^*)$ , where  $t^*$  is a fixed point in time. In the definition of  $FPR^S$  controls are static over time. Thus the time defining the controls differs from that defining those in the corresponding  $TPR$ , no matter what type of  $TPR$  is used.

#### Dynamic false positive rate $FPR^D$

Dynamic false positive rate for a biomarker value  $Y$  and any threshold  $c$ , denoted by  $FPR^D$  is given by:

$$FPR^D(c, t) = P(Y > c | T > t). \quad (3.7)$$



Using in this definition, the time defining the event is dynamic, thus the controls at time  $t$  are defined to be patients who were still event free at time  $t$ . The advantage of  $FPR^D$  is that it is based on the time defining the corresponding  $TPR$  in the ROC curve.

In summary we stress that a case  $i$  is said to be incident if  $T_i = t$  and cumulative if  $T_i \leq t$  for the two definitions of cases. It is also important to distinguish whether controls are static defined as subjects with  $T_i > t^*$  for fixed value of  $t^*$ , or whether controls are dynamic defined for time  $t$  as those subjects with  $T_i > t$ .

---

### 3.3. Combinations of time dependent $TPR$ and $FPR$

---

Since the time dependent ROC curve is a compound function of  $TPR$  and  $FPR$ , a combination of various types of the two rates need to be selected according to the purpose of study. Table 3.1 (see [100]) lists some combinations of  $TPR$  and  $FPR$  for constructing certain types of time dependent ROC curve.

Table 3.1: Combinations of time dependent TPR and FPR

TPR (Cases)	FPR (Controls)	Examples from literature
Cumulative	Dynamic	Etzioni et al. (1999) Heagerty et al. (2000) Zheng and Heagerty (2004) Song and Zhou (2008)
Incidence	Dynamic	Zheng and Heagerty (2004) Heagerty and Zheng (2005)
Incidence	Static	Cai et al. (2006)

After presenting definitions for time dependent sensitivity and specificity, ROC curves and AUC summaries can be computed and interpreted. We will present formulae for most commonly used ones.

### 3.3.1 Incident-Static combination

Using the following definitions

$$\begin{aligned} \text{Sensitivity}^I(c, t) &= P(Y_i > c | T_i = t) \quad \text{and} \\ \text{Specificity}^S(c, t^*) &= P(Y_i \leq c | T_i > t^*). \end{aligned}$$

each subject does not change disease status and is treated as either a case or a control. Cases are stratified according to the time at which the event occurs (incident) and controls are defined as those subjects who are event free through a fixed follow-up period  $(0, t^*)$  static. These definitions facilitate the use of standard regression approach for characterising sensitivity and specificity as the event time  $T_i$  can be used as a covariate.

The group of static controls mimics the group of individuals who never develop the disease, meaning patients with preclinical diseases are eliminated from the control group as far as possible if  $t$  is large enough. This can be viewed as the ideal control group in some situations. The cumulative  $TPR$  can be computed from the incident  $TPR$  when the distribution of the event time is known. Consider the incident TPR and static FPR as defined in Equations (3.4) and (3.6). Applying Bayes' Theorem, they can further be rewritten:

$$TPR^I(c, t) = \frac{\int_c^\infty f(t|y)g(y)dy}{\int_{-\infty}^\infty f(t|y)g(y)dy}$$

and

$$FPR^S(c, t^*) = \frac{\int_c^\infty P(T > t^* | Y = y)g(y)dy}{\int_{-\infty}^\infty P(T > t^* | Y = y)g(y)dy},$$

where  $g(y)$  is the probability density function of  $Y$  and  $f(t|y) = \partial P(T \leq t|Y = y)/\partial t$  is conditional density function of  $T$  given  $Y = y$ .

Heagerty and Zheng [67] estimated the  $TPR^I(c, t)$  using the Cox model of the form  $\lambda(t, Y) = \lambda_0(t) \exp(\beta(t)Y)$ , where  $\lambda(t, Y)$  stands for the conditional hazard rate of  $T$  given  $Y$  while  $\lambda_0$  is the unspecified base line hazard rate. Let the notation  $R(t)$  denote the risk set at time  $t$ . The authors mentioned that the distribution for the random variable  $Y \times \exp(\beta Y)$  for subjects in the risk set at time  $t$  is equal to the conditional distribution of  $Y$  given  $T = t$ . Setting  $R(t) = i : X_i \geq t$ , this leads to

$$TPR^I(c, t) = \frac{\sum_{i \in R(t)} I(Y_i > c) \exp \beta(t) Y_i}{\sum_{i \in R(t)} \exp \beta(t) Y_i}.$$

As for estimation of  $FPR^S(c, t^*)$ , they also proposed

$$FPR^S(c, t^*) = \frac{1}{n_{t^*}} \sum_{i \in S_t} I(Y_i > c),$$

where  $S_t = i : X_i > t^*$  is the control set and  $n_{t^*}$  is the cardinality of  $S_{t^*}$ .

### 3.3.2 Incident-Dynamic combination

We recall that

$$Sensitivity^I(c, t) = P(Y_i > c | T_i = t) \quad \text{and} \quad (3.8)$$

$$Specificity^D(c, t) = P(Y_i \leq c | T_i > t). \quad (3.9)$$

Using the approach of incident-dynamic combination, a subject can play the role of a control for an early time  $T_i > t$ , but then play the role of a case when  $T_i = t$ . Sensitivity is a measure of the expected fraction of subjects with a marker greater than the threshold  $c$  among the subpopulation of individuals who truly have the event at time  $t$ , while specificity measures the

fraction of subjects with a marker less than or equal to  $c$  among those who survive or do not experience the event beyond time  $t$ .

Incident-Dynamic ROC curves are defined following Equation (3.8) as the function  $ROC_t^{I/D}(\rho)$ , where  $\rho$  denotes the corresponding incident true positive rate. Let  $c^\rho$  be defined as the threshold that yields a false positive rate of  $\rho : P(Y_i > c^\rho | T_i > t) = 1 - specificity^D(c^\rho, t) = \rho$ . The true incident-dynamic positive rate,  $ROC_t^{I/D}(\rho)$  is the sensitivity that is obtained using this threshold or  $ROC_t^{I/D}(\rho) = sensitivity^I(c^\rho, t) = P(Y_i > c^\rho | T_i = t)$ . Using the true and false positive rate functions allows the ROC curve to be written as:

$$ROC_t^{I/D}(\rho) = TPR_t^I[FPR_t^D]^{-1}(\rho),$$

for  $\rho \in [0, 1]$ ,

The area under the  $I/D$  ROC curve for time  $t$  is

$$AUC(t) = \int_0^1 ROC_t^{I/D}(\rho) d\rho.$$

So the ROC curve is simply the plot of  $TPR(c, t) = [P(Y > c | D(t) = 1)]$  and  $FPR(c, t) = [P(Y > c | D(t) = 0)]$ . The area under the  $I/D$  ROC curve for time  $t$  denoted by  $AUC_t^{I/D}$  is then defined:

$$AUC^{I/D}(t) = \int_0^1 ROC_t^{I/D}(\rho) d\rho. \tag{3.10}$$

### 3.3.3 Cumulative-Dynamic AUC $AUC^{C/D}(t)$

For a baseline marker value, Heagerty and et al. [66] proposed versions of time dependent sensitivity and specificity under the cumulative case definition as

$$\begin{aligned} Sensitivity^C(c, t) &= P(Y_i > c | T_i \leq t) \quad \text{and} \\ Specificity^D(c, t) &= P(Y_i \leq c | T_i > t). \end{aligned}$$

Using this approach, at any fixed time  $t$  the entire population is classified as either a case or a control on the basis of vital status at time  $t$ . Also, each individual plays the role of control for times  $T_i > t$ , but then contributes as a case for taken times  $T_i \leq t$ . Cumulative and Dynamic accuracy summaries are most appropriate when one is interested in discriminating between subjects who experience an event of interest such as death prior to time  $t$  and those survive beyond  $t$ .

The setting of cumulative cases and dynamic controls may be regarded as most natural when specific evaluation times are of particular interest. It simply corresponds to defining cases at time  $t$  as subjects who experienced the event prior to time  $t$ , and controls at time  $t$  as subjects who were still event free at time  $t$ .

$AUC_t^{C/D}(t)$  is then obtained by using these definitions of  $TPR^C$  and  $FPR^D$ . However  $I(T \leq t)$  is not observed for all subjects due to presence of censoring before time  $t$ . To handle censoring, Baye's Theorem can be used to rewrite  $AUC^{C/D}(t)$  as a function of the conditional survival function  $P(T > t|Y = y)$ . There are other approaches called *Inverse Probability of Censoring Weighted* (IPCW) estimates. We first mention the method based on primary estimates of  $P(T > t|Y = y)$ , using Bayes' Theorem where

$$TPR^C(c, t) = \frac{\int_c^\infty P(T \leq t|Y = y)g(y)d(y)}{P(T \leq t)},$$

$$FPR^D(c, t) = \frac{\int_c^\infty P(T > t|Y = y)g(y)d(y)}{P(T > t)}.$$

From Equation (3.3), it follows that

$$AUC^{C,D}(t) = \int_{-\infty}^\infty \int_c^\infty \frac{P(T \leq t|Y = y)P(T > t|Y = c)}{P(T \leq t)P(T > t)} g(y)g(c)dydc.$$

Since  $P(T > t) = \int_{-\infty}^\infty P(T > t|Y = y)g(y)dy$ , we let  $\hat{S}_n(t|y)$  to be the estimator of the conditional survival function  $P(T > t|Y = y)$ . Then

$$AUC^{C/D}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \hat{S}_n(t|Y_j)[1 - \hat{S}_n(t|Y_i)]\mathbf{I}(Y_i > Y_j)}{\sum_{i=1}^n \sum_{j=1}^n \hat{S}_n(t|Y_j)[1 - \hat{S}_n(t|Y_i)]}.$$

Heagerty [66] suggested a conditional Kaplan-Meier estimator to derive estimates for  $\hat{S}_n(t|y)$ .

Also Hung et al. [69] suggested to use IPCW estimates, so we have:

$$TPR^C(c, t) = \frac{\sum_{i=1}^n \mathbf{I}(Y_i > c, T_i \leq t) \frac{\Delta_i}{n\hat{S}_C(T_i)}}{\sum_{i=1}^n \mathbf{I}(T_i \leq t) \frac{\Delta_i}{n\hat{S}_C(T_i)}}$$

and

$$FPR^D(c, t) = \frac{\sum_{i=1}^n \mathbf{I}(Y_i > c, T_i > t)}{\sum_{i=1}^n \mathbf{I}(T_i > t)},$$

where  $\hat{S}_C(\cdot)$  is Kaplan-Meier estimator of survival function of the censoring time  $C$ . Then the  $AUC^{C/D}(t)$  estimator is given by:

$$AUC^{C/D}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbf{I}(T_i \leq t) \mathbf{I}(T_i > t) \mathbf{I}(Y_i > Y_j) \frac{\Delta_i}{\hat{S}_C(T_i)} \hat{S}_C(t)}{n^2 \hat{S}(t) [1 - \hat{S}(t)]},$$

where  $\hat{S}(t)$  is Kaplan-Meier estimator of  $P(T > t)$ . To conclude this section we would like to mention that all time dependent ROC curves definitions can be used to evaluate and compare biomarkers in classifying subjects based on their survival times. The Incidence-Static ROC curve is useful in distinguishing subjects that fail at a given time from those failing after another time. The Incidence-Dynamic ROC curve is useful in distinguishing subjects that fail at a given time from those failing after that time. The Cumulative-Dynamic ROC curve is useful in distinguishing subjects that fail by a given time from those failing after another time.

# Missing data and imputation methods

Missing data are quite common in both designed clinical trials and observational research studies. Some methodologists have described missing data as one of the most important statistical and design problems in research. The problem of missing data is of a greater concern when decisions are to be made about the appropriateness of the care a patient should receive and also when there is interest in using a predictive model to discriminate subjects as likely to have a certain characteristic from those who do not.

Missing values can severely affect the results if there is dependence between the outcome and the missing data process, therefore dealing with missingness in the data becomes necessary. Despite the important nature of the problem, a large number of researchers routinely employ old standby techniques that have been criticized in the methodological literature. A simple and common strategy is to ignore cases with missing values, which means reducing the size of the original data set and can introduce substantial biases in the analysis and inference. Deletion methods are among the worst methods available for practical applications [113] and this can lead to severe bias if especially the missing data are not occurring in a purely random manner.

Define the complete data set as  $Y = (y_{ij})$ , where  $y_{ij}$  denotes the  $j^{th}$  observation for individual  $i = 1, 2, \dots, N$  and  $j = 1, 2, \dots, n_i$ . Note that  $Y$  includes both observed and unobserved values. Thus the data  $Y$  can be partitioned as  $Y = (Y_{obs}, Y_{mis})$ , where  $Y_{obs}$  and  $Y_{mis}$  denote the observed and missing part of the complete dataset respectively. To the dataset  $Y$  we also associate a matrix  $I = I_{ij}$  and we refer to this matrix as the missing data indicator matrix. It is defined as:

$$I_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed,} \\ 0 & \text{if } y_{ij} \text{ is missing.} \end{cases}$$

A common modeling approach for missing data is to assume the missing data mechanism is characterised by the conditional distribution of  $I$  given  $Y$ , that is  $f(I|Y, \phi)$ , where  $\phi$  denotes missing data parameters. The joint probability distribution of the response variables and the missing data indicator variables can be expressed as

$$f(Y, R|\theta, \psi) = f(Y|\theta)f(R|\psi, Y), \quad (4.1)$$

where  $f(Y|\theta)$  and  $f(R|\psi, Y)$  denote the marginal distribution of the response variable and the conditional distribution of missing data, conditional on the response variable, respectively. The probability model 4.1 has two sets of parameters  $\theta$  and  $\psi$  representing the parameters of interest and the missing data parameters, respectively. In model (4.1), the correct inferences on  $\theta$  in general need to be conducted.

It is important to have a clear understanding of the so-called missing data mechanisms. Rubin et al. [113, 115] introduced three missing data mechanisms. These mechanisms describe the relationships between measured variables and the probability of missing data. While these terms have a precise probabilistic and mathematical meaning, there are different reasons for why the data were missed.



---

**4.1. Missing reasons**

---

According to [115] there are three missingness mechanisms. In this section we give a conceptual description of each mechanism and for more details on these mechanisms, we refer the readers to (Allison [5], Enders [36], Little and Rubin [81], Rubin [113], Schafer and Graham [118]).

**4.1.1 Missing completely at random (MCAR)**

The first mechanism is called **missing completely at random** (MCAR) and it happens when the probability of missing observations is unrelated to the value of that observation or to the value of any other variable. That is  $f(I|(Y_{obs}, Y_{mis}), \phi) = f(I|\phi)$  for all  $Y$  and  $\phi$ .

There are many and varied reasons for the data to be missed completely at random (MCAR). It can happen, for example, as a result of equipment malfunction, inclement weather, illness incapacitating subjects or testers; or incorrectly entered data. When we say data are missing completely at random, we mean that the probability that an observation  $X_i$  being missing is unrelated to the value of  $X_i$  or to the value of any other variables. Thus data on family income would not be considered MCAR if people with low incomes were less likely to report their family income than people with higher incomes. Similarly, for example, if in the USA Whites were more likely to omit reporting income than African Americans, we again would not have data that were MCAR because missingness would be correlated with the factor of ethnicity. However if a participant's data were missing because s/he was stopped for a traffic violation and missed the data collection session, his/her data would presumably be missing completely at random. This is supported by the fact that being stopped due to traffic violation can occur to any participant regardless of the value his/her outcome.

An interesting feature of data that are MCAR is that the analysis remains unbiased even if complete cases only are used. We may lose power for our design, but the estimated parameters are not biased by the absence of data. However the key concluding comment here is that practically the MCAR assumption is hard to justify.

### 4.1.2 Missing at random (MAR)

The second mechanism of missing data is when data are missing at random (MAR). This happens when the probability of a missing observation depends only on available information. Thus the MAR mechanism can be expressed as:  $f(I|(Y_{obs}, Y_{mis}), \phi) = f(I|Y_{obs}, \phi)$  for all  $Y_{mis}$  and  $\phi$ . The MAR mechanism requires a less stringent assumption about the reason for missing data. This terminology is often confusing because of the use of the word random. The MAR mechanism is in fact is not random at all and it describes systematic missingness where the propensity for missing data is related to other measured variables in the analysis model, but not to the underlying values of the incomplete variables [113]. Sometimes we refer to MAR as *ignorable missingness*. MCAR missingness also falls under the ignorable missingness. Cases of missing not at random (MNAR), to be introduced next, could be labeled as cases of nonignorable missingness.

### 4.1.3 Missing not at random (MNAR)

Data that are not MCAR or MAR are classified as *Missing Not at Random* (MNAR). As an example if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data is missing not at random. Clearly the mean mental status score for the available data will not be an unbiased estimate

of the mean that we would have obtained with complete data. The same thing happens when people with low income are less likely to report their income on a data collection form. In this case the probability that income is not reported depends on the unobserved income value itself.

The data are classified as Missing Not at Random (*MNAR*) when the presence of missing data depends on variable values, which are themselves subject to missingness. That is  $f(I|(Y_{obs}, Y_{mis}), \phi) \neq f(I|Y_{obs}, \phi)$ . To obtain an unbiased estimate of parameters we have to model the missingness itself. In other words we need to formulate and estimate a model that accounts for the missing data. That model could then be incorporated into a more complex model for estimating missing values (see [35] for an example).

#### **4.1.4 Ignorable and non-ignorable missingness**

Difficulties appear when we have data that are MNAR. We say that the mechanism controlling missing data is non-ignorable. That means we cannot sensibly solve a model unless we are also able to write a model that controls missingness. Modeling the missingness is not an easy task and most discussions, including this one, do not discuss the treatment of data whose missingness is non-ignorable. On the other hand, if data are at most MAR, the mechanism for missingness is ignorable. Thus we can proceed without worrying about the model for missingness. The intention is to find better estimators of the parameters in our model, but we do not have to write a model that incorporates missingness. In the next section we introduce some strategies seeking an improvement in the estimation.

---

**4.2. Imputation strategies**

---

There are several ways to deal with missing data. One of them is to discard subjects with incomplete sequences and then analyse only the units with complete data [101]. Methods that use this approach are called *deletion methods*. These methods do not replace or impute missing values and do not make other adjustments to account for missing values.

The main advantages of deletion methods are their simplicity and that they can be applied easily with much of the available statistical software. Some of the deletion methods are good, but are applicable only under certain conditions [18]. Ideally for the analysis to be valid one strong condition or assumption is that the data need to be missing completely at random. These conditions do not generally hold, therefore McKnight et al. [94] proposed that deletion methods be avoided whenever is possible. Furthermore Little and Rubin [81] do not recommend any of the available deletion methods except if the amount of missing data is limited. The simplest deletion approach is the *complete case analysis* or *list-wise deletion analysis* in which the analysis uses only those subjects with completely recorded observations, that is complete observations. Some of the advantages of complete case analysis are:

- It is simple, in the sense that the method can be quite effective and may be satisfactorily used with small amounts of missing data. However, it is important to make sure that, even in such a situation, the deleted cases are not unduly influential [118].
- It is easy to carry out. It is used by default routines in most statistical software packages, but implementation details vary.

The primary disadvantages of this method - which clearly outweigh the advantages - are that:

- It can produce inefficient estimates, in the sense of loss of statistical power specifically when drawing inferences for sub-populations.
- When data are not MCAR, then the method can lead to seriously biased results. In other words, this method is valid only when data are MCAR [80]. We remark that even when MCAR holds, it can still be inefficient [118].
- If the units with missing values differ systematically from observed cases, this could bias the complete-case analysis.
- If many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a simple analysis. Thus McKnight et al. [94] state that one should give careful consideration before the use of this method regardless of its ease of use.
- It is easy to imagine situations where complete case analysis can be very misleading. Kenward et al. [74] and Wang-Clow et al. [149] presented examples where the complete case has led to misleading results.

Next we discuss further deletion methods that can be considered as a replacement for listwise deletion.

***Pairwise deletion*** is a well-known deletion method. Under this approach each element of the intercorrelation matrix is estimated using all available data. As an example if one participant reports his/her income and life satisfaction index, but not his/her age, s/he is included in the correlation of income and life satisfaction, but not in the correlations involving age. The problem with this approach is that the parameters of the model will be based on different sets of data, with different sample sizes and different standard errors.

It is known that if there are only few missing observations, it does not affect the use pairwise deletion. If there are many missing observations, this may lead to inappropriateness in the analysis.

Another simple approach is *available case analysis*, where different aspects of a problem are studied with different subsets of the data. For example, in the 2001 Social Indicators Survey carried out in New York, USA, all 1501 respondents stated their education level, but 16% refused to state their earnings. We could thus summarize the distribution of education levels of New Yorkers using all the respondents and the distribution of earnings using the 84% of respondents who answered that question. A major problem of this approach is that different analyses will be based on different subsets of the data and thus will not necessarily be consistent with each other. In addition, as with complete case analysis, if the nonrespondents differ systematically from the respondents, this will bias the available case summaries.

We now turn to discuss methods that generate possible values for the missing data. These alternative methods are called *imputation methods*, where one “fills-in” (imputes) the missing data to obtain a full data set. Then the resultant data are analysed by standard statistical methods without concern as if the new set represented the true and complete data set [80, 115]. This is the key idea behind commonly used procedures for imputation which include *simple* and *multiple* imputation [80]. Multiple imputation fills in more than one value for each missing item to allow for the appropriate evaluation of imputation uncertainty [80, 115]. In contrast to multiple imputation, simple imputation techniques substitute one value for every missing value in the data set [80, 81]. Simple imputation methods are valid under the ignorability assumption [5, 115, 118]. Simple imputation methods that were used in the current research are

- *Mean imputation*, where missing observations are replaced with the estimated mean of

the data set;

- *Hot Deck imputation*, where the missing data can be replaced with the observed data taken from matched data from the variables that contain non-missing values.

Simple imputation methods are general and flexible for handling missing data and can be implemented quickly in several statistical software packages.

We now discuss in details some of the simple imputation methods that have been used in this work.

#### **4.2.1 Mean imputation**

The single imputation method is a simple technique for handling missing data and consists of replacing any missing observation with a plausible value. The most common single imputation techniques are the overall mean imputation for continuous variables and the mode imputation for categorical variables. The mean imputation can be used either by using a conditional mean based on other variables in the data set or by using the unconditional mean of the variable of interest. The mean and mode summary statistics are used because they seem to provide reasonable point estimates. However it is important to mention some of the advantages and disadvantages of the mean imputation method.

##### **Advantages of mean imputation:**

- This technique is easy to implement for any type of variable.
- Once missing values are imputed and incorporated into the data set, multiple users can use the data with consistent results.

- If knowledge regarding the mechanism producing missing values is available, imputed values can often be improved to reflect this additional information.

**Disadvantages of mean imputation:**

- Rubin [115] showed that one imputed value cannot reflect sampling variability and marginal distributions and associations are distorted as there is no residual variance after the imputation.
- The tendency of this technique to reduce the overall variance and increase the significance of individual covariates within a regression model leads to type II modelling errors. This problem can be controlled only under the strong mostly unattainable MCAR assumption, where the variance estimation is consistent with the true variance adjusted by a correction factor [108].
- Imputed missing data do not represent additional uncertainty when the reason for non-response is unknown.
- All observed values are considered as actual observations.

**4.2.2 Hot deck imputation**

The second simple imputation strategy we discuss is the *nearest neighbor hot deck imputation* (also known as *distance function matching*). The term hot deck indicates that the information of responding units (donors) come from the same dataset as the recipients. Following this approach [7, 108] the missing values of one or more variables for the non-respondents are replaced by values from observed closest similar donors in the sample. This is a donor method



where the donor is selected by minimising a specified distance. This method involves defining a suitable distance measure, where the distance is a function of the auxiliary variables.

There are several reasons for the popularity of the hot deck method among survey practitioners. As with all imputation methods, the result is a rectangular data set that can be used by secondary data analysts employing simple complete-data methods. It avoids the issue of cross-user inconsistency that can occur when analysts use their own missing-data adjustments. The hot deck method does not rely on model fitting for the variable to be imputed and thus is potentially less sensitive to model miss specification error.

**Advantages of hot deck imputation:** [23]:

- Missing values are imputed with real observed values.
- Nearest neighbor is more efficient than other hot deck methods as it uses the information of the auxiliary variables.
- It makes no distributional assumptions, in other words it is a distribution free method.
- We can use standard analysis for the imputed dataset.

**Disadvantages of hot deck imputation:**

- Requires some programming to be implemented
- Requires complete information on auxiliary variables
- Estimated values depend on the selected auxiliary variables.
- Most implementations don't provide an uncertainty assessment.

The nearest neighbor hot deck imputation (NNI) method has some interesting features:

- It is a hot deck method in the sense that non-respondents are substituted by a value of the same variable from a respondent of the same pool; the imputed values are actually occurring values, not constructed values, and they may not be perfect substitutes, but are unlikely to be nonsensical values.
- It is more efficient than other hot deck methods in the sense that non-respondents are imputed by deterministic values, given the  $y$ -respondents and  $x$ -values [108]. It is important to keep in mind that the hot deck method makes implicit assumptions through the choice of the metric to match donors to recipients, and the variables included in this metric, so it is far from assumption free.
- Only plausible values can be imputed, since values come from observed responses in the donor pool.
- There may be a gain in efficiency relative to complete-case analysis, since information about the incomplete cases is being retained.
- There is also a reduction in non-response bias, to the extent that there is an association between the variables defining imputation classes and both the propensity to respond and the variable to be imputed.
- It makes use of auxiliary information and does not use an explicit model and hence it is expected to be more robust against model violations than methods based on explicit models, such as ratio imputation and regression imputation.
- The NNI method provides an asymptotically valid distribution.

Let  $x_i = (x_{i1}, \dots, x_{iq})$  be the values for subject  $i$  of  $q$  covariates that are used to create adjustment cells and let  $C(x_i)$  denote the cell in the cross classification in which subject  $i$  falls.

Then matching the recipients  $i$  to donors  $j$  in the same adjustment cell is the same as matching based on the metric:

$$d(i, j) = \begin{cases} 0 & \text{if } i \in C(x_i), \\ 1 & \text{if } j \notin C(x_i). \end{cases}$$

The other measure of potential closeness of potential donors to recipients can be defined to be the maximum deviation,  $d(i, j) = \max_k |x_{ik} - x_{jk}|$  where  $x_k$  have been suitably scaled to capture differences comparable (e.g. by using ranks and then standardizing). The *Mahalanobis distance* [108],

$$d(i, j) = (x_i - x_j)^T \widehat{\text{var}}(x_i)^{-1} (x_i - x_j),$$

where  $\widehat{\text{var}}(x_i)$  is an estimate of the covariance matrix of  $x_i$ , or the predictive mean,

$$d(i, j) = (\hat{Y}(x_i) - \hat{Y}(x_j))^2,$$

where  $\hat{Y}(x_i) = x_i^T \hat{\beta}$  is the predicted value of  $Y$  for non-respondent  $i$  from the regression of  $Y$  on  $x$  using only the respondents' data. One way to define the donor set for non-respondent  $i$  is as the set of respondents with  $(d(i, j) < \delta)$ , for a pre-specified maximum distance  $\delta$ . If the closest respondent to  $j$  is selected, the method is called nearest neighbor hot deck.

Nonetheless despite the availability of the single imputation techniques (e.g. mean and hot deck), they are not at all recommended when the rate of missing values and number of parameters are large. A major shortcoming in using them is that single imputation does not account for imputation error.

### 4.2.3 Multiple imputation

In *multiple imputation* (MI) each missing value is replaced with several imputed values that reflect the uncertainty of the imputation model. Multiple imputation is a method to handle

incomplete data in statistical inferences and was first proposed by [115]. Under MI,  $m \geq 2$  independent imputations are carried out for each missing value to create  $m$  complete data sets. Next each complete imputed data set is analysed separately using an appropriate standard analysis method and the results are finally combined to produce estimates and confidence intervals for parameter values. Multiple imputation operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR), which means that the probability that a value is missing depends only on observed values and not on unobserved values [118]. In other words, after controlling for all of the available data (i.e., the variables included in the imputation model) any remaining missingness is completely random [57]. MI procedures are very flexible and can be used in a broad range of settings. Because MI involves creating multiple predictions for each missing value, the analysis of multiple imputed data take into account the uncertainty in the imputations and yield accurate standard errors. On a simple level, if there is not much information in the observed data (used in the imputation model) regarding the missing values, the imputations have high variability, leading to high standard errors in the analyses. In contrast, if the observed data are highly predictive of the missing values the imputations will be more consistent across imputations, resulting in smaller, but still accurate standard errors [58]. The method of multiple imputation ensures high efficiency even for a small number of imputations. The efficiency of this method is given by

$$q = (1 + \frac{\gamma}{m})^{-1},$$

where  $\gamma$  is the fraction of missing information due to non-response. For  $m$  imputations we have  $m$  estimates  $\hat{\theta}_i$ ,  $i = 1, 2, \dots, m$  each with an estimated sampling variance. Then the overall MI estimator for the parameter of interest (which can be vector-valued) is simply given by the average of the  $m$  estimators obtained from each of the  $m$  complete data sets and this is

computed through the formula

$$\bar{\theta} = \frac{1}{m} \sum_i^m \hat{\theta}_i.$$

The standard error is obtained by taking into account of within imputation variance, as well as between imputations variance. These two variances are then combined together and the square root of the sum determines the standard error. Average sampling variance of  $m$  estimates result in *within imputation variance*  $W$  expressed as

$$W = \sum_i^m \frac{\text{var}(\hat{\theta}_i)}{m}.$$

The variability of estimates across  $m$  imputations namely *between imputation variance*  $B$  is given by

$$B = \sum_i^m \frac{(\hat{\theta}_i - \bar{\theta})^2}{(m - 1)}.$$

The *total imputation variance*  $T$  of  $\bar{\theta}$  is then given by

$$T = W + (1 + \frac{1}{m})B.$$

Multiple imputation has a number of advantages over the other missing data imputation approaches. Multiple imputation involves filling in the missing values multiple times, creating multiple complete datasets. Following [118], the missing values are imputed based on the observed values for a given individual and the relations observed in the data for other participants, assuming the observed variables are included in the imputation model. The MI inference assumes that the model used in analysing the multiple imputed data (the analysis model) is the same as the model used to impute missing values in MI (the imputation model). However, practically, the two models might not be the same [116]. The quality of the imputation model will influence the quality of the analysis model results, so it is important to carefully consider the design of the imputation model. Therefore, in order to obtain high quality imputations for a particular variable, the imputation model should include variables that are potentially

related to the imputed variable and variables that are potentially related to the missingness of the imputed variable [116]. Van Buuren et al. [143] recommended including the following covariates in the imputation model:

- Variables in the analysis model.
- Variables associated with missingness of the imputed variable.
- Variables correlated with the imputed variable.

However, one can include auxiliary variables which may or may not have missing values. Generally, including variables that do not have missing values is recommended in the imputation model. For more details of the imputation model, the reader may consult [117, 118, 143].

**Advantages of MI:**

- It is applicable to any type of variables.
- It represents missing data uncertainty.
- It takes into account the variability given by the multiple imputed data set with appropriate statistical inference.
- It yields robust estimates.
- The use of standard analysis in each imputed data set.

Imputation methods keep the full sample size, which can be advantageous for bias and precision; however, they can yield different kinds of bias. Whenever a single imputation strategy is used, the standard errors of estimates tend to be too low. The intuition here is that we have substantial uncertainty about the missing values, but by choosing a single imputation we

in essence pretend that we know the true value with certainty. Although MI is intuitively appealing, it still has some defects. **Disadvantages of MI:**

- Analysis of multiple data sets is time consuming and requires statistical expertise.
- MI can introduce bias over a complete case analysis if not carried out appropriately.

The MI is the most commonly used approach to deal with missing data. MI is generally preferred to Inverse-Probability Weighting (IPW) as it is more efficient. If the imputation model is correctly specified, MI should work well. Furthermore, we were interested in comparing between single and multiple imputation methods in order to estimate time-dependent AUC. We are planning to use the IPW in the future work.

### 4.2.4 Multiple imputation via chained equations

*Multiple Imputation via Chained Equations* (MICE) is a particular multiple imputation technique [111, 144]. The name chained equations refers to the fact that the Gibbs sampler can be easily implemented as a concatenation of univariate procedures to fill out the missing data. Implementing MICE when data are not MAR could result in biased estimates. Many of the initially developed multiple imputation procedures assumed a large joint model for all of the variables, such as a joint normal distribution. In large datasets, with hundreds of variables of different types, this is rarely appropriate. MICE is an alternative, flexible approach to these joint models. In fact, MICE approaches have been used in datasets with thousands of observations and hundreds of variables [65, 129]. In the MICE procedure, a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. Thus MICE falls under the general class of models called the fully conditional specification (FCS) of the joint distribution. This means that each variable can

be modeled according to its distribution for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression. Let  $X_1, X_2, \dots, X_k$  be a set of variables some or all having missing values, the MICE algorithm can be described as follows:

- If  $X_1$  has missing values, it will be regressed on other variables, the estimation is restricted to individuals with observed  $X_1$ . The missing values in  $X_1$  are then replaced by simulated draws from the posterior predictive distribution of  $X_1$ .
- The following variable with missing values is regressed on all other variables, thus the estimation is restricted to individuals with observed values for that variable and uses the imputed values of  $X_1$ .
- This process is repeated for all other variables for  $c$  cycles, which is suggested to be more than 10 for the convergence of the sampling distribution of imputed values. The entire process is repeated independently  $m$  times.

The number of cycles to be performed can be specified by the researcher. At the end of these cycles, the final imputations are retained, resulting in one imputed dataset. Generally, ten cycles are performed [112]. The idea is that by the end of the cycles, the distribution of the parameters governing the imputations (e.g., the coefficients in the regression models) should have converged in the sense of becoming stable. This will, for example, avoid dependence on the order in which the variables are imputed. In practice, researchers can check the convergence by, for example, comparing the regression models at subsequent cycles, as discussed in [65]. Different MICE software packages vary somewhat in the exact implementation of this algorithm but the general strategy is the same. To make the MICE approach more concrete, imagine a simple example where we have three variables in our dataset: age, income, and gender,



and all three have at least some missing values. The MAR assumption would imply that the probability of a particular variable being missing depends only on the observed values and that, for example, whether someone's income is missing does not depend on their (unobserved) income. In step 1 of the MICE process, each variable would first be imputed using, for example, mean imputation, temporarily setting any missing value equal to the mean observed value for that variable. Then in step 2 the imputed mean values of age would be set back to missing. In step 3, a linear regression of age predicted by income and gender would be run assuming all cases were observed [11]. In step 4, predictions of the missing age values would be obtained from that regression equation and imputed. At this point, age does not have any missingness. Steps 2 to 4 would then be repeated for the income variable. The originally missing values of income would be set back to missing and a linear regression of income predicted by age and gender would be run using all cases with income observed. Imputations (predictions) would be obtained from that regression equation for the missing income values. Then steps 2 to 4 would again be repeated for the variable gender. The originally missing values of gender would be set back to missing and a logistic regression of gender on age and income would be run using all cases with gender observed. Predictions from that logistic regression model would be used to impute the missing gender values. This entire process of iterating through the 3 variables would be repeated until convergence. The observed data and the final set of imputed values would then constitute one complete data set. The process is repeated again to yield the second complete data set and again until  $m$  complete data sets ready for analysis are created via this simulation and estimation algorithm.

**Advantages of MICE:**

- It is considered a flexible approach because it gives flexibility to the researcher having a multivariate structure on the data.

- It can handle variables of different types.
- It can handle arbitrary missing-data patterns.
- It can accommodate certain important characteristics of the observational data.

**Disadvantages of MICE:**

- If the imputation model has too many variables it may lead to multicollinearity problems.
- It also requires comprehensive computational skills.
- Implementing MICE when data are not MAR could result in biased estimates.

Although facilitating computation is very important, such a viewpoint ignores the imputer's assessments and information inaccessible to the users [96]. In [96] it was mentioned that "This view underlies the recent controversy over the validity of multiple-imputation inference when a procedure for analyzing multiply imputed data sets cannot be derived from (is "uncongenial" to) the model adopted for multiple imputation". The uncongeniality arises when the analyst and the imputer have access to different amounts and sources of information, and have different assessments (e.g., explicit model, implicit judgement) about both responses and non-responses. If the imputer's assessment is far from reality, Rubin [115] stated "all methods for handling non-response are in trouble". Based on such assessment, all statistical inferences need underlying key assumptions to hold at least approximately. If the imputer's model is reasonably accurate, the multiple imputation prevents the analyst from producing inferences with serious non-response biases.

An issue of uncongeniality is that it reveals a unique feature of multiple imputation inferences that has not been studied systematically and is therefore unfamiliar to some analysts [96]. For

an analyst, conducting multiple imputation inferences remove two major burdens of analysing incomplete data: the difficulty of modeling missing data mechanisms and the computational complications of incomplete data analyses. It is therefore recommended that the imputation model should be rich enough and include all the variables that are used in the analysis model including auxiliary variables if any. This condition was met in the current research because our imputation model used all the information that was used in the analysis model.

# **Combining multiple biomarkers in diagnostic testing for cross-sectional data**

---

## **5.1. Introduction**

---

In medicine a biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal, pathogenic processes or a pharmacologic response to therapeutic intervention [128]. More specifically, a biomarker indicates a change in expression or state of a protein that correlates with the risk or progression of a disease, or with the susceptibility of the disease to a given treatment. Biomarkers have gained immense scientific, clinical value and interest in the practice of medicine. For example clinical signs, symptoms, laboratory tests, gene expression technology and combinations of the afore-mentioned rely on the use of biomarkers.

Complex organ functions or general characteristic changes in biological structures can also serve as biomarkers. Although the term biomarker is relatively new, they have been used

in pre-clinical research and clinical diagnosis for a considerable length of time. Biomarkers measure the progress of disease and assist in the evaluation of the most effective therapeutic regimes for the disease. In medicinal biology, they play major roles and help in early diagnosis, disease prevention, drug target identification and evaluation as well as drug response. Before diagnosis, markers may be used for screening and risk assessment. During diagnosis, markers can determine the staging, grading, and selection of initial therapy. During treatment, they can be used to monitor therapy, select additional therapy, or monitor recurrent diseases. An ideal biomarker should be safe and easy to measure. If the biomarker is to be used as a diagnostic test, it should be sensitive and specific and have a high predictive value. In other words, most patients without the disease should have negative test results and vice-versa [72]. Biomarkers may be used alone or in combination to allow classification of an individual to a unique group with defined characteristics. Biomarkers are also used in fields like geology, astronomy and chemistry.

A critical challenge in clinical research is the combination of multiple biomarkers into models to improve disease or outcome predictive accuracy. In medical research data, scientists are evaluating a number of biomarkers simultaneously, which introduces an added complexity to the analysis. In addition to providing an improved understanding of factors associated with infection and disease development, a combination of relevant markers is important to the diagnosis and treatment of disease. In this chapter 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 do so. Combining biomarkers can be used to identify important disease features, diagnosis and prognosis. Therefore it is of interest to develop methods that can achieve this goal.

Correct diagnosis of disease is a real challenge and medical researchers invest considerable time and effort to the enhancement of accurate disease diagnosis. In the field of genomics or genetic analysis, genotyping methods are being advanced to enhance accurate detection of disease presence or disease stage generally. The ROC is a commonly used statistical tool for describing the discriminatory accuracy and performance of a diagnostic test. The prediction error can be used for model comparisons and evaluation, but it is not a meaningful indicator for disease discriminatory capacity and may not necessarily represent the targeted population. Thus we proposed the use of the AUC estimator as an evaluation method.

Features selection methods must be taken into account to improve the inferior univariate selection of features based on traditional inference tests. Although univariate models have some appealing strengths and are comparatively easy to fit. However, correlation cannot be modeled using univariate process thus multivariate models provide more comprehensive analysis. It is well-known that evaluating the model on the same data that was used to build it will cause an over-fitting problem, thus resampling methods should be used. There has been much recent work on developing methods for combining multiple biomarkers. Su and Liu [130] proposed linear combination of markers to maximise sensitivity over the entire specificity range. They also provided a solution of the best linear combination of markers in the sense that the AUC of this combination is maximised among all possible linear combinations. Pepe and Thompson [103] proposed a distribution free rank based approach for optimising the AUC. In [93] McIntosh and Pepe showed that the risk score defined as the probability of disease given data on multiple markers is the optimal function that maximises the ROC curve at every point. Etzioni et al. [43] proposed screening rules based on the consideration of logical combinations of biomarker measurements. Yuan and Ghosh [153] proposed a novel model combining algorithms for classifying biomarkers in studies.

In this chapter, multivariate stepwise logistic regression is used to select the biomarkers. Then the bootstrap leave one out cross-validation (LOOCV) technique is applied to evaluate the diagnostic performance of the combined markers using the ROC analysis. The process of variable selection is applied to each training set. In a simulation study, we developed a statistical model based on the multivariate normal distribution and used it to show that accounting for statistical correlation among the biomarkers is useful to help improving the predictive accuracy. The method is applied to a real data set collected to study the occurrence of TB-IRIS in patients from Cape Town, South Africa. The method is designed for the analysis of cross-sectional data.

---

## 5.2. Variable selection

---

Feature selection, also known as variable selection is the technique of selecting a subset of relevant features or variables for building models. Variable and feature selection have become the focus of much research when tens or hundreds of thousands of variables are available. Feature selection is the common first step when developing a class predictor based on microarray data [122]. In fact it is reasonable to assume that only some subsets of many of measured biomarkers contribute useful information for distinguishing the phenotype classes. By removing most redundant biomarkers or variables from the data, feature selection helps improve the performance of models. However, prior to variable selection it is important to ensure that distorting features such as identification of outliers are properly handled. This may include exclusion of outliers and transformation of variables appropriately. The objective of variable selection is to avoid over-fitting, improve model performance and provide faster and more effective predictors. Faraway [46] states that among several plausible explanations for a phenomenon, the simplest and smallest that fits the data is best. It is well known that unnecessary predictors will add noise to the estimation of other quantities that are of interest and degrees of freedom will be

---

unnecessarily wasted. Moreover, collinearity is caused by having too many variables trying to achieve the same goal. Finally, if the model is correctly identified, we can save time and cost by not measuring and including redundant 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 the  $t$ -test or Wilcoxon rank-sum test can be used to assess univariate statistical significance [122]. Then those variables considered statistically significant are to be identified for inclusion in the multivariate model.

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

### 5.2.1 Backward elimination

Backward elimination is the simplest of all variable selection procedures and can be easily implemented without special software. 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. Below are the steps for backward variable selection:

1. Start with all the predictors in the model.
2. Remove the predictor with highest  $p$ -value greater than the pre-specified one.
3. Refit the model and return to step 2.
4. Stop when all  $p$ -values for the remaining predictors are less than the pre-specified level



such as  $\alpha = 0.05$ .

Backward elimination does not perform well in the presence of multicollinearity and it cannot be used if there are more variables than observations,  $p > n$ . Additionally it may be computationally expensive if there are many variables. A classical alternative is forward selection.

### **5.2.2 Forward selection**

Forward selection reverses the backward selection and begins with an empty or null model. Variables are added sequentially to the model until a pre-specified stopping rule is satisfied. At a given step in 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. Below are the key steps under the forward selection procedure.

1. Start with no variables in the model.
2. For all predictors not in the model, check their  $p$ -value if they are added to the model and choose the one with lowest  $p$ -value less than a pre-specified threshold value such as 0.05.
3. Continue until no new predictors can be added.

### **5.2.3 Stepwise regression**

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 combination of backward elimination and 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. In other words, this addresses the situation where variables are added or removed early in the process and we want to change our mind about them later. At each stage a variable may be added or removed and there are several variations on exactly how this is done. It is important to realize that a stepwise approach is not guaranteed to lead to the best possible model. But it almost always leads to a good model.

Various model selection methods such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are available and commonly used criteria. The AIC is a measure of the goodness of fit of an estimated statistical model. The AIC is a method 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 AIC and BIC are respectively given by

$$AIC = -2 \log L + 2p$$

while

$$BIC = -2 \log L + p \log n,$$

where  $L$  is the likelihood of the data given the model parameters,  $p$  is the number of model parameters and  $n$  is the number of observations. Note that BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC. Both AIC and BIC can be used as selection criteria for all models both nested and non-nested, although for nested models the likelihood ratio test is preferred.

### 5.3. Resampling methods in the context of combining multiple biomarkers and estimation of the AUC

---

Diagnostic tests are important components in modern medical practice. Recall 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 with varying information about a condition or disease, it might be beneficial to combine them in order to obtain better diagnostic accuracy with a goal of maximising the AUC over all possible combinations (Fang et al. [44]).

As pointed out in Fang et al. [44], the procedure of combining multiple test results has been well studied. For example, Su and Liu [130] discussed the optimal linear combination under the multiple-normal assumption; Pepe and Thompson [103], Pepe, Cai, and Longton [106]; and Ma and Huang [84] discussed this procedure under the generalised linear model (GLM) assumption or formulation. Copas and Corbett [25] addressed the *over-fitting problem* (arguing that using the same data both to fit the prognostic score and to calculate its ROC tends to give an over optimistic estimate of the performance of the score) when combining tests through a logistic regression model. In this thesis we use the logistic regression model, which is often used to find a linear combination of covariates that best discriminates between two populations. The purpose of this section is to briefly discuss the resampling methods with application to estimating the AUC in the context of variable selection.

In recent years many emerging statistical analytical tools, such as resampling methods have been gaining attention among psychological and educational researchers. However, many researchers tend to embrace traditional statistical methods rather than experimenting with these new techniques, even though the data structure does not meet certain parametric assumptions.

Resampling techniques are rapidly entering mainstream data analysis; some statisticians believe that resampling procedures will soon overtake common traditional non-parametric procedures and may displace most parametric procedures as well.

Resampling procedures are statistical inference methods based on generating repeated samples drawn from the original sample. Compared to standard methods of statistical inference, these modern methods often are simpler and more accurate, require fewer assumptions. Resampling provides clear advantages when assumptions of traditional parametric tests are not met, as with small samples from non-normal distributions. Additionally, resampling can address questions that cannot be answered with traditional parametric or non-parametric methods, such as comparisons of means, medians or ratios. Thus, resampling also has the advantage of conceptual simplicity. Classical parametric tests compare observed statistics to theoretical sampling distributions. Resampling is a revolutionary methodology because it departs from theoretical distributions. Rather, the inference is based upon repeated sampling within the same sample. Indeed, the resampling method is tied to the Monte Carlo simulation, in which researchers “make up” data and draw conclusions based on many possible scenarios [83]. Monte Carlo simulations are widely used by statisticians to study the actions of different statistical procedures. In resampling one could explore all possible combinations, but such a strategy can be too time-consuming and computing-intensive.

### **5.3.1 Over-fitting**

In statistics, over-fitting occurs when a statistical model describes random error or noise instead of the underlying systematic relationship. An over-fitting problem can also be defined as fitting a statistical model with too many degrees of freedom in the modeling process. Thus over-fitting leads to users being too optimistic about the performance of the model. Over-fitting

unrealistically leads to a complex model making the training data set too noisy and too small, in addition it gives a very rich hypothesis space. The possibility of over-fitting exists because the criterion in the training data model may not be the same as the criterion used to judge the efficacy of a model. In particular, a model is typically trained by maximising its performance on some set of training data. However, its efficacy is determined not by its performance on the training data but by its ability to perform well on unseen data. It may occur that a model begins to memorize training data rather than learning to generalise to new data. In other words, validating a model using the same data used to develop it is no evidence of prediction accuracy for the data.

One very common way of selecting variables for a regression model is to start with a series of univariate models to study the relation between each variable and the response. Then one selects only those variables significant for entry into the subsequent multivariate regression analysis. However the process still leads to degrees of freedom being spent against the sample and leading to increased risk of over-fitting. Using univariate prescreening also creates other problems in the context of multivariable modeling. Variables in isolation may behave quite differently with respect to the response variable than when they are considered simultaneously with one or more other variables. The relation between a variable and an outcome may not appear to be important at all in the univariate case, but may become quite important after adjustment for other covariables and vice-versa.

In order to avoid over-fitting, it is necessary to use additional techniques such as cross-validation and bootstrapping. The basis model validation techniques is either to explicitly penalize overly complex models, or to test the model's ability to generalise by evaluating its performance on a set of data not used for training, which is assumed to approximate the typical unseen data that a model will encounter.

Bootstrap validation of models also has been shown to be superior to older techniques of model validation, such as splitting the data set into training and validating data sets. Enhancements to models - such as shrinkage techniques - allow us to understand the extent of over-optimism and generate an estimate of how well the model might fit in a new sample. In the following we discuss some of the above mentioned resampling techniques.

### 5.3.2 Cross-validation

Cross-validation is useful in dealing with the problem of over-fitting. Validation techniques are motivated by fundamental problems such as model selection and performance estimation. As we mentioned before, over-fitting is one aspect of the larger issue of what statisticians refer to as shrinkage. Cross-validation techniques are one way to address this over-fitting bias and it is a model evaluation method that is better than simply looking at the residuals. Residual evaluation does not indicate how well a model can make new predictions on cases it has not already handled. Cross-validation techniques tend to focus on not using the entire data set when building a model but rather on subdividing the data into training and validation or testing subsets. Some cases are removed before the data is modelled; these removed cases are often called the testing set. Once the model has been built using the cases - often called the training set - the cases which were removed - the testing set - can be used to test the performance of the model on the “unseen” data.

Recall that the prediction error ( $PE$ ) is a quantity that measures how well the model predicts 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 [39]. In regression models it is referred to as the expected squared difference between a future response and its prediction from the model that is  $PE = E(y - \hat{y})^2$ .

Cross-validation 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 is a strong standard tool for estimating prediction error and it is a specialized resampling procedure that is designed specifically for application in model validation problems. 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 only valid if the test set is not used in any way in the development of the model. Using the complete set of samples to select markers violates this assumption and invalidates cross-validation. With proper cross-validation, the model must be developed from scratch for each leave-k-out training set. This means that feature selection must be repeated for each leave-k-out training set. The objective of variable selection is to avoid over-fitting, improve model performance and provide faster and more effective predictors.

As we have indicated earlier; if many diagnostic tests are available and some of them are redundant, then we want to seek an optimal subset of diagnostic tests where the combined test has the largest AUC. Note here the term test has same meaning as a biomarker. 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 give the largest - or maximises - 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* [39].

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.

There are many types of cross-validation. We will discuss some of them below.

### 5.3.3 $K$ -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 to 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  $k = 1, 2, \dots, K$  data subsets 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  $K$  is 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.



#### 5.3.4 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.

Cross-validation is a method for estimating the error rate given test data not used in the training stage. Regardless of what value you set for  $K$ -fold cross-validation, using a  $K$  of 10-20 gives better results than using a smaller number, but each number could result in a slightly different error estimate.

#### 5.3.5 Bootstrap method

In 1979 Efron [37] 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* [122] 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 misclassifications per prediction is used as an estimate of the misclassification rate [122]).

To understand bootstrap, suppose it were possible to draw repeated samples of the same size from the population of interest, a large number of times. Then it is possible to get a fairly good idea about the sampling distribution of a particular statistic from its estimated values arising from these repeated samples. The purpose of a sample study is to gather information cheaply in a timely fashion. The idea behind bootstrap is to resample with replacement from the sample data at hand and create a large number of bootstrap samples. The sample summary is then computed on each of the bootstrap samples.

The bootstrap method has been shown to be successful in many situations, therefore being accepted as an alternative to the asymptotic methods. In fact, it is better than some other asymptotic methods, such as the traditional normal approximation.

In this method estimates  $\theta_b^*$ ,  $b = 1, 2, \dots, B$  of the parameter of interest  $\theta$  are calculated from  $B$  pseudo samples. Then an estimate of the bootstrap variance of the parameter of interest is

calculated as:

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

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[39] stated that a large  $B$  would be 200 replicates and generally variance decreases as the number of replicate samples increases. For this reason we use 1000 bootstrap replicates in both simulation studies and application to real dataset.

### 5.3.6 Bootstrap standard errors and confidence intervals

Let  $\hat{F}$  be the empirical distribution. A bootstrap sample is defined to a random sample of size  $n$  drawn from  $\hat{F}$ , say  $x^* = (x_1^*, \dots, x_n^*)$ . The bootstrap data points  $(x_1^*, \dots, x_n^*)$  are a random sample of size  $n$  drawn with replacement from the population of  $n$  objects  $(x_1, \dots, x_n)$ . Thus some members of original data may not appear in the bootstrap sample and others may appear more than one times. A bootstrap replicate estimate  $\hat{\theta}^*$  is given by:

$$\hat{\theta}^* = s(x^*),$$

where  $s(x^*)$  is the estimating function  $S(\cdot)$  applied to  $x^*$  as was applied to  $x$ . The bootstrap estimates of standard error  $se_{\hat{F}}(\hat{\theta}^*)$  is the standard error of  $\hat{\theta}$  for data sets of size  $n$  randomly sampled from  $\hat{F}$ . Below is the algorithm for estimating the standard error of  $\hat{\theta} = s(x)$  from the observed data  $x$ .

- Select  $B$  independent bootstrap samples  $x^{*1}, \dots, x^{*B}$  each consisting of  $n$  data values drawn with replacement from  $x$ . The number of bootstrap replicates  $B$  will ordinarily be in range 25 – 200.

- Evaluate the bootstrap replication estimate corresponding to each bootstrap sample,

$$\hat{\theta}^*(b) = s(X^{*b}), b = 1, 2, \dots, B.$$

- Estimate the standard error  $se_F(\hat{\theta})$  by the sample standard deviation of the  $B$  replications:

$$s\hat{e}_B = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left( \hat{\theta}^*(b) - \hat{\theta}^*(.) \right)^2},$$

where

$$\hat{\theta}^*(.) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^*(b).$$

Confidence intervals for a given population parameter  $\theta$  are sample based whose range  $[\theta_1, \theta_2]$  given out for the unknown number  $\theta$ . The range possesses the property that  $\theta$  would lie within its bounds with a high (specified) probability. The latter is referred to as confidence level. Of course this probability is with respect to all possible samples, each sample giving rise to a confidence interval which thus depends on the chance mechanism involved in drawing the samples. The two mostly used confidence coefficients are 95% and 99%. We limit ourselves to the level 95% for our work here. Traditional confidence intervals rely on the knowledge of sampling distribution of  $\hat{\theta}$ , exact or asymptotic as  $n \rightarrow \infty$ .

There are two rules for the number of replicates:

- Even a small number of bootstrap replicates, for example  $B = 25$ , is usually informative and  $B = 50$  is often enough to give a good estimate of  $se_F(\hat{\theta})$ .
- Very seldom are more than  $B = 200$  replications needed for estimating a standard error, Much bigger values of  $B$  are required for bootstrap confidence intervals.

Standard errors are often used to assign approximate confidence intervals to a parameter  $\theta$  of interest. Given an estimate of  $\hat{\theta}$  and estimated standard error  $\hat{se}$  the usual 95% **confidence**

interval for  $\theta$  is

$$\hat{\theta} \pm 1.96\hat{se}.$$

The number 1.96 comes from standard normal table.

### 5.3.7 Bootstrap cross-validation

The method is proposed by Fu et al. [51] to handle small sample problems. The procedure generates  $B$  bootstrap samples of size  $n$  from the observed sample and then calculates a leave-one-out cross-validation estimate on each bootstrap sample. Averaging the  $B$  cross-validation estimates gives the bootstrap cross-validation estimate for the prediction error. The authors of [51] did not carefully address the issue of feature selection when the method is applied to high dimensional gene expression data. The bootstrap cross-validation method tends to underestimate the true prediction error.

### 5.3.8 Leave-one-out bootstrap

The leave-one-out bootstrap procedure [39] generates a total of  $B$  bootstrap samples of size  $n$ . Each observation is predicted repeatedly using the bootstrap samples in which the particular observation does not appear. In this way, the method avoids testing a prediction model on the observations used for constructing the model. The leave-one-out bootstrap is basically a smoothed version of the leave-one-out cross-validation. The leave-one-out bootstrap estimate has much smaller variability than the leave-one-out cross-validation estimate. A bootstrap sample of size  $n$  contains roughly  $0.632n$  distinct observations from the original sample. It is often inadequate to represent the distribution of the original data when the sample size  $n$  is small. Hence the leave-one-out bootstrap estimate tends to overestimate the true prediction

error.

For estimation of bootstrap leave one out 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, and then we calculated the mean for these 1000 AUCs in order to obtain a single AUC.

Bootstrapping was used to get the variance estimates for the cross-validated AUC as the bootstrap estimates have smaller variances, especially for small sample sizes.

In our case we used *leave-one-out cross-validation* (LOOCV) as it is nearly unbiased, easy to implement and to understand. 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 together with stepwise variable selection method based on AIC criteria. This gives prediction values between 0 and 1. The process is repeated until all the possible sets are selected. Finally the tested observations are then pooled together to estimate the AUC. Formally, the AUC is calculated with LOOCV as ([2])

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

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.

### 5.3.9 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.

### 5.3.10 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.

---

## 5.4. Logistic regression

---

Logistic regression is part of a broader family of generalised linear models (GLMs), where the conditional distribution of the response falls in the Bernoulli or Binomial distribution, and the parameters are set by the linear predictor. Ordinary least-squares regression is the case where response is Gaussian, with the mean equal to the linear predictor, and constant variance. Logistic regression is the case where the response is binomial, in R, any standard GLM can be fit using the (base) `glm` function. The major wrinkle is that, of course, one needs to specify the family of probability distributions to use, by the family option “binomial” defaults to 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. The most common regression model used to model binary outcomes such as disease outcomes is the logistic regression model given by

$$\log \frac{P(Y)}{1 - P(Y)} = \beta_0 + \sum_{i=1}^p \beta_i Y_i,$$

where  $Y_i, i = 1, 2, \dots, p$  are the  $p$  disease markers of interest measured from each subject or case and  $Y = (Y_1, \dots, Y_p)'$ ,  $\beta_i, i = 1, 2, \dots, p$  is the regression coefficient for  $Y_i$ , and  $\beta_0$  is the intercept, the value of the log odds when  $Y_i = 0$  for all  $i$  i.e. the null model with no additional information on the odds of the outcome. The model falls under a broader class of models called generalized linear models [92]. We can compute the probability  $P$  from the regression



equation also. So, if we know the regression equation, we could, theoretically, calculate the expected probability that  $D = 1$  for a given value of the vector  $Y = (Y_1, Y_2, \dots, Y_p)'$ . Logistic regression is one of the most commonly used tools for applied statistics.

**Advantages of logistic regression:**

- The technique is traditional and easy to understand and implement. It is very useful for understanding the influence of several independent variables on a single dichotomous outcome variable.
- The quantity  $\log p/(1 - p)$  plays an important role in the analysis of contingency tables (the log odds). Classification is a bit like having a contingency table with two columns (classes) and infinitely many rows (values of  $y$ ).
- It is closely related to the exponential family which arises in many contexts in statistical theory, thus there are lots of problems which can be turned into logistic regression.
- It often works surprisingly well as a classifier.
- The dependent and independent variables do not have to be normally distributed.
- It does not assume a linear relationship between the dependent and independent variables.
- It may handle nonlinear effects.
- There is no homogeneity of variance assumption.
- Normally distributed error terms are not assumed.

**Disadvantages of logistic regression:**

- Logistic regression cannot predict continuous outcomes.
- Logistic regression requires that each data point be independent of all other data points.

When using the logistic distribution, we need to make an algebraic conversion to arrive at our usual linear regression equation. The goal of logistic regression is a bit different, because we are predicting the likelihood that  $Y$  is equal to 1 rather than 0 given certain values of  $Y$ . That is, if  $Y$  and  $D$  have a positive linear relationship, the probability that a person will have a score of  $D = 1$  will increase as values of  $Y$  increase. So we are concerned about predicting probabilities rather than the scores of dependent variable. As mentioned before logistic regression predicts probabilities and since the modeling is based on a given distributional assumption, the Bernoulli model, we can fit or estimate the model using a likelihood approach. Since the probabilities of the two possible outcomes are either  $P$  if  $D = 1$  or  $1 - P$  if  $D = 0$ , then the likelihood is given by:

$$L(\beta_0, \beta) = \prod_{i=1}^n P(Y)^D (1 - P(Y))^{1-D}.$$

The log likelihood is then given by:

$$\begin{aligned} \ell(\beta_0, \beta) &= \sum_{i=1}^n D \log P(Y) + (1 - D) \log(1 - P(Y)) \\ &= \sum_{i=1}^n \log(1 - P(Y)) + \sum_{i=1}^n D \log \frac{P(Y)}{1 - P(Y)} \\ &= \sum_{i=1}^n \log(1 - P(Y)) + \sum_{i=1}^n D(\beta_0 + Y\beta) \\ &= \sum_{i=1}^n -\log(1 + \exp(\beta_0 + Y\beta)) + \sum_{i=1}^n D(\beta_0 + Y\beta). \end{aligned}$$

Iterative methods such the Newton-Raphson, Iterated (Re-)Weighted Least squares and the Fisher scoring can easily be used to estimate the parameters of the model including their asymptotic variances.

---

### 5.5. Linear discriminant analysis

---

Discriminant analysis is a statistical technique that allows one to understand the differences between two or more groups with respect to several variables simultaneously. In other words, the aim of discriminant analysis is to classify an observation, or several observations, into these known groups. In general, discriminant analysis is concerned with the development of a rule for allocating objects into one of some distinct groups. Then the constructed classification rule will be used to determine a group membership for some future objects.

In different papers (see for example [42, 103, 130]), linear combinations of markers that maximise the area under the receiver operating characteristic curve have been proposed. However, none of them can be applied in all possible scenarios.

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

$$AUC_{LD} = \Phi \left[ \sqrt{(\mu^D - \mu^{\bar{D}})' (\Sigma^D + \Sigma^{\bar{D}})^{-1} (\mu^D - \mu^{\bar{D}})} \right], \quad (5.1)$$

where  $\Phi$  denotes the standard normal cumulative distribution function [103].

---

### 5.6. Algorithm

---

In this section we supply an algorithm for the computations of the AUC.

- The first step is to evaluate the diagnostic performance of biomarkers: We estimate the

diagnostic accuracy of each biomarker by computing and plotting the ROC curve and estimating its area under the ROC curve using `roc` and `auc` functions in `pROC` package in R software. The area under an ROC curve captures the overall diagnostic accuracy of the test. In our proposed algorithm, a non-parametric estimate of the area using trapezoidal rule is used in this step for the classical ROC curve (binary response). AUC with high values indicate diagnostically informative biomarkers and low values suggesting a low discriminatory performance of the biomarkers. We also obtained the confidence interval for each biomarker AUC using bootstrapping. However, since the diagnostic performance of one biomarker may be correlated with that of others the single biomarker may not give a good AUC and thus the biomarker combination method was used. This leads to step two below.

- The second step is undertaken to optimise the set of biomarkers with high and independent diagnostic information content in a multivariate setting. There are critical issues that need to be considered in this step: choosing an appropriate statistical method for multivariate analysis, choosing the number of diagnostically informative biomarkers to be entered into the multivariate model and using an appropriate method to optimise the number of finally selected biomarkers. As the outcome variable is by definition dichotomous, the likely choices can be methods like logistic regression or probit regression. In our algorithm we use logistic regression (implemented using the `glm` function in `stats` R package) for binary response outcomes and linear discriminant function model which is an extension of the linear regression model and can also be used in place of logistic regression. Therefore the linear discriminant score is used. Stepwise regression using backward elimination procedure based on `stepAIC` function within the `MASS` package was used.
- Before applying the algorithm to derive a discriminatory rule and in order to avoid the

over-fitting problem associated with model selection, we split each original dataset into a training set and a validation set. This split is done as many times as the number of individuals, because we use leave one out cross-validation (LOOCV) as described in Subsection 5.3.9. We also report bootstrapping LOOCV estimator of AUC and its variance as in Subsection 5.3.10. We wrote our own code in R for estimating the bootstrap LOOCV AUC and its variance. Feature Selection was done from scratch for each training set.

---

### 5.7. Simulation studies

---

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. Our simulation is based on different assumptions of biomarkers correlations in order to understand the effect of different correlations on the AUC estimation.

We simulate datasets under the following group settings: Assume that there are  $K$  diagnostic tests (corresponding to  $K$  biomarkers)  $Y_1, Y_2, \dots, Y_K$ . In our case, we let  $K = 5$ , that is five biomarkers  $Y_1, Y_2, Y_3, Y_4$  and  $Y_5$ . Let the mean vector of the  $K$  biomarkers in diseased and non-diseased be denoted by  $\mu_k^D$  and  $\mu_k^{\overline{D}}$  respectively.

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

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

and

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

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 the index for the set  $\{1, 2, \dots, n\}$  (resp.  $\{1, 2, \dots, m\}$ ),
- $k$  is the index for a biomarker,
- $a_i^D$  (resp.  $a_j^{\overline{D}}$ ) is the subject specific random effect which is assumed to follow the normal distribution that is  $a_i^D \sim N(0, 0.5)$  (resp.  $a_j^{\overline{D}} \sim N(0, 0.5)$ ) and
- $\varepsilon_{ik}^D$  (resp.  $\varepsilon_{jk}^{\overline{D}}$ ) is the random error effect also assumed to follow the normal distribution  $\varepsilon_{ik}^D \sim N(0, 0.25)$  (resp.  $\varepsilon_{jk}^{\overline{D}} \sim N(0, 0.25)$ ).

We use small variances (0.25 and 0.5) in simulation since low variance means that, in general, samples will be close to the mean and hence to each other.

The outcome vectors  $y^D$  and  $y^{\overline{D}}$  are respectively generated from three multivariate normal distributions with means given by  $\mu_k^D = (0.5, 0.25, 0, 0, 0)$  and  $\mu_k^{\overline{D}} = (0, 0, 0, 0, 0)$  for three group settings defined by three different variance-covariance matrices. The three variance-covariance matrices are as follow:

- For the first setting (Model 1) we assume independence between all the biomarkers and consequently we will have variances in the main diagonal and zeros elsewhere.
- For the second setting (Model 2) we add dependence for the biomarkers  $Y_1$  and  $Y_2$ .
- For the third setting (Model 3) we assume the same dependence across all the biomarkers.

The resulting covariance structure is the exchangeable or compound symmetry.

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 individuals after applying logistic regression to predict the disease outcome for different biomarker combinations and applying re-sampling methods. 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. In the simulations and application we applied stepwise variable selection using the AIC criterion to select the biomarkers implemented using the stepAIC function from MASS R package.

An original R program was written to carry out this process for the simulated data. We calculated the AUC (from the auc function available in pROC package) after combining the biomarkers using bootstrap Cross-Validation which we denote by  $AUC_{bcv}$ . Computing coverage probability is complex in this setting, therefore we evaluated the performance of the fixed predictor model on a large simulated dataset of sample sizes set at 10 000.

Table 5.1 shows a summary of different types of quantities which were estimated from the analysis. These include first: The true AUC based on LDA ( $AUC_{TLD}$ ), the mean of Cross-Validated AUCs ( $AUC_{cv}$ ) across 1000 simulations, the mean of Bootstrap Cross-Validated AUCs ( $AUC_{bcv}$ ) across 1000 simulations, in columns 1, 2 and 3 respectively.

Columns 4, 5 and 6 include respectively the confidence interval (CI) of Cross-Validated AUC ( $AUC_{cv}$ ) across 1000 simulations, the confidence interval for  $AUC_{cv}$  based on standard errors obtained from the Hanley and McNiel [61] method, the confidence interval of  $AUC_{bcv}$  based on asymptotic normality using bootstrap standard errors.

Proportion of times lower confidence limits of  $AUC_{cv}$  and  $AUC_{bcv}$  excludes 0.5, are listed

in columns 7 and 8 respectively. In columns 9 and 10 we show respectively the coverage probabilities of  $AUC_{cv}$  and  $AUC_{bcv}$  that include the true AUC (TAUC).

Columns 11, 12, 13 and 14 display respectively the empirical standard errors for  $AUC_{bcv}$  ( $SE_{b,e}$ ), bootstrap standard errors ( $SE_b$ ), the empirical standard errors for  $AUC_{cv}$  ( $SE_{e,cv}$ ) and the standard errors for  $AUC_{cv}$  ( $SE_{cv}$ ) using the Hanley and McNeil Equation [61].

Finally the last three quantities reported are the prediction errors ( $PE$ ), true prediction errors ( $TPE$ ) and the true AUC (TAUC) obtained from a large dataset.

With a total sample size  $N = 200$ , bootstrap replicates  $B = 1000$  and number of simulations  $nsim = 1000$ , we used the proposed three variance-covariance matrices specified under Models 1, 2 and 3 respectively in Table 5.1 together with the mean vectors of diseased and non-diseased  $\mu^D$  and  $\mu^{\overline{D}}$  to perform our simulation. In the simulations 1000 AUC values were estimated.



Table 5.1: Mean of the AUC from the 1000 simulated samples for each scenario,  $B = 1000$  bootstrap replications are performed for computing the cross-validation and the bootstrap cross-validation for the AUCs

	True AUC Based on LDA	CV-AUC Mean across 1000 simulations	BS-CV AUC Mean across 1000 simulations	CI from Hanley & McNeil based 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 limit of CV-based CI bound excludes 0.5	Proportion of times lower limit 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 5.1 (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

From Table 5.1, we can see that:

- For Model 1, the  $AUC_{TLD}$  equals to 0.6772 while the TAUC equals to 0.662 which means that using a large dataset gives AUC values (TAUC) nearly close to true AUC from LDA ( $AUC_{TLD}$ ). The 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 CIs 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 limits of CIs for  $AUC_{cv}$  and  $AUC_{bcv}$  exclude 0.5 are 0.903 and 0.875 respectively. Clearly, both methods (cross-validation and bootstrap cross-validation) perform well given only 9.7% and 12.5% of the times do the lower limits respectively include the threshold of 0.5.

The coverage probabilities (proportion of times the CI's include the true AUC values) for  $AUC_{cv}$  and  $AUC_{bcv}$ , 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. This indicates that the bootstrap cross-validated AUC estimation method performs better than just cross-validated AUC estimation. We also found that the bootstrap method produces larger variances therefore yielding larger standard errors while using the Hanley and McNeil method gives smaller standard errors. Finally, both PE and TPE values are similar to each other.

- 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 0.639 and 0.655 respectively. This result indicates that values of  $AUC_{cv}$  and  $AUC_{bcv}$  based on a model with some correlation are close to each other and nearly unbiased since their values are close to the true AUC values.

Based on CIs for both  $AUC_{cv}$  and  $AUC_{bcv}$  we deduce that the two values of  $AUC_{cv}$  and  $AUC_{bcv}$  are statistically significant because they both exclude 0.5 the value under  $H_0$ . However the AUC values tend to be lower here than in Model 1. The results from Model 2 also show that the bootstrapping is better than just cross-validation for estimating the coverage probability and it gives AUC values close to true AUC.

- From Model 3 CIs 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 CIs for both  $AUC_{cv}$  and  $AUC_{bcv}$  that exclude 0.5 are 0.982 and 1 respectively. 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 affords confidence intervals that most of them include the true AUC.

From the above results we can see that LOOCV is nearly unbiased as a method of estimating the AUC for the three models. It appears that the bootstrap cross-validated AUC values are larger than the cross-validated AUC values. Most of the bootstrap cross-validated confidence

intervals contained the true values of AUC. We conclude by remarking that using Model 3 (with correlation) is preferable since it yielded the highest AUC values and smallest variance compared to other two models. Furthermore the bootstrap method gives larger variances compared to empirical variances.

---

### **5.8. Application to TB-IRIS**

---

The dataset that is used was collected in a study to investigate the occurrence of tuberculosis immune reconstitution inflammatory syndrome (TB-IRIS). According to a recent paper by Marais et al., [90], paradoxical TB-IRIS occurs in 8 – 43% of HIV-infected patients receiving TB treatment after starting antiretroviral therapy (ART) in South Africa. It was reported that TB-IRIS results from rapid restoration of *Mycobacterium tuberculosis* (*M. tuberculosis*)-specific immune responses. A prospective observational study targeting adults ( $\geq 18$  years) ART-naïve 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. For more details about these biomarkers, we refer the reader to [9, 27, 82, 88, 107, 142, 148]. There are not many published studies describing tuberculous meningitis TBM-IRIS. The authors of [90] investigated clinical and laboratory findings in ART-naïve HIV-infected patients who presented with TBM. They based their study on serial cerebrospinal fluid (CSF) samples in patients who did and did not develop TBM-IRIS.

Five biomarkers, namely Interleukin 6 (il6), interleukin 10 (il10), interleukin 12p40 (il12p40), interferon gamma (ifng), and tumor necrosis factor alpha (tnfa) were selected as candidate markers of TBM-IRIS. These biomarkers were measured in CSF at the time of TBM diagnosis.

The authors of [90] suggested an analysis model for evaluating the multivariate biomarkers model. They selected significant biomarkers using Wilcoxon rank sum tests for a logistic regression model and the final model was found by dropping the non-significant biomarkers. The authors of [90] used bootstrap cross-validation method to build the model and the permutation test to provide a cross-validated estimate of the AUC and confidence intervals. In this manuscript we used bootstrapped cross-validation to estimate AUCs and CIs and in addition to that we did not pre-specify the biomarkers but allowed the most informative biomarkers come out of the model.

First we considered AUC estimation individually for each biomarker in order to evaluate the performance of each biomarker in distinguishing between IRIS and Non-IRIS groups. However our main purpose was to calculate the 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.

Graph 5.1 plots the estimated disease risk versus the risk distribution. Risk estimates are based on a generalised linear binary model for disease risk as a function of the specified marker.

Table 5.2 contains the AUC values for each of the above biomarkers, their 95% confidence intervals (CI) and standard errors SE. The CI and SE were estimated using the bootstrap method with 1000 replicates.

As already stated a larger AUC value would suggest that the marker or 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 (for example from Table 5.2 the biomarker with the highest AUC value has the smallest SE). As can be seen in Table 5.2, the results show that

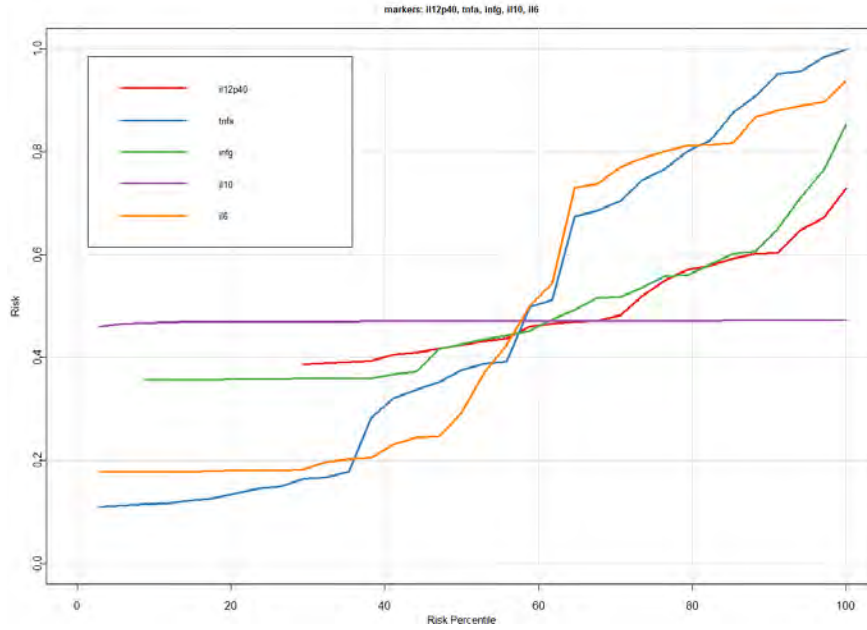


Figure 5.1: Predictiveness curves for TB-IRIS biomarkers

Table 5.2: AUC values for TB-IRIS 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>infg</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

fitting a logistic regressions for each biomarker, the smallest AUC values were for *il12p40* and *il10* equal to 0.67 and 0.66 respectively and the CIs both include 0.5. The bootstrap technique provided 95% confidence interval of (0.48, 0.84) and (0.45, 0.84) respectively and their respective standard errors equal to 0.09 and 0.1. Based on the null hypothesis of  $H_0 : AUC = 0.5$  and from the confidence intervals for the *il12p40* and *il10*, we conclude that *il12p40* and *il10* AUCs

are not statistically significant and therefore the two markers individually do not have a high discriminatory ability between the two groups of IRIS and non-IRIS patients. In addition their variances which are larger compared to other biomarkers is an indication of less precise tests. The CI for *infg* indicates that the AUC for this biomarker is statistically significant; the estimated AUC value suggests that this biomarker does have a fairly good ability to discriminate between two groups of IRIS and non-IRIS.

From Table 5.2 we conclude that the two biomarkers *tnfa* and *il6* have the best ability to distinguish between the two groups of IRIS and non-IRIS patients compared to other biomarkers as they have the highest statistically significant AUCs values and less variable estimates.

Similar to simulation studies, we calculated the AUC from combination of the five biomarkers using bootstrap cross-validation with repeated variable selection in each training set. We also calculated the variance for bootstrap cross-validated AUC the same way we did for the simulated data.

An R program similar to the one used in simulation studies, was written in order to apply the bootstrap cross-validation technique of AUC and variance estimation to the IRIS data. In our case we used 1000 bootstrap replicates.

Table 5.3 shows the results obtained from the application of the bootstrap cross-validation technique from the model that combines the biomarkers. The  $AUC_{bcv}$  is estimated as 0.956 with

Table 5.3: Bootstrap cross-validation of AUC for composite marker from TB-IRIS

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

the corresponding bootstrap cross-validated standard error of 0.036. The confidence interval was 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 yields a high AUC value and small variability. This shows that the combination of biomarkers may have a high ability to distinguish between diseased and non-diseases samples. 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 as  $PE = 0.059$ , indicating that the model performance is good. It is clear 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.

Figure 5.2 contains the ROC curves, AUCs and CIs for two of the biomarkers, namely *tnfa* with a high significant AUC of 0.87 (95%CI : 0.74, 0.96) and *il10* with a low non-significant AUC of 0.66 (95%CI : 0.45, 0.84).

Figure 5.3 shows empirical and smoothed ROC curves for *tnfa*. It also shows the AUC value corresponding to the best threshold.

We conclude this section by remarking that resampling methods (e.g., cross-validation and bootstrap) in terms of variable selection for the purpose of estimating the AUC for a model that combine biomarkers for both simulated and real datasets (TB-IRIS) gives deep insight in understanding the disease and provide a more accurate diagnosis of the disease.



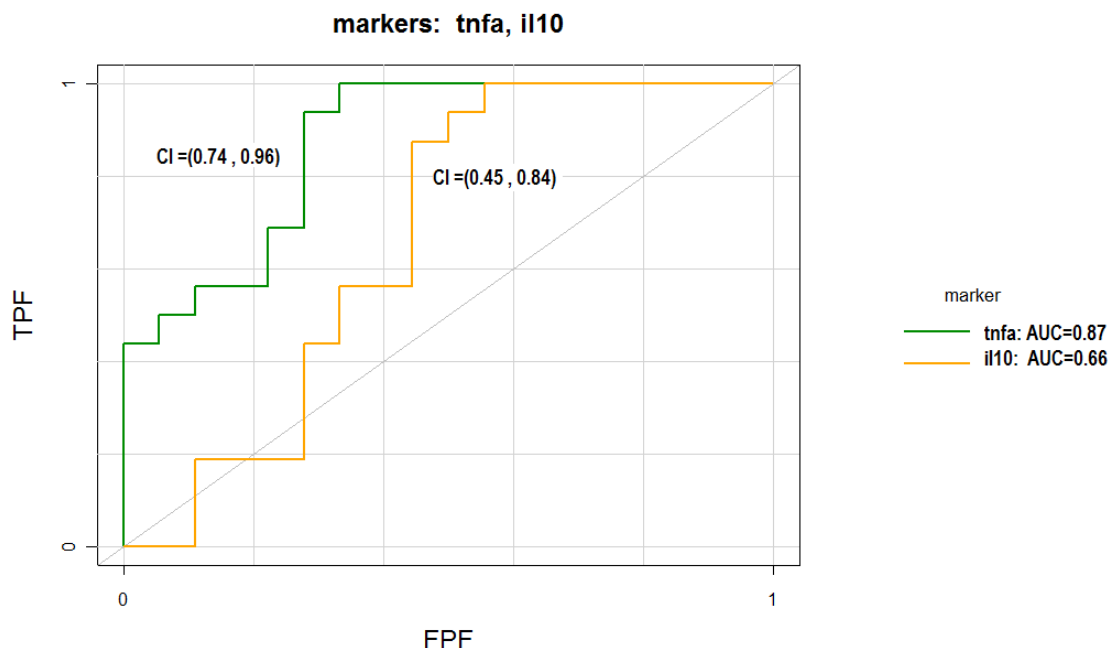


Figure 5.2: ROC curves for *tnfa* and *il10*

### 5.8.1 Conclusion

In this chapter, we used stepwise logistic regression combined with LOOCV bootstrapping in the estimation of the AUC. The AUC which is the area under the ROC is a popular summary index to evaluate the discriminatory accuracy of a diagnostic method. It can also be used to assess the ability of a prognostic factor to correctly distinguish patients who have an event such as a disease from those who do not. In the application to TB-IRIS data, for example, our estimated AUC value using the composite marker was 0.96. This means that the combination of biomarkers has high ability to predict TB-IRIS.

We studied the proposed methodology by simulating three correlation scenarios (no correlation, two biomarkers with some correlation, same correlation between all pairs of biomarkers or the exchangeable correlation structure).

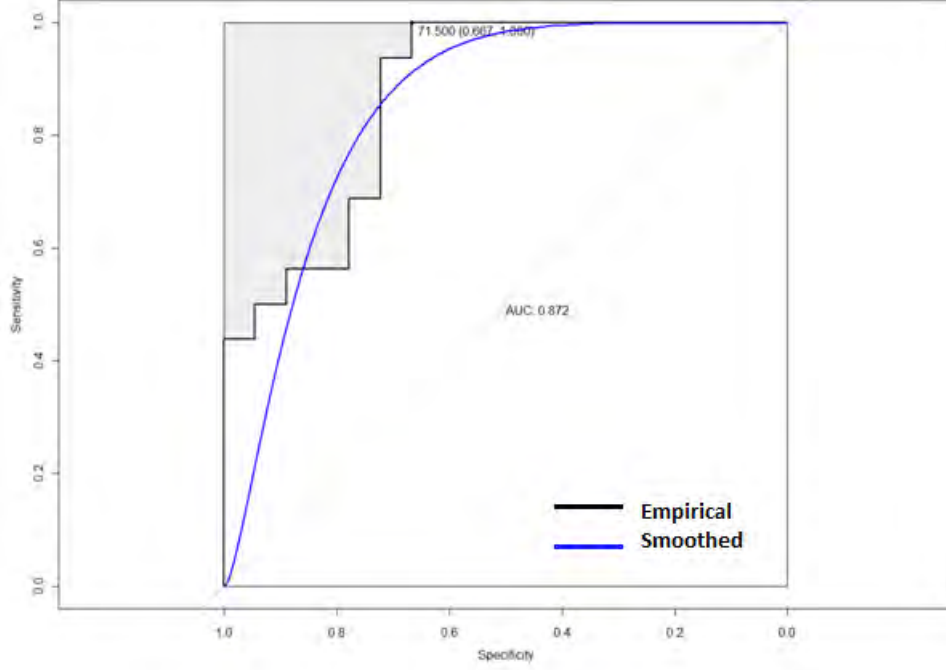


Figure 5.3: ROC curve for *tnfa*

We found that Model 3 gives higher AUC values compared to the other models. This shows that accounting for correlation between biomarkers gives a better predictive model than a model that ignores it. We also deduced that LOOCV gives nearly unbiased AUC estimates and using bootstrap LOOCV estimator gives high coverage probability. We note that the bootstrap method gives larger variances compared to empirical variances, and its coverage probability is at or above the nominal level. In addition AUC estimates based on bootstrap cross-validation are larger than those based on cross-validation alone. An interesting point noted is that our proposed method gives similar AUC estimates to those using either large independent dataset or LDA.

An application to IRIS dataset reveals that the bootstrap LOOCV for combining TB-IRIS biomarkers gave higher AUC value than using a single biomarker in the model. Both cross-

validation and bootstrap cross-validation yield AUC estimations closely to each other.

We would like to mention that in [90], the authors produced a cross-validated AUC of 0.91 which is less than our estimated cross-validated AUC (with or without bootstrapping of 0.956 and 0.967 respectively). In our opinion this is because in this earlier analysis the biomarkers were pre-specified. This agrees to what we mentioned before that variable selection should be done in each training set.

# **Predictive accuracy of multiple time dependent biomarkers with missing values in diagnostic testing**

---

## **6.1. Introduction**

---

Current available methods in analysing ROC curves are limited to complete data sets and classical ROC analysis. However in the development of prognostic models the presence of missing data is a frequently encountered problem and cannot be overlooked. Using the complete case analysis to deal with missing values will reduce the sample size for analysis considerably if the missing rate is high. This might distort the results by introducing bias into the estimation of model parameters and the prediction accuracy in a predictive model. Thus it is necessary to consider some of the methods for handling missing data in order to mitigate this problem in diagnostic testing. It is also important to make robust assumptions regarding the missing data mechanism. In this work, we use three imputation techniques namely, mean, nearest neighbor hot-deck and multiple imputation (discussed in Chapter 4) to impute variables containing missing values.

The classic ROC curve deals with dichotomous diagnostic outcomes (presence or absence of disease) but in the real world we often deal with time-dependent disease outcomes and thus ROC curves that vary as a function of time become necessary [66].

The aim of this chapter is to improve the discriminating accuracy and performance of a diagnostic test, and compare imputation methods when estimating the time dependent AUC in the presence of missing values. That is to discriminate between subjects who may have an event of interest and those who may not. In addition it is appealing to use resampling methods to adjust for over-fitting associated with model selection that have been applied to choose important biomarkers or covariates. We applied different imputation methods when handling missing values and evaluated different resampling methods to estimate the time dependent area under ROC curves. Such an approach can be summed up as imputation before AUC estimation.

In this chapter, the biomarkers are combined using the Cox and logistic regression models in order to come up with a predictor index. We then predicted the status of cases in the validation set and then used the predicted values to calculate the time dependent AUC at time  $t$ . Bootstrap cross-validated time dependent AUC values were computed using the nearest neighbor estimation to measure the predictive accuracy [3, 66]. The bootstrapping technique was used to obtain the variances of the estimators of interest. The estimation methods were evaluated using simulations and illustrated using data on primary biliary cirrhosis PBC.

---

## **6.2. Missing data and imputation methods**

---

Recall from Chapter 4 that missing data are quite common in biomedical research studies. Some methodologists have described missing data as one of the most important statistical and design problems in research. The problem of missing data is of a greater concern when

decisions are to be made about the appropriateness of the care a patient should receive and also when we are interested in discriminating subjects as likely to have a certain characteristic from those who do not.

Recall that Rubin [113, 115] came up with three missing data mechanisms (discussed in Section 4.1). These mechanisms describe the relationships between measured variables and the probability of missing data. While these terms have a precise probabilistic and mathematical meaning, they essentially have different explanations for why the data were missed.

In this chapter we consider the three imputation methods (mean, nearest neighbor hot-deck and multiple imputation) discussed in Section 4.2.

---

### 6.3. Methods for estimation of $AUC(t)$

---

Let  $Y$  be a continuous biological biomarker whose values are an indicator of disease. Suppose if  $Y > c$  for some cut off  $c$  it implies an individual is diseased and disease free otherwise. With survival data we take the time to the event into account since the accuracy may be higher when the markers are measured closer to the onset of disease. ROC curves that vary as a function of time may be more appropriate to derive the corresponding time dependent ROC curves [67]. Definitions of time dependent ROC curves rely on first defining time dependent sensitivity ( $TPR(c, t)$ ) and specificity ( $1 - FPR(c, t)$ ). Then using simple plots of  $TPR(c, t)$  versus  $FPR(c, t)$  to get the ROC curve at a specific time point,  $t$ . The sensitivity and specificity are considered as time dependent functions and are given by Equations (3.1) and (3.2).

The time dependent area under the ROC curve denoted by  $AUC(t)$  is given by:

$$AUC(t) = \int_{-\infty}^{\infty} TPR(c, t) \left| \frac{\partial FPR(c, t)}{\partial c} \right| dc. \quad (6.1)$$

As mentioned before in Section 3.1 and according to Heagerty and Zheng terminology [67],

there are two ways of defining cases; cases are said to be *incident* if  $T_i = t$  and *cumulative* if  $T_i \leq t$  is used instead. On the other hand, controls are said to be *static* if  $T_i > t^*$ , where  $t^*$  is a fixed point in time and controls are said to be *dynamic* if  $T_i > t$ .

In this chapter we are interested in distinguishing between people who may experience the event of interest (which in our case is death) before a given time and those who may still be event free after that time. Thus we use the theory introduced by Heagerty [66], where they proposed summarizing the discrimination potential of a marker  $Y$  measured at baseline ( $t = 0$ ), by calculating *cumulative/dynamic ROC*( $t$ ) curves. They proposed an estimator that can accommodate censored survival data, which is based on a nearest neighbor estimator for the bivariate distribution function of  $(Y, T)$ , where  $T$  represents survival time. This estimator guarantees monotonicity and moreover the censoring process is allowed to depend on the diagnostic biomarker. Using these definitions (cumulative/dynamic), they defined the corresponding ROC curve for any time  $t$ ,  $ROC(t)$  by using an estimator of the bivariate distribution function  $F(c, t) = P(Y \leq c, T \leq t)$ , or equivalently  $S(c, t) = P(Y > c, T > t)$ , provided by Akritas [3]. This estimator is based on the representation  $S(c, t) = \int_c^\infty S(t|Y = s)dF_Y(s)$ , where  $F_Y(s)$  is the distribution function for  $Y$ .

We used the Heagerty et al. [66] approach which is briefly presented below. The authors proposed summarizing the discrimination potential of a marker  $Y$ , measured at baseline ( $t = 0$ ), by calculating ROC curves for cumulative disease incidence by time  $t$ , denoted as  $ROC(t)$ . A typical complexity with survival data is that observations may be censored. Two ROC curve estimators that can accommodate censored data were proposed by Heagerty et al. [66]. A simple estimator is based on using the Kaplan-Meier estimator for each possible subset  $\{i : Y_i > c\}$ . However, this estimator does not guarantee the necessary condition that sensitivity and specificity are monotone in  $Y$ . Another problem with KM-based ROC estimator

is that the conditional Kaplan-Meier estimator  $\hat{S}_{KM}(t|Y > c)$  assumes that the censoring does not depend on  $Y$ . This assumption may be violated as the intensity of follow-up efforts are influenced by the base line diagnostic marker measurements. An alternative estimator is based on a nearest neighbor estimator for the bivariate distribution function of  $(Y, T)$ , where  $T$  represents survival time. This estimator guarantees monotonicity in addition to the fact that it allows censoring to depend on  $Y$ .

The authors considered sensitivity and specificity as time-dependent functions defined as:

$$\begin{aligned} sensitivity(c, t) &= P(Y > c | D(t) = 1) \\ specificity(c, t) &= P(Y \leq c | D(t) = 0) \end{aligned}$$

In the first method Heagerty et al. [66] used the Bayes Theorem to rewrite the sensitivity and the specificity as

$$\begin{aligned} P(Y > c | D(t) = 1) &= \frac{1 - S(t|Y > c)P(Y > c)}{1 - S(t)}, \\ P(Y \leq c | D(t) = 0) &= \frac{S(t|Y \leq c)P(Y \leq c)}{S(t)}, \end{aligned}$$

where  $S(t)$  is the survival function  $S(t) = P(T > t)$  and  $S(t|Y > c)$  is the conditional survival function for the subset defined by  $Y > c$ . Define  $\tau_n$ , to be the unique values of  $X_i$ , for observed events,  $\Delta_i = 1$ . The Kaplan-Meier (KM) estimator is defined as:

$$\hat{S}_{KM}(t) = \prod_{s \in \tau_n, s \leq t} \left[ 1 - \frac{\sum_j 1(X_j = s) \Delta_j}{\sum_j 1(X_j \geq s)} \right].$$

The KM estimator uses all of the information in the data, including censored observations, to estimate the survival function. When obtaining the KM estimator, censoring is assumed to occur after an event therefore censored observation at or after an event time will be included in the risk set at that time.

A simple estimator for sensitivity and specificity at time  $t$  is then given by combining the KM



estimator and the empirical distribution function of the marker covariate,  $Y$ , as

$$\begin{aligned}\hat{P}_{KM}(Y > c|D(t) = 1) &= \frac{\{1 - \hat{S}_{KM}(t|Y > c)\}\{1 - \hat{F}_Y(c)\}}{\{1 - \hat{S}_{KM}(t)\}}, \\ \hat{P}_{KM}(Y \leq c|D(t) = 0) &= \frac{\hat{S}_{KM}(t|Y \leq c)\hat{F}_Y(c)}{\hat{S}_{KM}(t)},\end{aligned}$$

where  $\hat{F}_Y(c) = \sum 1(Y_i \leq c)/n$ . One problem with this simple estimator is that it does not guarantee that sensitivity (or specificity) is monotone. By definition, we require  $P(Y > c|D(T) = 1) > P(Y > c'|D(t) = 0)$  for  $c' > c$ . A valid ROC solution can be provided by using an estimator of the bivariate distribution function,  $F(c, t) = P(Y \leq c, T \leq t)$ , or equivalently  $S(c, t) = P(Y > c, T > t)$ , provided by Akritas [3]. This estimator is based on the representation  $S(c, t) = \int_c^\infty S(t|Y = s)dF_Y(s)$ , where  $F_Y(s)$  is the distribution function for  $Y$ . As shown by Akritas [3], an estimator can be provided by

$$\hat{S}_{\lambda_n}(c, t) = \frac{1}{n} \sum_i \hat{S}_{\lambda_n}(t|Y = Y_i)I(Y_i > c),$$

where  $\hat{S}_{\lambda_n}(t|Y = Y_i)$  is a suitable estimator of the conditional survival function characterised by a parameter  $\lambda_n$ .

Define the *weighted Kaplan-Meier estimator* as:

$$\hat{S}_{\lambda_n}(t|Y = Y_i) = \prod_{s \in \tau_n, s \leq t} 1 - \frac{\sum_j K_{\lambda_n}(Y_j, Y_i)I(Z_j = s)\delta_j}{\sum_j K_{\lambda_n}(Y_j, Y_i)I(Z_j \geq s)},$$

where  $K_{\lambda_n}(Y_j, Y_i)$  is a kernel function that depends on a smoothing parameter  $\lambda_n$ . Akritas [3] shows that the nearest neighbor estimator (NNE) is a semiparametric efficient estimator.

The resulting estimates of sensitivity and specificity mentioned by Heagerty et. al. [66] are given by:

$$\begin{aligned}\hat{P}_{\lambda_n}(Y > c|D(t) = 1) &= \frac{[(1 - \hat{F}_Y(c)) - \hat{S}_{\lambda_n}(c, t)]}{1 - \hat{S}_{\lambda_n}(t)}, \\ \hat{P}_{\lambda_n}(Y \leq c|D(t) = 0) &= 1 - \frac{\hat{S}_{\lambda_n}(c, t)}{\hat{S}_{\lambda_n}(t)},\end{aligned}$$

where  $\hat{S}_{\lambda_n}(t) = \hat{S}_{\lambda_n}(-\infty, t)$ .

The proposed estimators of Heagerty et al. [66] yield time-dependent ROC curve methods that provide a natural way for handling censored survival data, and involve no parametric assumptions. Heagerty et al. assume the measurement time for the prognostic marker,  $Y$ , is fixed at baseline and as a result the ROC curve is only a function of the disease ascertainment time  $t$ . When a marker is measured repeatedly over time, a method that also incorporates the time at which the measurement was obtained allows for an updating of the medical decision at the current follow-up time. Heagerty et al. estimators assume that the data are derived from a cohort study where sampling does not depend on the disease outcome  $D(t)$ .

---

#### 6.4. Models for predictive scores

---

Since we have several prognostic variables, there is need to use multivariate approaches. In this work we used two statistical models to estimate predictor scores, then use these scores to estimate the AUC(t). First, logistic regression (discussed in Section 5.4) was used to build the model when the outcome or response is defined as the presence or absence of a condition or disease.

Logistic regression cannot deal with censored observations and does not take account of time. Thus an alternative method is to use survival analysis models in the presence of censoring. One very popular model for survival data is the *Cox proportional hazards model* (CPH), which was proposed by Cox [6, 137]

The Cox model [6] is based on a modeling approach to the analysis of survival data. It is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. The Cox model provides an estimate of the treatment effect on survival

after adjustment for other explanatory variables. In addition, it allows us to estimate the hazard or risk of death for an individual, given their prognostic variables. In a clinical trial the Cox model is used to analyse the survival of patients. The model can also be used if it is known that there are other variables besides treatment that influence patient survival and these variables cannot be easily controlled in a clinical trial. Using this model may improve the estimate of treatment effect by narrowing the confidence interval. The regression method introduced by Cox is used to investigate the effect of several variables on the hazard function. It is also known as proportional hazards regression analysis. Briefly, the procedure models or regresses the hazard of an event on the explanatory variables. Thus final model from a Cox regression analysis will yield an equation for the hazard as a function of several explanatory variables.

The Cox proportional hazards model describes survival data with covariates in terms of a hazard function of the form:

$$h(t|X) = h_0(t) \exp(\beta' X),$$

where  $\beta$  is an unknown vector of parameters,  $h_0(t)$  is the baseline hazard and  $X$  is a vector of covariates. The unknown vector  $\beta$  can be estimated by solving the partial likelihood:

$$L(\beta) = \prod_{i=1}^f \frac{\exp(\beta' X(i))}{\sum_{j \in R(t_i)} \exp(\beta' z_j)},$$

where  $R(t_i)$  is the risk set at event time  $t_i$  and  $f$  different failure times  $t_1 < t_2 < \dots < t_f$  with exactly one failure at each time. The estimate of the survival function for an individual with covariates  $X$  may be obtained via:

$$S(t) = [S_0(t)]^{\exp(\beta' X)}.$$

Note in the above equation  $S_0(t)$  is known as the baseline survival when  $X = 0$ . The Cox regression analysis yields an equation for the hazard as a function of several explanatory

variables. A coefficient for a single covariate give the log-hazard for that variable given all other variables included in the model. A positive regression coefficient for an explanatory variable means that the hazard is higher, and thus the prognosis is worse. Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable. This model is a *semi-parametric model* because it makes no assumptions about the form of  $h_0(t)$  the non-parametric part of the model, but assumes a parametric form for the effect of the predictors on the hazard. Corresponding to  $h_0(t)$  is  $S_0(t)$  - also known as the baseline survival - which also remains unspecified. The beauty of this model, as observed by Cox, is that if one is to use such a model, and one is interested in the effects of the covariates on survival, then one does not need to explicitly specify the form of  $h_0(t)$ . A standard Cox regression model can be used to derive a composite marker effect as a weighted combination of biomarkers and clinical variables, in which the weights are determined by the estimated regression coefficients. Recall that a Cox regression model is specified via the hazard function, which is defined as the instantaneous rate at which failures occur for individuals that are surviving at time  $t$ , therefore it is formally defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T < t + \Delta t]}{\Delta t}.$$

The Cox regression model employs a log function to relate the hazard function to a linear combination of biomarkers and clinical variables:

$$\log h(t) = \log h_0(t) + \sum_{i=1}^p \beta_i X_i + \sum_{j=1}^q \gamma_j Z_j,$$

where  $\beta_i$  are regression parameters that correspond to biomarkers  $X_i$  and  $\gamma_j$  are regression parameters that correspond to clinical variables  $Z_j$ . Since  $h(t)$  is a product of  $h_0(t)$  and a term that is a function of the biomarkers and clinical variables this leads to a proportional hazards model if both the biomarkers and the clinical variables are baseline variables and are constant over time. Censoring can be accommodated in likelihood-based estimation of the regression

parameters, but censoring must be assumed to be independent of survival, i.e. non-informative censoring. The estimated regression coefficients can then be used to derive a composite marker index  $M$  as a weighted combination of biomarkers and clinical variables given by

$$M = \sum_{i=1}^p \beta_i X_i + \sum_{j=1}^q \gamma_j Z_j.$$

To quantify the predictive accuracy of the composite marker,  $M$  can be used as the input to a time-dependent ROC analysis. Note that  $M(t)$  is time dependent because now cases are defined according to a time to event outcome.

It is not desirable to use the same data to both develop and evaluate the models, thus cross-validation and bootstrapping were used to estimate the time dependent AUC denoted by  $AUC(t)$  in the context of variable selection. Using the same data both to fit the score and to calculate its ROC at a specific time leads to what is known as over-fitting and this problem tends to give an over optimistic estimate of the performance of the score [25]. We remark that variable selection should be done from scratch using a number of training sets not the complete data.

Leave-one-out cross-validation (LOOCV) was used together with multivariate Cox and logistic regression models to fit the prediction model. Bootstrapping was used to get the bootstrap cross-validated  $AUC(t)$  and its variance estimate. In this case we used 200 bootstrap replicates in both simulation studies and application to real data set.

For each of the training sets, both the logistic regression and CPH models were used to build the predictive model and for each variable selection was applied from scratch for each training set. We then predicted the part of the data left out. These procedures are repeated until every observation appears once in a validation set. Finally the predicted values were used to estimate the time dependent AUC. Bootstrapping was applied to estimate  $AUC(t)$  as well as its vari-

ances and therefore standard errors for confidence interval estimation. We were interested in comparing the two models in three aspects namely, obtaining the predictive scores, evaluation of two resampling methods in terms of variable selection as methods to estimate  $AUC(t)$  and finally performance models (when comparing between imputation methods discussed earlier) for handling missing values.

---

### **6.5. Algorithm**

---

The following is an algorithm to estimate the AUC.

- Considering the missing value problem. Most of the current research ignores this problem yet it is very important to deal with missing values. In our algorithm we consider single and multiple imputation methods as described in Chapter 4. For mean imputation we used mean function from “ForImp” package in R software. The function “impute.NN<sub>HD</sub>” in HotDeck Imputation was used to apply nearest neighbor hot deck imputation and MICE for multiple imputation via chained equations. We evaluate and compare these methods in order to estimate time dependent AUC.
- The second step needs to be undertaken to optimise the set of biomarkers with high and independent diagnostic information content in a multivariate setting. In our algorithm we use logistic regression using glm function in stats package and Cox model using coxph in survival package.
- Before applying the algorithm to derive a discriminatory rule and in order to avoid an over-fitting problem associated with model selection, we split each original dataset into a training set (to build the model) and a validation set (to test whether the score discriminates between the same diagnostic classes in an independent group of subjects).

This split was done as many as the number of individuals, as we use LOOCV. We also reported bootstrap LOOCV estimator of the  $AUC(t)$  and its variance. The diagnostic performance of the discriminant score was assessed by estimating sensitivity, specificity and diagnostic accuracy by plotting the ROC curves at our interested time. As stated earlier, variable or feature selection should be done from scratch for every training set.

- For the prediction score from the logistic regression or Cox model, a non-parametric estimator proposed by Heagerty [66] for time dependent AUC was used. The `survivalROC.C` function in R was used with nearest neighbor estimation of time dependent AUC.

---

## **6.6. Simulation studies**

---

In order to evaluate the performance of the suggested resampling methods and different imputation techniques for computing time dependent AUCs, we conducted simulation studies as follows. For each of  $N = 50$  subjects we simulated a survival time including five biomarkers with a possibility of missing values. The survival time  $T$  was generated from the exponential distribution. The censoring indicator was generated using the binomial distribution with 40% censoring rate.

Missing values on the biomarkers were generated using missing rates of 10% and 20% under MCAR and MAR assumptions. As a summary our simulated data sets contain the survival times, the censoring indicator and five biomarkers with missing values. Then we imputed these missing values using the mean, nearest neighbor hot-deck (NNE HD) and multiple imputation methods.

Biomarkers from the simulated data set were used to evaluate whether a model output based on a combination of these biomarkers can accurately discriminate between individuals who are

likely to experience the event before a given time (120 days) from those who may not after that time. This was accomplished by applying time dependent AUC analysis and resampling methods for imputed data sets.

Our analysis of simulated data was based on estimation of the following quantities: Cross-validated time dependent AUC denoted by  $AUC_C^t$ , bootstrap time dependent AUC denoted by  $AUC_B^t$ , bootstrap standard error denoted by  $SE_B$ , confidence interval using bootstrapping denoted by  $CI_B$ . We summarize these results in Table 6.1 below.

From Table 6.1, under MCAR assumption and when the missing rate is 10% using Cox model and mean imputation, the estimated value of  $AUC_B(t)$  is equal to 0.724 and according to its confidence interval (0.571, 0.877), this diagnostic test is statistically significant (CI excludes 0.5). Thus it has the ability to distinguish between patients who are likely to die during the first four months and those who are likely to survive beyond that time. The estimated value of  $AUC_{cv}(t)$  is equal to 0.687. Using the NNE HD imputation, the estimated value of  $AUC_{bcv}(t)$  is equal to 0.743 and according to its confidence interval (0.613, 0.902,) this diagnostic test is statistically significant. The standard error based on bootstrapping equals to 0.081. With the MI, the estimated value of  $AUC_{bcv}(t)$  equals to 0.746 and associated standard error equals to 0.083. The confidence interval (0.583, 0.909) shows that the diagnostic test is statistically significant and could predict the survival at 120 days.

Under MCAR assumption and when the missing rate is 10% using logistic regression and mean imputation, the estimated value of  $AUC_{bcv}(t)$  is equal to 0.638. According to its CI (0.434, 0.842), the combinations of biomarkers are not statistically significant (CI includes 0.5) and do not have any ability to discriminate between subjects who may die during 120 days and those who may not. The estimated value of  $SE_B$  is equal to 0.104. With the same settings but using NNE HD and MI we obtained the same conclusion.



Table 6.1: Mean of the time dependent AUC at 120 days obtained from 250 simulated samples under MCAR and MAR for each combination of censoring rate, predictive model and imputation method.  $B = 1000$  bootstrap replications are performed for computing the  $AUC_{cv}$ ,  $AUC_{bcv}$ ,  $SE$  and  $CI$ s for the AUCs

Mechanism of missingness	Missing rate	Models	Imputation methods	$AUC_{bcv}(t)$	$SE_B$	$CI_{bcv}$	$AUC_{cv}(t)$
MCAR	10%	Cox	Mean	0.724	0.078	(0.57112, 0.87688)	0.687
			HD	0.743	0.081	(0.6134, 0.90176)	0.711
			MI	0.746	0.083	(0.58332, 0.90868)	0.706
		Logistic	Mean	0.638	0.104	(0.43416, 0.84184)	0.592
			HD	0.659	0.106	(0.45124, 0.86676)	0.62
			MI	0.635	0.109	(0.42136, 0.84864)	0.59
	20%	Cox	Mean	0.732	0.078	(0.65712, 0.88488)	0.693
			HD	0.784	0.074	(0.63896, 0.92904)	0.757
			MI	0.770	0.072	(0.62888, 0.91112)	0.736
		Logistic	Mean	0.641	0.104	(0.43716, 0.84484)	0.598
			HD	0.734	0.086	(0.56544, 0.90256)	0.698
			MI	0.630	0.084	(0.46536, 0.79464)	0.661
MAR	10%	Cox	Mean	0.709	0.074	(0.56396, 0.85404)	0.673
			HD	0.714	0.073	(0.571, 0.85708)	0.707
			MI	0.704	0.078	(0.55112, 0.85688)	0.682
		Logistic	Mean	0.605	0.109	(0.39136, 0.81864)	0.542
			NNE HD	0.68	0.104	(0.47616, 0.88384)	0.602
			MI	0.614	0.107	(0.40428, 0.82372)	0.608
	20%	Cox	Mean	0.746	0.069	(0.61076, 0.88124)	0.738
			NNE HD	0.813	0.078	(0.66012, 0.96588)	0.802
			MI	0.800	0.069	(0.66476, 0.93524)	0.800
		Logistic	Mean	0.674	0.096	(0.48584, 0.86216)	0.644
			NNE HD	0.751	0.086	(0.58244, 0.91956)	0.726
			MI	0.719	0.1	(0.523, 0.915)	0.686

Under MCAR assumption for 20% missing rate, the estimated values of  $AUC_{bcv}(t)$  using the Cox model with the three imputation methods suggested that the combined markers have the discrimination ability.

Under MCAR assumption for 20% missing rate, the estimated values of  $AUC_{bcv}(t)$  using the logistic model with mean and multiple imputation methods suggested that the combined markers do not have the ability to discriminate between patients who will die during the first four months and those who may be alive after that time. However with NNE HD imputation the confidence interval (0.565, 0.903) excludes 0.5.

In all the above cases bootstrap cross-validation seems to perform better than just cross-validation. The NNE HD performed better in most cases when compared to other imputation methods in that there are slightly higher AUC(t) estimations. All imputations methods perform equally well if the missing mechanism is MCAR. It is the difference that the Cox model for time dependent AUC is clearly better than the logistic regression as the predictive model.

For the MAR mechanism when the missing rate is either 10% or 20%, using Cox model the estimated values of  $AUC_{bcv}(t)$  from the three imputation methods are similar (NNE HD obtained slightly higher AUC(t) estimations) and statistically significant. These results suggest that the combination of biomarkers have the ability for discrimination in 120 days.

Under MAR assumption and when the missing rate is 10% using logistic regression with the mean imputation, NNE HD imputation and MI, the estimated values of  $AUC_{bcv}(t)$  equal to 0.605 ,0.68 and 0.614 respectively. The combinations of biomarkers are not statistically significant and do not have any ability to discriminate between subjects who may die during 120 days and those who may not.

Under MAR assumption and when the missing rate is 20% using logistic regression and mean

imputation, the estimated value of  $AUC_{bcv}(t)$  equals to 0.674 with  $SE_B = 0.096$ . We found that the combinations of biomarkers are not statistically significant.

When the missing rate is 20%, it could be seen that with all imputation techniques- and for MAR- the resulting  $AUC_{bcv}(t)$  values are better than those when the mechanism is MCAR. This indicates that for higher missing rates- and specially if MAR was the case- then applying imputation methods will be necessary and may help improve the results.

The true simulated  $AUC(t)$  is 0.742 which is clearly close to the values estimated under the Cox model than those estimated under the logistic regression model. This emphasizes the point that the Cox model is best suited to handle the estimation of time dependent AUCs than the logistic regression.

From Table 6.1 we found that nearest neighborhood hot deck reveals higher  $AUC_{bcv}(t)$  and  $AUC_{cv}(t)$  estimates. However the difference between AUC estimates obtained using the three imputation methods were not statistically significant. The Cox regression as a predictive model is better than the logistic regression. Its AUC values are higher than those from the logistic model. This tells us that the appropriate model for the data should be the first step in the process. Since in this work the data set is time to event, it is obvious that Cox regression model performs better as it gives  $AUC(t)$  estimations similar to the true AUC. Bootstrapping performed better than cross-validation, however the differences in values of AUCs are small.

---

### **6.7. Application to primary biliary cirrhosis (PBC)**

---

This data is from the Mayo Clinical trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during a ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the

drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial while the other 112 cases did not. Here we consider the first 312 cases.

In Table 6.2 we show the missingness rate for the covariates that have missing values:

Table 6.2: Missing rate in PBC dataset

Covariate	chol	copper	trig	platelet
Missings rate	28 (9%)	2 (1%)	30 (10 %)	4 (1 %)

We used the PBC data set to develop the clinical prediction model. In our analysis we compare between model scores using the time dependent ROC analysis. Table 6.3 contains results for two models- namely that based on list wise deletion (Model 1) and the multiple imputation model (Model 2). We fitted the multivariate Cox model as the predictive model under both missing data handling method and this table gives the significant variables from each model, which are strong indicators of mortality. The number of observations is 276 after list wise deletion, meaning 36 patients were not used in analysis. From the table we can see that there is a covariate that appears under Model 1 and does not appears in the Model 2 results and vice-versa. This indicates that the method used to deal with missing data may affect the final predictive model even if not in a big way.

We considered time dependent AUC estimation distinguishing between people who may die and those who may not.

A larger AUC(t) value would suggest that the model score is more accurate in its discriminatory capacity. Based on Cox regression results for each model in Table 6.3, the estimated AUC(t) value for Model 1 is equals to 0.90 and the estimated AUC(t) value for Model 2 is equals to 0.92. These results suggest that both predictive models have high ability to distinguish

Table 6.3: Cox regression estimates for PBC dataset

	Covariate	Estimate	SE	Z-statistic	p-values	CI
Model 1	age	0.00007173	0.0000309	2.322	0.0203 *	(1.00, 1.0001)
	edema	0.9248	0.3782	2.445	0.0145 *	(1.2014, 5.2915)
	log(bili)	0.7221	0.1618	4.463	8.1e-06 ***	(1.4993, 2.8272)
	albumin	-0.678	0.3049	-2.223	0.0262 *	(0.2792, 0.9228)
	stage	0.3867	0.1757	2.201	0.0277 *	(1.0432, 2.0772)
Model 2	age	0.00007493	0.00002833	2.645	0.00818 **	(1.00, 1.0001)
	edema	0.8375	0.3283	2.551	0.01074 *	(1.2142, 4.3969)
	log(bili)	0.7342	0.1504	4.883	1.04e-06 ***	(1.5519, 2.7980)
	albumin	-0.7039	0.2802	-2.512	0.01199 *	(0.2857, 0.8566)
	log(protime)	2.798	1.191	2.349	0.01882 *	(1.5898, 169.5695)

where in Table 6.3, “\*” means that the  $p$ -value is less than 0.05, the “\*\*” means that the  $p$ -value is less than 0.01 and the “\*\*\*” means that the  $p$ -value is less than 0.001.

between patients who are likely to die from those who are not. However the imputation model (Model 2) is better than the model that ignores missing values (however the difference is very small).

We also obtained time dependent AUCs when we used the Cox and logistic regression models. Figure 6.1 contains time dependent ROC curves for the two models at the first year. It is clear that the two ROC curves are very close to each other at 365 days. We expect the Cox model to perform better, thus we investigate the performance of the two models over long period. Figure 6.2 explains the performance of the two models and their AUC values from minimum time until ten years.

In the first 1000 days the two models are quite similar which is supported by Figure 6.2.

However after over time, the Cox regression model seems to perform better and thus the score obtained has higher ability for discrimination purpose.

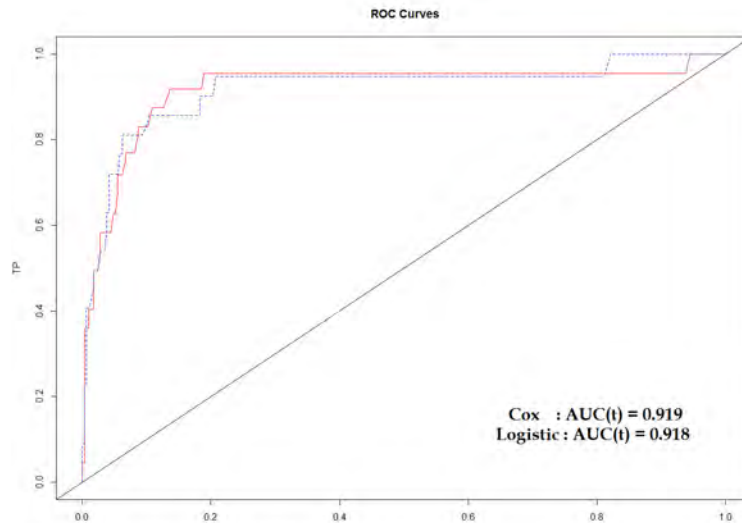


Figure 6.1: ROC curves for Cox model and logistic regression for  $t = 365$  days using PBC dataset

In survival data, the accuracy may be higher when the markers are measured closer to the onset of disease. In Figure 6.3 we plotted three  $ROC(t)$  curves at different times to investigate the performance of the model and its ability to discriminate with time.

Figure 6.3 shows that predictive accuracy decreases with increasing time since baseline. It is clear that the best  $ROC(t)$  is when time is 180 days. However the area under the  $ROC(t)$

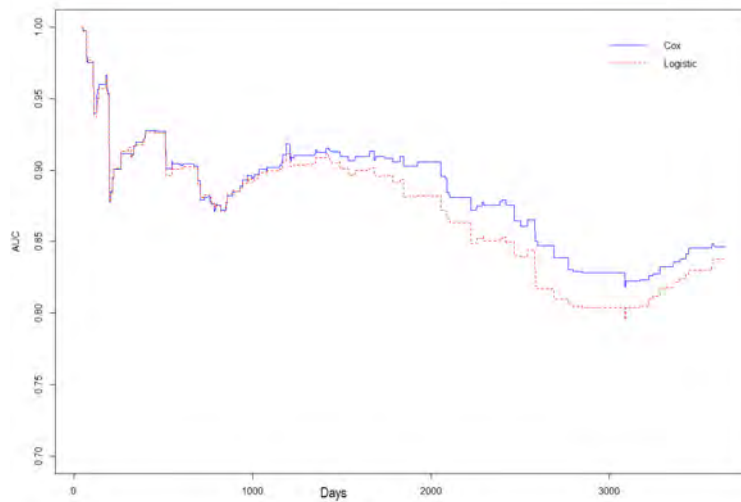


Figure 6.2: AUC estimates up to 10 years for PBC dataset

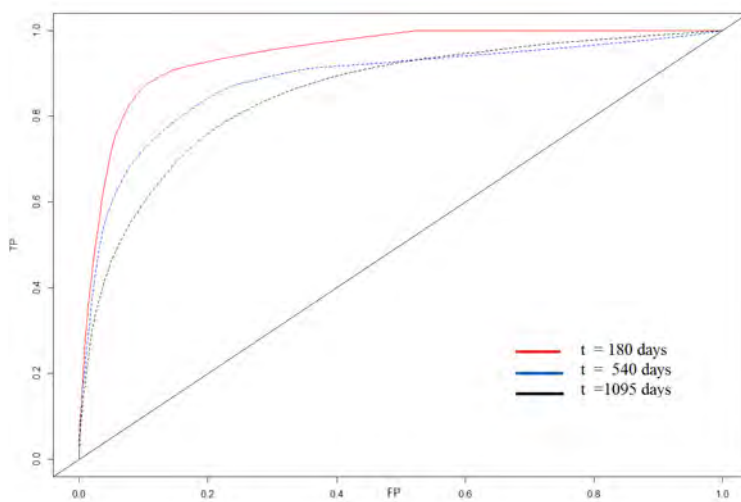


Figure 6.3: ROC curves at different times using the Cox regression model for PBC dataset

curves become smaller when it goes to 1000 days and thereafter. This indicates that the closer the biomarkers are measured to the event time the better is the  $AUC(t)$  and thus there is a better discrimination ability.

Table 6.4 shows the results for  $AUC(t)$  values based on resampling methods and when variable

selection were used to evaluate whether the obtained scores could accurately discriminate between people who are likely to die before the first year and those who may still be alive after that time. In this table we list the values of  $AUC_B(t)$ ,  $SE_B$ ,  $CI_B$  and  $AUC_C(t)$ . These estimates were based on the Cox and logistic regression models under three imputation methods for missing values.

Table 6.4: The  $AUC_{cv}$  and  $AUC_{bcv}$  estimations using mean imputation, nearest neighbor imputation and multiple imputation

Models	statistic	Mean imputation	Hot deck imputation	Multiple imputation
Cox	$AUC_B(t)$	0.932	0.0.930	0.931
	$SE_B$	0.034	0.034	0.035
	$CI_B$	(0.865, 0.999)	(0.863, 0.997)	(0.862, 1)
	$AUC_C(t)$	0.932	0.925	0.924
Logistic	$AUC_B(t)$	0.0.921	0.922	0.922
	$SE_B$	0.032	0.030	0.0.030
	$CI_B$	(0.858, 0.984)	(0.863, 0.981)	(0.863, 0.982)
	$AUC_C(t)$	0.932	0.0.932	0.932

The following is an explanation for the results in Table 6.4.

- Using the Cox model to build the predictive model for the predicted scores, we found the following:
  - With the mean imputation method, the estimated  $AUC_B(t)$  is equal to 0.932 and estimated  $SE_B$  equal to 0.034. The 95% bootstrap CI is (0.865, 0.999) which means that this score is statistically significant and has a high ability to discriminate at



365 days.

- With the NNE HD method the values of  $AUC_B(t)$ , their  $CI_B$ s, suggest that this predicted score is statistically significant. It is also appears that the obtained value of  $AUC_C(t)$  is also high and very close to  $AUC_B(t)$ .
- With multiple imputation method, the estimated  $AUC_B(t)$  equals to 0.931 and according to  $SE_B$  and  $CI_B$  the score has high ability to discrimination at the first year. The obtained value of  $AUC_C(t)$  is a bit smaller than  $AUC_B(t)$ , however the difference is very small.
- Using the logistic regression model to build the predictive model for predicted scores, we have the following:
  - With the mean imputation method, the estimated bootstrap AUC  $AUC_B(t)$  and its corresponding CI confirm that this score is statistically significant which shows it has a high ability to discriminate between subjects who are likely to die before 120 days and those who are likely to survive beyond that time.
  - With the NNE-HD and MI methods the estimated values are almost identical and very close to the estimate obtained by using the mean imputation as well. These results suggest that the score is statistically significant.

From all the above findings, we conclude that both resampling methods perform well and similarly. Bootstrapping seems to perform better but the difference is small. Estimates from all the imputation techniques are very similar to each other in both estimated  $AUC(t)$ s from resampling methods. According to the null hypotheses

$$H_0 : AUC_A = AUC_B, H_0 : AUC_A = AUC_C \text{ or } H_0 : AUC_B = AUC_C,$$

where A, B and C are three different imputation strategies, the obtained statistics suggest

that there is no evidence to reject  $H_0$  and thus it will not make difference to which imputation method we use in order to estimate the  $AUC(t)$ .

Parameter estimates for some covariates (with missing values) obtained from multiple imputation with  $m = 5$  are given in Table 6.5. In this table, W, B and T stand for: within, between and total imputation variances respectively.

Table 6.5: Pooled estimates from MI for some variables in PBC dataset

Covariate	Est	$W$	$B$	$T$
Copper	97.66	23.63	0.32	24.01
Platelet	261.98	29.20	0.57	29.88
Cholesterol	367.08	174.75	10.58	187.44
triglycerides	124.91	13.46	0.44	13.99

---

## 6.8. Conclusion

---

The time dependent area under receiver operating characteristic is an important summary index of discriminatory accuracy in modern clinical medicine. In this chapter we evaluated some of the imputation techniques and resampling methods for the purpose of estimating time dependent AUCs in the presence of missing data.

According to our simulation studies, we deduced that LOOCV gives good estimates of the AUC, however bootstrapping LOOCV seems to perform even better. An interesting observation is that the three imputation methods gave rise to similar discrimination accuracy where the  $AUC(t)$  estimates obtained by the nearest neighborhood hot deck method were slightly higher. However we recommend the use of multiple imputation method as it represents missing data

uncertainty and takes into account the variability given by the multiple imputed data set with appropriate statistical inference. The Cox model performed better than the logistic regression and specifically at 120 days most results suggest using the logistic regression as the combination method did not have ability to discriminate at that time.

An application to the PBC data set reveals that both cross-validation and bootstrap cross-validation yield  $AUC(t)$  estimations are close to each other. Cox regression outperforms the logistic regression for large period estimation time.

# **Use of resampling methods to predict the outcome of tuberculous meningitis in high HIV prevalence patients in South Africa**

---

## **7.1. Introduction**

---

Early diagnosis of disease has the potential to reduce morbidity and mortality. Biomarkers may be useful for detecting disease at early stages before it becomes clinically apparent. Our interest is to evaluate the predictive and discriminatory capacity for the diagnostic index.

Meningitis causes significant mortality and morbidity in HIV infected persons [16, 56, 71, 89, 131]. Tuberculous meningitis (TBM) accounts for a substantial proportion of deaths, particularly in high tuberculosis (TB) prevalence areas [71]. However, few studies have investigated the predictors of mortality in patients with HIV associated TBM.

Missing data is a common problem in many fields especially in medical clinic research. It is

very important to consider missing values in order to reduce the bias associated with parameter estimates. We applied nearest neighbor hot-deck imputation (discussed in Subsection 4.2.2) to consider missing values associated with TBM/HIV data set. This method is suitable for any kind of variable because it does not require strong distributional assumptions and standard statistical methods could easily be applied to the imputed data.

When we have many variables and the interest is to estimate the diagnostic test accuracy, we have to select the best variables in order to have better predictive accuracy. Moreover resampling methods are recommended to adjust for over-fitting associated with variable selection.

To estimate the  $AUC(t)$ , we used the proposed method by Heagerty [66] to calculate ROC curves for cumulative disease or death incidence by time  $t$ , which we denote as  $ROC(t)^{C/D}$ . See Section 3.1 for more details.

A retrospective study of Tuberculous Meningitis in a high HIV prevalence settings at GF Jooste Hospital in Cape Town is used to describe the application of cross-validation, 632 and 632+ bootstrapping to estimate  $AUC(t)$ . We also described the presentation and outcome of patients with TBM. The nearest neighbor hot-deck technique was used to impute the missing values in the variables used to construct the TBM-IRIS scores. These scores were then used to estimate the  $AUC(t)$  with respect to discriminate between TBM-IRIS patients who may die before six months and those who may survive after that time.

Penalized regression models provide a statistically appealing method to build prediction models, where the aim is to simultaneously select features and to fit the model [45, 121]. Since the introduction of the LASSO for linear regression models (Tibshirani [138]), the methodology has been extended to generalized linear regression models and time-to-event endpoints among others. In addition to the well-known  $L_1$ -norm (LASSO) and  $L_2$ -norm (ridge) penalty

functions, various other penalties have been proposed in recent years to select features and/or estimate their effects. In particular, we will use the  $L_1$ -norm and  $L_2$ -norm.

We used two methods to estimate the cross-validated, 632 and 632+ AUC(t). These are the proposed method of using *ridge-Cox regression* and *least absolute selection and shrinkage operator (LASSO)-Cox regression* [50]. Ridge-Cox regression and LASSO-Cox regression methods were evaluated through simulation studies. The AUC(t) was used to compare resampling methods with respect to predictive power. Estimation of the variances of the estimated AUC values are very necessary when the interest is to evaluate the performance of combined biomarkers or predictor index. Thus we considered and evaluated (through simulation studies) two level bootstrapping to estimate standard errors and then confidence intervals.

---

## **7.2. Penalized Cox methods**

---

The *Ordinary Least Squares* (OLS) estimates are obtained by minimising the residual squared error [138], which is the difference between the observed and estimated function value. There are some problems with estimates obtained by OLS, the first is the difficult interpretation. In addition to that OLS estimates demonstrate large variances despite having low bias.

There are two solutions to improving OLS estimation [138], namely *subset selection* and *ridge regression*. Subset selection has interpretable models but can be extremely variable because it is a discrete process and regressors are either retained or dropped from the model. A small change in data can result in different models being selected and this can reduce its prediction accuracy. Ridge regression has been proposed as an alternative, it is a continuous process that shrinks coefficients and hence is more stable. However it does not set any coefficients to zero and hence does not give an easily interpretable model [138].

### 7.2.1 LASSO-Cox regression

Tibshirani [138] proposed the least absolute selection and shrinkage operator (LASSO) for variable selection and shrinkage. The proposal by Tibshirani is based on minimising the log partial likelihood subject to the sum of the absolute values of the parameters being bounded by a constant. This technique shrinks coefficients and reduces some to zero.

As we mentioned in Section 6.4,  $\beta$  could be estimated through maximisation of the partial likelihood. Denote the partial log-likelihood of  $\beta$  as  $\ell(\beta)$  which is given by:  $\ell(\beta) = \log L(\beta)$ . Then there are two definitions for LASSO [138] to estimate  $\beta$ , these definitions are:

- $\hat{\beta} = \operatorname{argmax} \ell(\beta)$ , subject to  $\sum \|\beta_j\| \leq s$ , where  $s > 0$  is user specified parameter and  $\|\cdot\|$  is  $L_1$  norm,
- $\hat{\beta} = \operatorname{argmax} \ell(\beta) - \lambda \|\beta\|_1$ ,

where  $\beta$  is the vector of regression coefficients and  $\lambda$  is the tuning parameter for  $L_1$ . LASSO is attractive as a regularisation method because it simultaneously performs variable selection and shrinkage. It shrinks all regression coefficients towards zero and automatically sets many of them exactly to zero. Variable selection is desirable in order to obtain an interpretable prediction rule, and shrinkage is desirable to prevent over-fitting.

Geoman [53] proposed a method that presents a novel full gradient algorithm for maximising the LASSO-penalized likelihood. It follows the gradient of the likelihood from a given starting value of  $\beta$ . Their method uses the full gradient at each step, furthermore the algorithm can automatically switch to a Newton Raphson algorithm when it gets close to the optimum [53].

Let  $\beta = (\beta_1, \dots, \beta_P)^T$ , the target function is defined as:

$$\ell_{pen}(\beta) = \ell(\beta) - \lambda \sum_{i=1}^P |\beta_i|. \quad (7.1)$$

This function includes two terms, the first term is the log likelihood function and the second one is penalty  $p(\beta) = -\lambda \sum_{i=1}^P |\beta_i|$ . Geoman [53] mentioned two important points: Firstly the penalized likelihood function is the sum of two concave functions and it is itself a concave function. However this is not generally strictly concave. Secondly, the penalized likelihood function is not differentiable everywhere due to the lack of differentiability of the penalty function.

The gradient ascent algorithm is very simple to understand, but it requires a large number of steps to converge. Geoman [53] proposed the option of switching to the Newton Raphson algorithm to avoid too many steps. For more details about this two algorithms we refer the reader to [53].

It is possible to define a directional derivative

$$\ell'_{pen}(\beta, v) = \lim_{t \rightarrow 0} \frac{1}{t} [\ell_{pen}(\beta + tv) - \ell_{pen}(\beta)], \quad (7.2)$$

for every point  $\beta$  in every direction  $v \in \mathbb{R}$ . The gradient can then be defined for every  $\beta$  as the scaled direction of steepest ascent. Let  $v_{opt}$  be the direction that maximises  $\ell'_{pen}(\beta, v)$  among all  $v$  with  $\|v\| = 1$  then the gradient can be defined as  $\ell'_{pen}(\beta, v_{opt}) - v_{opt}$  if  $\ell'_{pen}(\beta, v_{opt}) \geq 0$  and  $\mathbf{0}$  otherwise, where  $\mathbf{0}$  is a  $p$ -vector of zeros. We can define the directional second derivative as:

$$\ell''_{pen}(\beta, v) = \lim_{t \rightarrow 0} \frac{1}{t} [\ell'_{pen}(\beta + tv, v) - \ell'_{pen}(\beta, v)]. \quad (7.3)$$

For the penalized log likelihood the directional second derivative is given for every  $\beta$  and  $v$  by:

$$\ell''_{pen}(\beta, v) = v' \frac{\partial^2 \ell(\beta)}{\ell(\beta) \ell(\beta')} v. \quad (7.4)$$



In practice it is hardly ever necessary to calculate the full Hessian matrix of  $\ell(\beta)$  to calculate the directional second derivative, as the direction  $v$  of interest, which is the direction of the gradient, will typically have many zeros. Furthermore, in the Cox proportional hazards model, as well as in generalised linear models with a canonical link function, the Hessian matrix is of the form

$$\frac{\partial^2 \ell(\beta)}{\ell(\beta)\ell(\beta^T)} = X^T W X, \quad (7.5)$$

where  $X$  is an  $n \times p$  design matrix and  $W$  an  $n \times n$  weights matrix. This structure of the Hessian matrix allows the algorithm to avoid construction of the full  $p \times p$  Hessian. The gradient ascent algorithm uses a series of Taylor approximations. At each step it approximates the penalty locally from  $\beta$  in the direction of the gradient by a directional second order Taylor approximation, given by:

$$\ell_{pen}(\beta + tg(\beta)) \approx \ell_{pen}(\beta) + t\ell'_{pen}(\beta, g(\beta)) + \frac{1}{2}t^2\ell''_{pen}(\beta, g(\beta)). \quad (7.6)$$

This approximation is meaningful only within a single sub domain of gradient continuity, for  $0 < t < t_{edge}$ , with

$$t_{edge} = \min \left\{ -\frac{\beta_i}{g(\beta)} : \text{sign}(\beta_i) = -\text{sign}[g_i(\beta)] \neq 0 \right\}, \quad (7.7)$$

where

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

The optimum of the Taylor approximation in the subdomain is at:

$$t_{opt} = -\frac{\ell'_{pen}(\beta, g(\beta))}{\ell''_{pen}(\beta, g(\beta))}, \quad (7.8)$$

provided  $t_{opt} < t_{edge}$ , otherwise it is at  $t_{edge}$ . The algorithm proceeds in every next step with a new directional Taylor approximation from the optimum found in the previous one.

Convergence occurs when  $g(\beta) = 0$ . If there is not a unique optimum, the algorithm will converge to a point in the optimal area.

- Start with some  $\beta^0$ .
- For steps  $i = 0, 1, \dots$  of the algorithm, iterate the following

$$\beta^{i+1} = \beta^i + \min(t_{opt}, t_{edge})g(\beta^i).$$

Let  $\tilde{\beta} = (\beta_{J_1}, \dots, \beta_{J_m})^T$  and let  $\tilde{g}(\beta) = (g_{J_1}(\beta), \dots, g_{J_m}(\beta))^T$  be the gradient in the constrained domain and  $\tilde{H}$  the  $m \times m$  Hessian of the constrained optimisation, given by

$$\tilde{H}_{K,l}(\beta) = \frac{\partial^2 \ell(\beta)}{\partial \beta_{J_k} \partial \beta_{J_l}}, k, l = 1, \dots, m. \quad (7.9)$$

A step of the Newton Raphson algorithm in the current subdomain would propose:

$$\tilde{\beta}^{i+1} = \tilde{\beta}^i - [\tilde{H}(\beta^i)]^{-1} \tilde{g}(\beta^i).$$

Then the Newton Raphson algorithm is implemented as follows:

- Start with some  $\beta^0$ .
- For steps  $i = 1, 2, \dots$  of the algorithm, iterate this until convergence:

$$\beta^{i+1} = \begin{cases} \beta^i + t_{edge}g(\beta^i) & \text{if } t_{opt} \geq t_{edge}, \\ \beta_{NR}^{i+1} & \text{if } t_{opt} < t_{edge} \text{ and } \text{sign}(\beta_{NR}^{i+1}) = \text{sign}(\beta_+^i), \\ \beta^i + t_{opt}g(\beta^i) & \text{otherwise.} \end{cases}$$

### 7.2.2 Ridge-Cox regression

We also proposed the use of ridge-Cox regression model in order to estimate the AUC using 632+ and compare it to LASSO to estimate the 632+ AUC as proposed by [50]. Ridge

regression scales all the coefficients towards 0, but sets none to exactly zero. Ridge regression better handles correlated predictors and achieves a stable fit even in the presence of strongly correlated predictors, shrinking each coefficient based on the variation of the corresponding variable. If two predictors are very correlated, ridge regression will tend to give them equal weight. Starting from the Cox model, the estimated coefficients can be constrained to satisfy the condition  $\beta'\beta < C$  for some choice of  $C$ . This is known as ridge penalty which is only applied to linear and generalised linear models. To control the estimates, a penalty term with an appropriate weight  $w$  is subtracted from the log-likelihood, which now becomes the penalized partial likelihood:

$$\ell_{pen}^{ridge}(\beta) = \sum_{i=1}^D \left[ X_i\beta - \sum_{j \in R(t_i)} \log(\exp(X_j\beta)) \right] - \frac{1}{2}w\beta'\beta. \quad (7.10)$$

The penalized likelihood is maximised by taking the scores and using the Newton Raphson method. At the end of the procedure, the Breslow estimator can be used to obtain an estimate for the baseline hazard [139]. The authors of [145] suggested the use of the full likelihood to fit the same model. The authors of [151] used an iterative weighted least squares procedure based on the result that the partial likelihood is equivalent to the likelihood function of independently sampled poisson random variables. The same procedure is used in [139] for fitting the LASSO method.

---

### **7.3. Resampling methods**

---

In this section we discuss the use of *cross-validation* and different *bootstrap methods* for estimating AUC(t). Also the proposed method for variance estimation is described.

The bootstrap family was introduced by Efron and is fully described in Efron and Tibshirani [138]. Efron [38] introduced the bootstrap, double bootstrap and the 632 estimator (all variations on the bootstrap) and compared them to the leave-one-out estimate using a variety of

small sample simulations with Gaussian features.

### 7.3.1 632 bootstrap

The 632+ bootstrap was proposed by Efron and Tibshirani [138] in order to reduce the upward bias of the leave-one-out bootstrap estimator. A bootstrap 632 estimator has been shown to be superior to leave-one-out estimator in a variety of situations with small training sample sizes [138].

*Bootstrap cross-validated* FN and FP fractions are obtained by:

$$FNF^{BCV}(c, t) = \frac{1}{B} \sum_{b=1}^B FNF^b(c, t), \quad (7.11)$$

$$FPF^{BCV}(c, t) = \frac{1}{B} \sum_{b=1}^B FPF^b(c, t). \quad (7.12)$$

The corresponding ROC curve is  $FPF^{BCV}(c, t), 1 - FNF^{BCV}(c, t), c \in R$  and  $AUC^{BCV}(c, t)$  is the corresponding bootstrap cross-validation estimation of the AUC at time  $t$ . For a large sample size ( $n > 40$ ), the probability that the individual appears in the training set is approximately equal to  $1 - (1 - 1/n)^n \approx 0.632$ . The proportion of 0.368 is composed of completely independent data from the replicated data included in the training set, which causes an underestimation of the prognostic capacity. Efron [38] proposed the **632 estimator** to correct this underestimation:

$$FNF^{632}(c, t) = 0.368 \overline{FNF(c, t)} + 0.632 FNF^{BCV}(c, t), \quad (7.13)$$

$$FPF^{632}(c, t) = 0.368 \overline{FPF(c, t)} + 0.632 FPF^{BCV}(c, t), \quad (7.14)$$

where  $\overline{FNF(c, t)}$  and  $\overline{FPF(c, t)}$  are the apparent rates, calculated using the  $B$  training sets. More precisely, the apparent FNF and FPF can be calculated using only the data included in

the bootstrap sample

$$\overline{FNF(c, t)} = \frac{1}{B} \sum_{b=1}^B \overline{FNF^b(c, t)}, \quad (7.15)$$

$$\overline{FPF(c, t)} = \frac{1}{B} \sum_{b=1}^B \overline{FPF^b(c, t)}. \quad (7.16)$$

The 632 ROC curve is  $FPF^{632}(c, t), 1 - FNF^{632}(c, t), c \in R$  and  $AUC^{632}(c, t)$  is the corresponding 632 estimation of the AUC at time  $t$ .

### 7.3.2 632+ bootstrap

The 632 rates may be associated with over estimations if the apparent estimations are very small when over-fitting data. Efron and Tibshirani [38, 138] improved the correction with the 632+ estimator [50]. The no-information rates associated with FNF and FPF may be estimated using all the data and considering the independence between  $Y$  and  $T$ :  $\gamma_N(c, t) = 1 - \gamma_P(c, t) = \hat{F}(c)$ .

These no-information probabilities are used to define the over-fitting rates:

$$r_N(c) = \frac{FNF^{BCV}(c, t) - \overline{FNF(c, t)}}{\gamma_N(c, t) - \overline{FNF(c, t)}}, \quad (7.17)$$

$$r_P(c) = \frac{FPF^{BCV}(c, t) - \overline{FPF(c, t)}}{\gamma_P(c, t) - \overline{FPF(c, t)}}. \quad (7.18)$$

The authors of [50] assigned these rates to 0 for negative values and to 1 for values higher than

1. The 632+ estimations of the false negative and positive rates are thus defined by:

$$FNF^{632+}(c, t) = [1 - \varphi(\gamma_N(c, t))] \overline{FNF(c, t)} + \varphi(\gamma_N(c, t)) FNF^{BCV}(c, t), \quad (7.19)$$

$$FPF^{632+}(c, t) = [1 - \varphi(\gamma_P(c, t))] \overline{FPF(c, t)} + \varphi(\gamma_P(c, t)) FPF^{BCV}(c, t), \quad (7.20)$$

where  $\varphi(f) = \frac{0.632}{1-0.368f}$ .

The corresponding 632+ ROC curve is  $FPF^{632+}(c, t), 1 - FNF^{632+}(c, t), c \in R$  and  $AUC^{632+}(c, t)$  is the corresponding 632+ estimation of the AUC at time  $t$  [50].

### 7.3.3 Estimation of standard errors

Estimation of the time dependent AUC together with its standard error are very important to help in computing the confidence intervals. We can then use this to decide if the diagnostic index has ability to discriminate between two populations or not. We applied the two-level bootstrapping, where we performed  $B$  bootstrap samples from which  $B$   $AUC(t)$ 's are computed by any of the re-sampling methods discussed in Subsections 7.3.1 and 7.3.2. Then confidence interval will be the 2.5<sup>th</sup> and the 97.5<sup>th</sup> percentiles of the  $B$   $AUC(t)$  estimates at time  $t$ . In this way we can clearly declare whether the associated predictor index has the ability to distinguish or not. In this method, estimates  $\theta_b^*$ ,  $b = 1, 2, \dots, B$  of the parameter of interest  $\theta$  are calculated from  $B$  pseudo samples. Then an estimate of the bootstrap variance of the parameter of interest is calculated as:

$$V_B(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}_{\cdot}^*)^2, \quad (7.21)$$

where  $B$  is the number of replicate samples and  $\hat{\theta}_{\cdot}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*$ . Note that in the two-level bootstrapping approach the  $\hat{\theta}_b^*$  is estimated from another bootstrap sample as in previous function. We wrote a specific R program for this function.

---

## 7.4. Algorithm

---

The following is an algorithm to estimate the  $AUC(t)$  using different methods of imputation, variable selection and resampling.

- Step 1: Considering missing values. In the algorithm we use nearest neighbor hot deck imputation method.
- Step 2: Evaluation of diagnostic performance of biomarkers by estimating the AUC. The AUC captures the overall diagnostic accuracy of the combined biomarker test.
- Step 3: This important step needs to be undertaken to optimise the set of biomarkers with high and independent diagnostic information content in a multivariate setting. There are a number of critical issues to consider in this step: These are choosing an appropriate statistical method for multivariate analysis, choosing the number of diagnostically informative biomarkers to be entered into the multivariate model and using an appropriate method to optimise the number of finally selected biomarkers. The Cox proportional hazards model was used for the regression of hazard of experiencing an event of interest according to quantitative factors. LASSO penalty was used as it is very reliable; it is a variable selection method and results in considerably better prediction performance. We also suggested the ridge-Cox regression model to estimate the  $AUC(t)$  as a comparative competing model to LASSO-Cox penalty.
- Before applying the algorithm to derive a discriminatory rule and in order to avoid the over-fitting problem associated with model selection, we split each original data set into a training set and a validation set. This split was done using ordinary bootstrap, 632 and 632+ estimators of time dependent AUC and their variances as we mentioned before. We used “boot.ROct” function in “ROC632” package for estimating  $AUC(t)$  using various resampling methods. Bootstrap and cross-validated  $AUC(t)$  estimates were estimated using our own updated R code.

## 7.5. Simulation studies

---

Simulation studies represent an important statistical tool to investigate the performance, properties and adequacy of statistical models, test statistics and estimation techniques considering pre-specified conditions. In this work we apply the proposed two level bootstrapping method with respect to estimating the AUC variance. In addition we also used simulation studies to compare between different resampling methods to estimate  $AUC(t)$ .

We simulate datasets under the following group settings: Assume that there are  $K$  diagnostic tests (corresponding to  $K$  features)  $Y_1, Y_2, \dots, Y_K$ . In our case, we let  $K = 5$ , that is five features  $Y_1, Y_2, Y_3, Y_4$  and  $Y_5$ . Denote the mean vector of the  $K$  features by  $\mu_k$ .

With the above settings, the features outcomes  $y_{ik}$  is given by

$$y_{ik} = \mu_k + a_i + \varepsilon_{ik},$$

where the notation and assumptions in our case are

- $n$  (resp.  $m$ ) is the number of individuals and  $i$  is the index for the set  $\{1, 2, \dots, n\}$ ,
- $k$  is the index for a feature,
- $a_i$  is the subject specific random effect which is assumed to follow the normal distribution that is  $a_i^D \sim N(0, 0.5)$  and
- $\varepsilon_{ik}$  is the random error effect also assumed to follow the normal distribution  $\varepsilon_{ik} \sim N(0, 0.25)$ .

The outcome vector  $y$  are respectively generated using three model assumptions from the normal distribution with mean given by  $\mu_k = (1, 0.5, 0.25, 0, 0)$  for three different covariance



structures given by:

- For the first setting (Model 1) we assume independence among the five biomarkers or features.
- For the second setting (Model 2) we assume the same dependence across all the features. The resulting covariance structure is the exchangeable or compound symmetry.
- For the third setting (Model 3) we remove dependence for the first three features.

We simulated data sets with sample size  $N = 50$  and bootstrap replicates  $B = 200$  for both levels of bootstrapping. The time to event was simulated from the exponential distribution and the censoring time was generated independently such that we fix the censoring rates to 30%.

Table 7.1 shows a summary of different types of quantities which were estimated from the analysis. These include the AUC and its standard deviation based on different methods using apparent, bootstrap cross-validation, 632 bootstrap and 632+ bootstrap, which is given in column 3. Column 4 includes standard error for AUC using our proposed two level bootstrapping. Columns 5 and 6 include the lower and upper confidence limits of the AUC based on bootstrap standard errors. Proportion of times lower confidence limits of AUC excludes 0.5, are listed in column 7 for different estimation methods within a given model assumption. Finally, the coverage probabilities that the CIs for different resampling methods include the true AUC are listed in column 8.

From Table 7.1 we can see that:

For Model 1, the apparent AUC equals to 0.848 while the bootstrap cross-validated one equals 0.858. The values of 632 and 632+ AUC are almost the same. All estimated AUC's are very

---

**Chapter 7 – Use of resampling methods to predict the outcome of tuberculous meningitis in high HIV prevalence patients in South Africa**

---

Table 7.1: Mean of the time dependent AUC at 6 months obtained from 250 simulated samples for each scenario B =200, bootstrap replications are performed for computing the apparent, BCV, 632, 632+ and 2-level bootstrap standard error(SE) for the AUCs

Model	Estimation method	AUC(STD)	SE	Lower limit	Upper limit	Proportion of lower limit excludes 0.5	Coverage probabilities
Model1	Apparent	0.848(0.0661)	0.053	0.744	0.952	0.985	
	BCV	0.858(0.044)	0.037	0.785	0.931	0.985	0.854
	632	0.855(0.047)	0.036	0.784	0.926	0.995	0.968
	632+	0.855(0.047)	0.0.04	0.777	0.933	0.991	0.911
Model2	Apparent	0.812(0.064)	0.064	0.687	0.937	0.975	
	BCV	0.823(0.057)	0.055	0.714	0.930	0.985	0.950
	632	0.819(0.055)	0.054	0.712	0.924	0.997	0.995
	632+	0.821(0.057)	0.0.056	0.731	0.931	0.995	0.955
Model3	Apparent	0.809(0.056)	0.064	0.684	0.934	0.980	
	BCV	0.82(0.055)	0.054	0.714	0.926	0.985	0.950
	632	0.816(0.050)	0.055	0.708	0.924	0.995	0.990
	632+	0.818(0.055)	0.0.054	0.712	0.924	0.975	0.935

close to each other. Based on the CI's, which do not include 0.5, for all AUCs from resampling methods, we deduce that the combined biomarker yields a significant discriminatory ability. In all scenarios, the BCV, the 632 and the 632+ estimates were similar to the true apparent estimations. This conclusion can be made regardless of the simulation models.

In addition to comparing AUC values from different resampling methods, it is more important to evaluate the performance of the score index or diagnostic tests in each resampling method. Thus we obtained variances and confidence intervals for AUC estimations. We also investigated the level of discrimination of the two level bootstrapping methods by firstly looking at how often the lower limits exclude an  $AUC = 0.5$ . Secondly, we looked at how often the CIs for AUC estimators (using resampling methods) include the true AUC. For example, the proportion of

times lower limits of CI's for 632 AUC that exclude 0.5 is 0.995 under Model 1. The coverage probabilities (under Model 1) for AUC using 632 show that 242 out of 250 CI's include the true AUC values, while 228 out of 250 CI's, using 632+, include the true AUC values. The coverage probabilities for 632 estimator were 0.968, 0.995 and 0.990 for Model 1, Model 2 and Model 3 respectively. This indicates that the two level bootstrap methods with 632 estimators perform better compared to BCV and 632+ estimators. Two level bootstrap methods for all scenarios perform very well in terms of proportion and coverage probabilities. However this solution is time consuming and we need to investigate it with more simulations.

From Table 7.1, BCV, 632 and 632+ yielded significant diagnostic AUC results. We also found that there is no statistical difference for resampling methods and true value which indicates these estimators are nearly unbiased. Moreover we obtained the two level estimations of AUCs with the same conclusion. We conclude by remarking that using Model 2 (with correlation) is preferable since it yielded the highest coverage probabilities and proportion of times lower limits of CI's for AUC that exclude 0.5.

We are also interested in comparing the use of LASSO estimation of AUC [50] and our proposed method of using ridge-Cox regression to estimate the AUC. We simulated data sets with sample size  $N=50$  and bootstrap replicates  $B = 200$ . Censoring rates were fixed to 10% and 30%. We distinguished the following two scenarios:

First, among five biomarkers three are associated with the time to event,  $\beta = (\log(1.2), -\log(1.2), 0, \log(2), 0)$ . Indeed, there is no over-fitting because only five biomarkers are analysed by using at least 50 individuals.

The second scenario involves 100 biomarkers, no biomarker is associated with the time to event,  $\beta = (0, 0, \dots, 0)$ . It is clear that this case leads to severe over-fitting as the number

of biomarkers is larger than the number of individuals and there is no association between biomarkers and time to event.

Table 7.2: Mean of the time dependent AUC at 6 months obtained from 250 simulated samples for each combination of over-fitting level, censoring rate and penalized model, B =200 bootstrap replications are performed for computing the apparent, BCV, 632 and 632+ estimations

Missing rate	Model	Method	Apparent	BCV	632	632+
10%	model 1 (5 variable)	LASSO	0.808	0.815	0.813	0.813
		Ridge	0.809	0.825	0.819	0.821
	100 variables (over-fitting)	LASSO	0.829	0.834	0.832	0.832
		Ridge	0.839	0.843	0.842	0.844
30%	model 2 (5 variable)	LASSO	0.833	0.832	0.832	0.831
		Ridge	0.846	0.853	0.850	0.852
	100 variables (over-fitting)	LASSO	0.834	0.827	0.829	0.828
		Ridge	0.844	0.848	0.847	0.848

From Table 7.2 and - when the censoring rate is 10%, - LASSO-Cox regression and ridge-Cox regression obtained very closed AUC(t) values using the two scenarios. When the censoring rate is 30%, it is clear that our proposed estimator using ridge-Cox regression model from the first scenario resulted in a slightly higher estimation of AUC(t). As for the LASSO method, the 632+ estimator was 0.831, whereas for ridge-Cox it was 0.852. Both methods performed very well as their estimators of AUC using 632+ are very close to each other. The ridge-Cox regression model and LASSO-Cox regression model revealed similar AUC results under the second scenario. As for LASSO-Cox method the 632+ estimator was 0.828 whereas for ridge-Cox regression it was 0.848. Thus using ridge-Cox to estimate 632+ bootstrap AUC(t)

resulted in a slightly higher estimations than the LASSO-Cox method.

In the first scenario (no over-fitting) with 10% censoring rate, both ridge-Cox and LASSO-Cox regressions obtained similar AUC(t) estimates. However the 632 estimator appeared to give AUCs more close to the true AUC (apparent). The same conclusion can be made with 30% censoring rate.

---

## **7.6. Application of tuberculous meningitis (TBM) in high HIV prevalence**

---

This section is an application to TBM/HIV data set. The first subsection describes tuberculous meningitis HIV data. The next subsection contains the statistical analysis, results and discussion.

### **7.6.1 The TBM/HIV description**

*Mycobacterium tuberculosis* is a common, devastating cause of meningitis in HIV-infected persons. Meningitis causes significant mortality and morbidity in HIV infected persons. Tuberculous meningitis (TBM) accounts for a substantial proportion of cases, particularly in high tuberculosis (TB) prevalence areas. Due to international rollout programs, access to antiretroviral therapy (ART) is increasing globally. Starting ART during TB treatment is associated with reduced mortality in TB/HIV co-infected patients [1, 28]. However, few studies have reported the influence of ART on the outcome of patients with HIV-associated TBM [26, 140]. The data of the study was collected at GF Jooste Hospital in Cape Town, South Africa. This hospital serves a high density low income community. It is a 200-bed public sector referral hospital that serves adult patients from a community of approximately 1.3 million people.

This predominantly low income, high density population is at the epicenter of the TB/HIV

pandemic. In some parts of the referral area the reported TB case notification rate exceeds 1500 cases per 100 000 people per year and the HIV seroprevalence at antenatal clinics reaches 30% [89]. All patients accessing public sector care facilities with suspected meningitis are referred to GF Jooste Hospital for investigations, including a lumbar puncture (LP). Adult patients (18+ years) who had a LP performed over a six-month period (1 March 2009–31 August 2009) were identified from laboratory logs and included in the study. Definite, probable and possible TBM were diagnosed according to published case definitions [140].

### **7.6.2 Statistical analysis, results and discussion**

The demographic, clinical and investigative findings for patients with definite, probable and possible TBM are detailed in Table 7.3. The majority of TBM cases (68%) presented with advanced TBM disease (Stage 2 or 3). The percentage of patients who were receiving TB treatment at the time of presentation was 23% (26/115). Median age for definite probable and possible patients were 35, 36 and 38 respectively. The percentage of females who were diagnosed as definite was of 47%. The percentage of patients who were HIV positive - of which 41% of them were definite TBM - was 88% (106/120). For HIV status a typical cerebrospinal fluid (CSF) findings in patients with definite TBM ( $n = 47$ ) included, a polymorphonuclear cell predominance (median = 6.50,  $SE = 9.43$  and  $CI = (12.26, 50.26)$ ), a glucose level with median 1.59 mmol/L.

The percentage of patients with probable TBM who received corticosteroids was 70%, which is similar to that of the definite TBM group. The percentage of overall inpatient mortality amongst patients who were hospitalized (for four days after LP) was 38% (45 out of 120 patients). The percentage of patients (discharged from hospital) who were HIV-infected and not on ART at time of presentation was 57% (31 out of 54 patients). For those patients who

---

## Chapter 7 – Use of resampling methods to predict the outcome of tuberculous meningitis in high HIV prevalence patients in South Africa

---

Table 7.3: The demographic, clinical and investigative findings for patients with definite, probable and possible TBM

	Definite TBM ( $n = 47$ )	Probable TBM ( $n = 35$ )	Possible TBM ( $n = 38$ )
Age median SE (CI)	35.00 1.54 (32.98, 39.19)	36.00 2.4 (34.35,44.11)	38.00 1.65 (33.73,40.43)
Female, $n/N$ (%)	22/47 (47)	16/35 (46)	22/38 (58)
HIV status, $n/N$ (%)			
Infected	43/47 (91)	27/35 (77)	36/38 (95)
Uninfected	2/47 (4)	5/35 (14)	1/38 (3)
Unknown	2/47 (4)	3/35 (9)	1/38 (3)
CD4+ median SE (CI)	63 10.22 (57.19,98.94)	103.00 21.49 (79.33,168.26)	109 17.12 (91.83,161.43)
On ART , $n/N$ (%)	2 9/41 (22)	6/27 (22)	5/35 (14)
Previous TB, $n/N$ (%)	15/43 (35)	7/34 (21)	12/38 (32)
On TB treatment, $n/N$ (%)	9/43 (21)	8/34 (24)	9/38 (24)
BMRC TBM Disease Grade			
Stage1	10/42 (24)	7/34 (21)	16/38 (42)
Stage2	9/42 (69)	23/34 (68)	20/38 (53)
Stage3	3/42 (7)	4/34 (12)	2/38 (5)
Hemoglobin median SE (CI)	10.45 0.36 (10.09, 11.53)	12.0 0.38 (11.13,12.68)	10.000 10 .39 (9.14,10.72)
Wcc	6.00 0.63 (6.04,8.57)	5.60 1.00 (4.95,8.98)	7.65 0.68 (6.77, 9.53)
Sodium (CI)	126.0 0.79 (124.51,127.71)	129.0 1.00 (127.46,131.54)	130.0 1.05 (128.79,133.05)
Protein (CI)	2.60 2.4 (2.50,12.16)	2.41 2.59 (2.3, 12.82)	1.21 0.15 ( 1.19, 1.81)
Glucose (CI)	1.59 .14 ( 1.36, 1.92)	1.87 .20 (1.6, 2.39)	2.66 .17 (2.41,3.09)
Lymphocytes (CI)	77.0 26.39 (93.02, 199.33)	58.50 27.69 (61.78,174.45)	12.00 21.56 (.85, 88.47)
Polymorphonuclear (CI)	6.50 9.43 (12.26, 50.26)	0.0 9.83 (4.5 ,44.50)	0.000 2.86 (.34, 11.95)

initiated ART during the six months of TB treatment at six month follow-up, the percentage of TBM patients who had died was 48% while 10% were lost to the follow-up process.

Table 7.4 describes the management and outcome in patients with TBM.

Table 7.5 shows factors analysed for association with inpatient mortality for all patients ( $n=120$ ) in univariate and multivariate analysis. A higher BMRC TBM disease stage (2 or 3 versus

Table 7.4: Management and outcome in patients with TBM

Outcome	
History of previous TB	No 81/115 (70%) yes 34
TB treatment On treatment at time of presentation	yes 26/115 (23%) No 89
Corticosteroids started	yes 64/113 (57%) No 48
on ART	No 81/100(81%) yes 19
ART started 6 months after starting TB treatment	Yes 31/54(57%) No 89
Inpatient mortality	yes 45/120 (38%) No 75

1) remained predictive of mortality in multivariate analysis.



Table 7.5: Univariate and multivariate analyses of association between variables and in hospital mortality in all patients

variables	Univariate analysis				Multivariate analysis			
	log odds	SE	z-value	p-value	log odds	SE	z-value	p-value
age	0.00232	0.01626	0.143	0.887	-2.640e-03	2.899e-02	-0.091	0.9275
sex	0.2136	0.3778	0.565	0.572	6.392e-01	5.859e-01	1.091	0.2753
history of previous TB	0.05106	0.42129	0.121	0.9035	3.156e-01	5.468e-01	0.577	0.5638
on TB treatment	-0.3773	0.4771	-0.791	0.4291	-2.438e-01	6.205e-01	-0.393	0.6944
HIV status	1.405	1.088	1.291	0.1966	1.742e+01	1.499e+03	0.012	0.9907
wcc	-0.008517	0.041255	-0.206	0.836	3.634e-04	4.995e-02	0.007	0.9942
TBM stage	0.9933	0.3774	2.632	0.00849 **	1.149e+00	4.731e-01	2.428	0.0152*
Diagnosis of TBM	0.1201	0.2250	0.534	0.593	4.874e-01	3.854e-01	1.265	0.2060
sodium	-0.007868	0.031404	-0.251	0.802	-1.260e-02	4.793e-02	-0.263	0.7926
hemoglobin	-0.11207	0.08041	-1.394	0.163	-2.155e-01	1.298e-01	-1.661	0.0968 .
glucose	-0.22698	0.17433	-1.302	0.193	-1.308e-01	2.674e-01	-0.489	0.6248
protein	0.03824	0.01896	2.017	0.043733 *	4.338e-02	2.443e-02	1.776	0.0757 .
steroids	-0.4654	0.3944	-1.180	0.238	-3.162e-02	5.691e-01	-0.056	0.9557
lymp	0.0005222	0.0011656	0.448	0.65418	2.023e-04	1.787e-03	0.113	0.9098
poly	-0.0003397	0.0037591	-0.09	0.92799	-5.665e-04	5.281e-03	-0.107	0.9146

Table 7.6 shows factors analysed for association with inpatient mortality for HIV-infected patients ( $n = 106$ ) only. In a univariate analysis, CD4 count ( $p$ -value=0.0296) and a higher BMRC TBM disease stage ( $p$ -value=0.00952) remained predictive of mortality same as was with a multivariate analysis.

Table 7.6: Univariate and multivariate analyses for association association with inpatient mortality for HIV-infected patients

variables	Univariate analysis				Multivariate analysis			
	log odds	SE	z-value	p-value	log odds	SE	z-value	p-value
age	0.004371	0.020929	0.209	0.835	0.0164032	0.0325116	0.505	0.61389
sex	0.3216	0.4043	0.795	0.426	0.6754317	0.7250326	0.932	0.35155
history of previous TB	0.2491	0.4398	0.566	0.5711	0.1452846	0.7060201	0.206	0.83696
on TB treatment	-0.3496	0.4866	-0.718	0.4725	-0.1051131	0.7219156	-0.146	0.88423
CD4	-0.006548	0.003010	-2.175	0.0296 *	-0.0113005	0.0048680	-2.321	0.02027 *
wcc	0.001282	0.041127	0.031	0.975	-0.0175340	0.0640863	-0.274	0.78439
TBM stage	1.0962	0.4228	2.593	0.00952 **	1.6259955	0.5916534	2.748	0.00599 **
Diagnosis of TBM	0.1959	0.2346	0.835	0.4036	0.7349338	0.5041899	1.458	0.14494
sodium	0.001597	0.033281	0.048	0.962	-0.0069866	0.0579301	-0.121	0.90401
hemoglobin	-0.15027	0.09044	-1.662	0.0966 .	-0.2401738	0.1615598	-1.487	0.13712
glucose	-0.21973	0.18265	-1.203	0.229	0.1665182	0.3105304	0.536	0.59179
protein	0.02614	0.01803	1.450	0.14715	0.0320985	0.0358383	0.896	0.37044
steroids	-0.2142	0.4215	-0.508	0.611	0.4598135	0.7000183	0.657	0.51127
lymp	0.0005235	0.0011795	0.444	0.65716	0.0010891	0.0023793	0.458	0.64715
poly	0.002032	0.004033	0.504	0.61433	-0.0007876	0.0058973	-0.134	0.89376
on ART	0.3216	0.5182	0.621	0.53491	0.6067015	0.9056628	0.670	0.50292

Table 7.7 shows factors analysed for association with mortality for all patients ( $n=120$ ) in univariate and multivariate analysis. A higher BMRC TBM disease stage (2 or 3 versus 1) ( $p$ -value=0.0417) and protein ( $p$ -value=0.00585) remained predictive of mortality in the multivariate analysis.

Table 7.7: Univariate and multivariate analyses for association with survival for all patients

variables	Univariate analysis					Multivariate analysis				
	coef	exp(coef)	se(coef)	z	p-value	coef	exp(coef)	se(coef)	z	p-value
age	0.006935	1.006959	0.011229	0.618	0.537	0.002587	1.002590	0.015906	0.163	0.87082
sex	0.06339	1.06544	0.26053	0.243	0.808	0.205751	1.228447	0.368987	0.558	0.57711
history of previous TB	0.2538	1.2889	0.2777	0.914	0.361	0.582031	1.789670	0.334833	1.738	0.08216 .
on TB treatment	-0.1689	0.8446	0.3250	-0.52	0.603	0.199178	1.220399	0.389308	0.512	0.60892
HIV status	0.7557	2.1290	0.7207	1.049	0.294	1.792391	6.003792	1.095818	1.636	0.10191
wcc	-0.01360	0.98649	0.02839	-0.479	0.632	-0.036236	0.964412	0.035847	-1.011	0.31208
TBM stage	0.4841	1.6226	0.2376	2.037	0.0417 *	0.631159	1.879788	0.307157	2.055	0.03989 *
Diagnosis of TBM	0.1941	1.2143	0.1571	1.236	0.217	0.474783	1.607666	0.240215	1.976	0.04810 *
sodium	0.003314	1.003319	0.021782	0.152	0.879	0.004072	1.004081	0.028114	0.145	0.88483
hemoglobin	-0.1017	0.9033	0.0532	-1.911	0.0561 .	-0.071604	0.930900	0.081104	-0.883	0.37731
glucose	-0.1735	0.8407	0.1238	-1.402	0.161	-0.303545	0.738197	0.186231	-1.630	0.10311
protein	0.021406	1.021637	0.007766	2.756	0.00585 **	0.030245	1.030707	0.010497	2.881	0.00396 **
steroids	-0.5331	0.5868	0.2676	-1.992	0.0464 *	-0.225463	0.798147	0.370992	-0.608	0.54337
lymp	-0.0002250	0.9997751	0.0008663	-0.26	0.795	-0.001287	0.998714	0.001097	-1.173	0.24083
poly	0.001040	1.001040	0.002649	0.393	0.695	0.004820	1.004831	0.003132	1.539	0.12388

Table 7.8 shows factors analysed for association with mortality for HIV-infected patients ( $n=106$ ) only and similar results to Table 7.7 were obtained - in which a higher BMRC TBM disease stage and protein remained predictive of mortality.

Analysis of factors associated with six-month mortality is reported only for HIV-infected hospital survivors for whom the outcome was known at the six-months follow-up. Patients on ART at presentation, or having started ART during TB treatment, were negatively associated with six-month mortality.

Table 7.8: Univariate and multivariate analyses for association with survival for HIV-infected patients

	Univariate analysis					Multivariate analysis				
variables	coef	exp( <i>coef</i> )	se(coef)	z	p-value	coef	exp( <i>coef</i> )	se(coef)	z	p-value
age	0.006683	1.006706	0.013820	0.484	0.629	0.008772	1.008810	0.017848	0.491	0.62310
sex	0.08642	1.09027	0.27827	0.311	0.756	0.049979	1.051249	0.402405	0.124	0.90116
history of previous TB	0.3211	1.3786	0.2915	1.102	0.271	0.609063	1.838708	0.429982	1.416	0.15663
on TB treatment	-0.1525	0.8585	0.3313	-0.46	0.645	0.287393	1.332948	0.434159	0.662	0.50800
CD4	-0.005054	0.994958	0.002106	-2.4	0.0164 *	-0.005964	0.994054	0.002554	-2.335	0.01954 *
wcc	-0.005956	0.994061	0.027419	-0.217	0.828	-0.063020	0.938925	0.044510	-1.416	0.15681
TBM stage	0.5520	1.7367	0.2755	2.004	0.0451 *	0.919506	2.508051	0.354254	2.596	0.00944 **
Diagnosis of TBM	0.2387	1.2696	0.1659	1.439	0.15	0.542467	1.720245	0.308557	1.758	0.07873 .
sodium	0.01416	1.01426	0.02257	0.628	0.53	0.002682	1.002686	0.031363	0.086	0.93185
hemoglobin	-0.11829	0.88844	0.05869	-2.015	0.0439 *	-0.019196	0.980987	0.093270	-0.206	0.83694
glucose	-0.1525	0.8585	0.1282	-1.19	0.234	-0.064214	0.937804	0.211158	-0.304	0.76105
protein	0.014671	1.014779	0.009707	1.511	0.131	0.033532	1.034101	0.026389	1.271	0.20384
steroids	-0.4281	0.6518	0.2860	-1.497	0.134	-0.099619	0.905183	0.417117	-0.239	0.81124
lymp	-0.0002163	0.9997837	0.0008753	-0.247	0.805	-0.001280	0.998721	0.001473	-0.869	0.38487
poly	0.003385	1.003391	0.002862	1.183	0.237	0.003414	1.003420	0.003725	0.916	0.35948
on ART	-0.1358	0.8730	0.3683	-0.369	0.712	-0.107428	0.898141	0.567485	-0.189	0.84985

Figure 7.1 is a Cox proportional hazard model showing survival curves of TBM/HIV patients On/Not ART and it is clear that the survival curves for patients on ART is higher than the curve for patients did not take or start ART, suggesting that the survival experience is possibly slightly better for patients on ART.

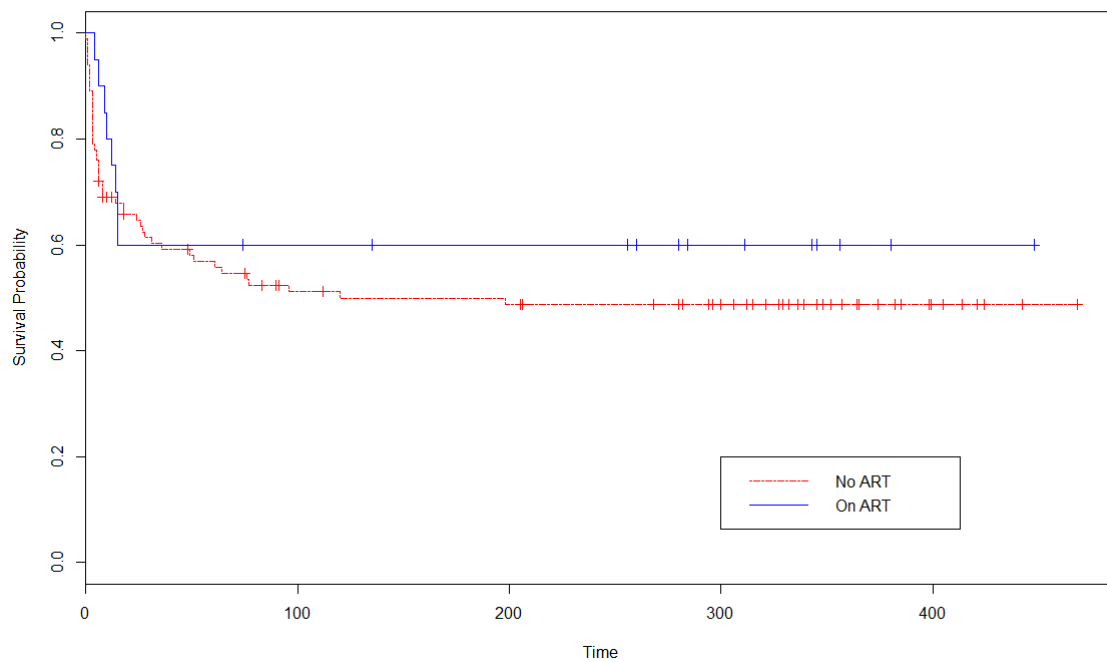


Figure 7.1: Survival curves for patients on ART and non ART for TBM/HIV dataset

Figure 7.2 is a Cox proportional hazard model showing survival curves of TBM stage for patients (stage 1, stage 2 and stage 3). As expected the resulting survival curves for stage 1 is higher than the curves for stage 2 and stage 3, suggesting that the survival experience is better for patients with stage 1.



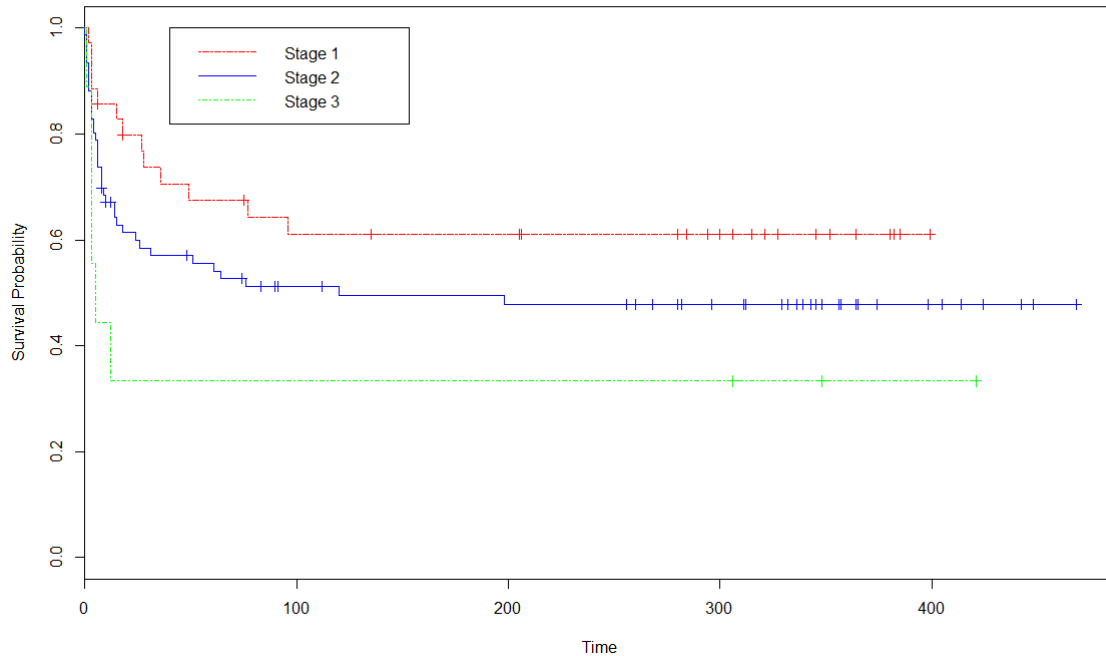


Figure 7.2: Survival curves for TBM stages for TBM/HIV dataset

Figure 7.3 is Cox proportional hazard model survival curves of TBM diagnosis as definite, probable and possible TBM. The survival curve for possible diagnosed subjects is lower than the curves for definite and probable patients, suggesting that the survival experience is worse for possible TBM patients.

From figure 7.4 we can see that the survival curve for negative HIV status is higher than the curve for positive HIV status, confirming that the survival experience for negative HIV status is better for positive status.

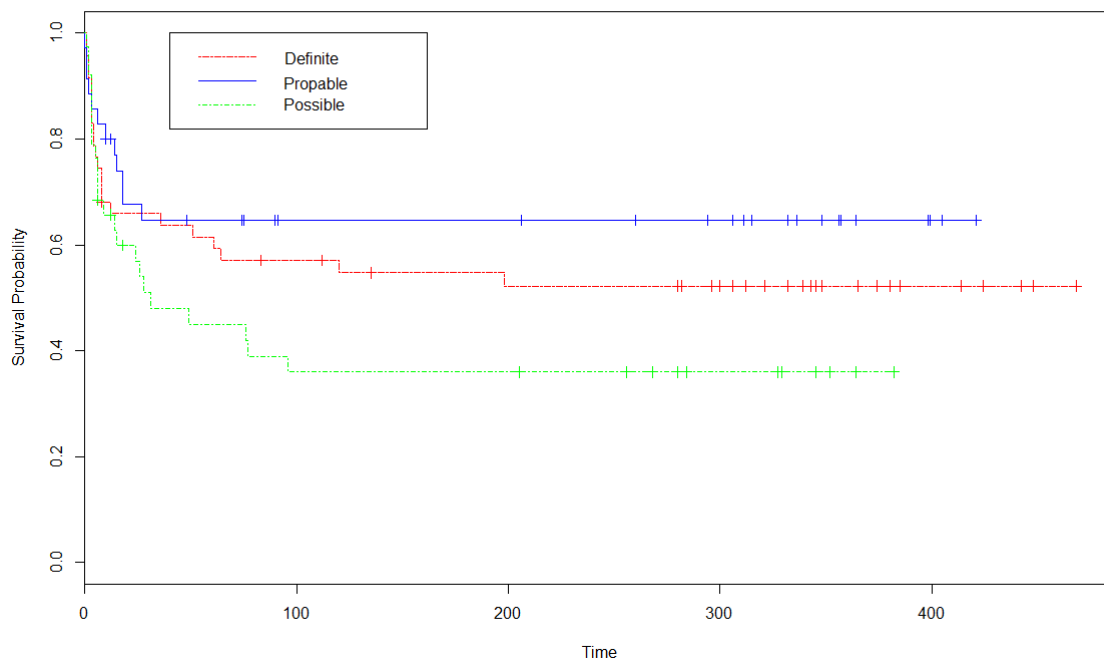


Figure 7.3: Survival curves for TBM diagnosis for TBM/HIV dataset

We are interested in investigating which variables are important and should be used to obtain better AUC. We used time dependent AUC estimation with penalized LASSO-Cox regression for variable selection. The use of the penalty LASSO-Cox regression model appears to be the best approach as it balances between prediction and interpretation [50]. We used the cross-validation, 632 and 632+ estimations of the ROC(t) and, the missing data were imputed according to the nearest neighbor hot deck strategy (see Subsection 4.2.2).

In addition to comparing AUC values from different resampling methods, it is more important to evaluate the performance of score index or diagnostic tests in each resampling method. Thus we obtained variances and confidence intervals for AUC estimations. From Table 7.9, CV, 632 and 632+ estimators obtained significant diagnostic test - which means that the index scores have an ability to discriminate between the TBM/HIV subjects who are likely to die before

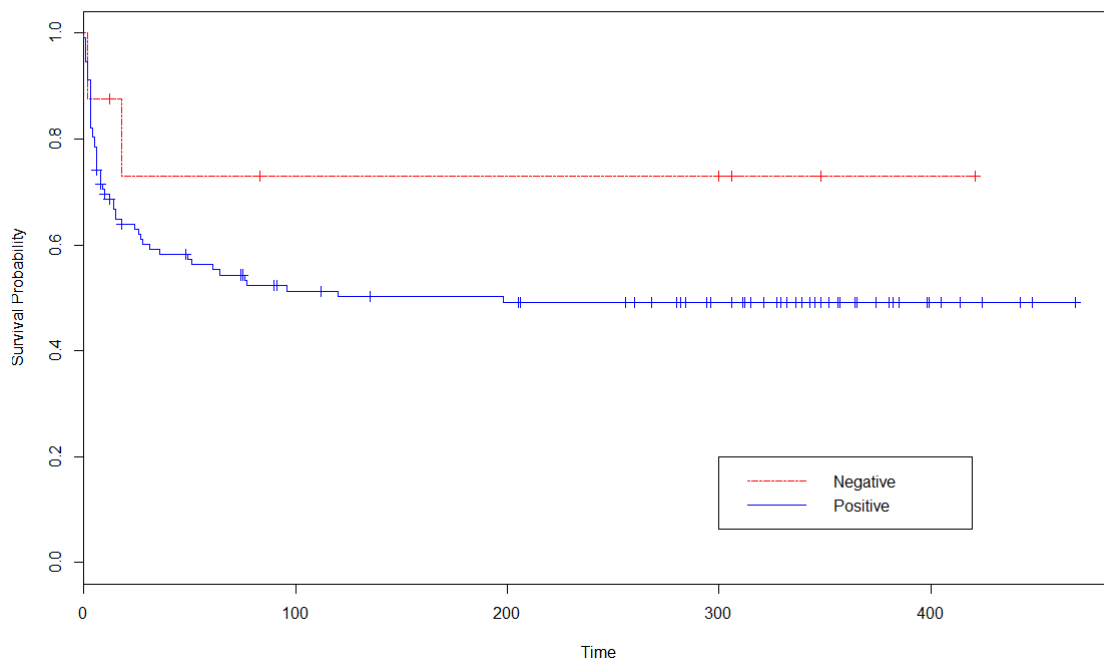


Figure 7.4: Survival curves for HIV status for TBM/HIV dataset

the first 6 months and those who may still alive beyond that time.

In Table 7.9, all the estimation methods yielded similar AUC values. The smallest among them is that estimated using the BCV method and the 632+, with respective values 0.846 and 0.849. The highest value was from the apparent method followed by the 632. The standard error was the smallest under the 632 resampling method. The BCV and apparent method yielded similar SEs and it was highest under the 632+ method.

We estimated the 632+  $AUC(t)$  as 0.85 for a prognosis up to 6 months. Thus a patient who will die before 6 months has a 85% chance of having a score higher than a patient who will be alive at this time. The AUC estimates from all methods are high which confirms the

Table 7.9: AUC values from different resampling methods for composite biomarker from TBM/HIV dataset

Methods	AUC	$SE$	Lower-limit	Upper-limit
<i>apparent</i>	0.858	0.076	0.709	0.998
<i>BCV</i>	0.846	0.075	0.774	0.993
632	0.850	0.069	0.715	0.985
632+	0.849	0.087	0.678	0.997

discriminatory capacity between patients who will die before six months and those who are alive after that time according to confidence intervals. The resulting  $p$ -values were strongly statistically significant and suggested that they all agree that the prognostic score has a high discriminatory capacity.

We propose the use of Ridge-Cox regression model to estimate the AUC as a competing method to LASSO-Cox regression. Both the LASSO ( $L_1$ ) and Ridge ( $L_2$ ) penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates [138]. The purpose of this shrinkage is to prevent over-fitting that arise from either collinearity of the covariates or high-dimensionality. Although both methods are shrinkage based methods, the effects of  $L_1$  and  $L_2$  penalization are quite different in practice. Applying an  $L_2$  penalty tends to result in all small but non-zero regression coefficients, whereas applying an  $L_1$  penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage.

The results are presented in Table 7.10, the AUCs obtained from LASSO using the 632+ estimator were 0.658 and 0.630 in 9 months and the first year respectively. The area decreased

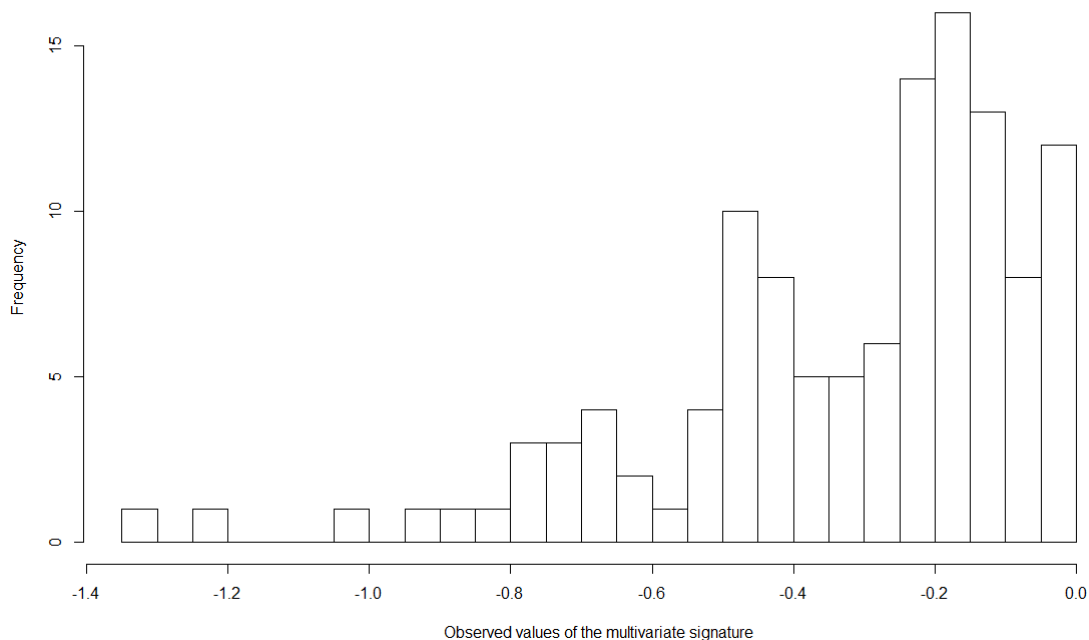


Figure 7.5: Observed values of multivariate signature for TBM/HIV dataset

Table 7.10: AUC estimations using LASSO and ridge methods for TBM/HIV dataset in different time points

method	time	BCV	632	632+
LASSO	270 days	0.653	0.661	0.658
	365 days	0.610	0.637	0.630
Ridge	270 days	0.744	0.777	0.769
	365 days	0.735	0.772	0.760

with prognostic time, illustrating that long-term failures are often more difficult to predict. This illustrates the utility of this signature to predict mortality over different times; however it explains that these signatures alone are not sufficient for medical decision making.

From Table 7.10, it is clear that our proposed estimator using ridge-Cox regression model resulted in higher AUC estimates. As for LASSO-Cox method the 632+ estimators in the first year was 0.630 whereas for ridge-Cox regression it was 0.774, means that the ridge-Cox regression appeared to be overoptimistic. From Figure 7.6, the over-fitting was high with an apparent AUC around 0.86 (using the ridge-Cox regression). In contrast, the prognostic capacity appeared to be underestimated when using the BCV estimator. The 632+ estimations were less optimistic than the 632 estimations.

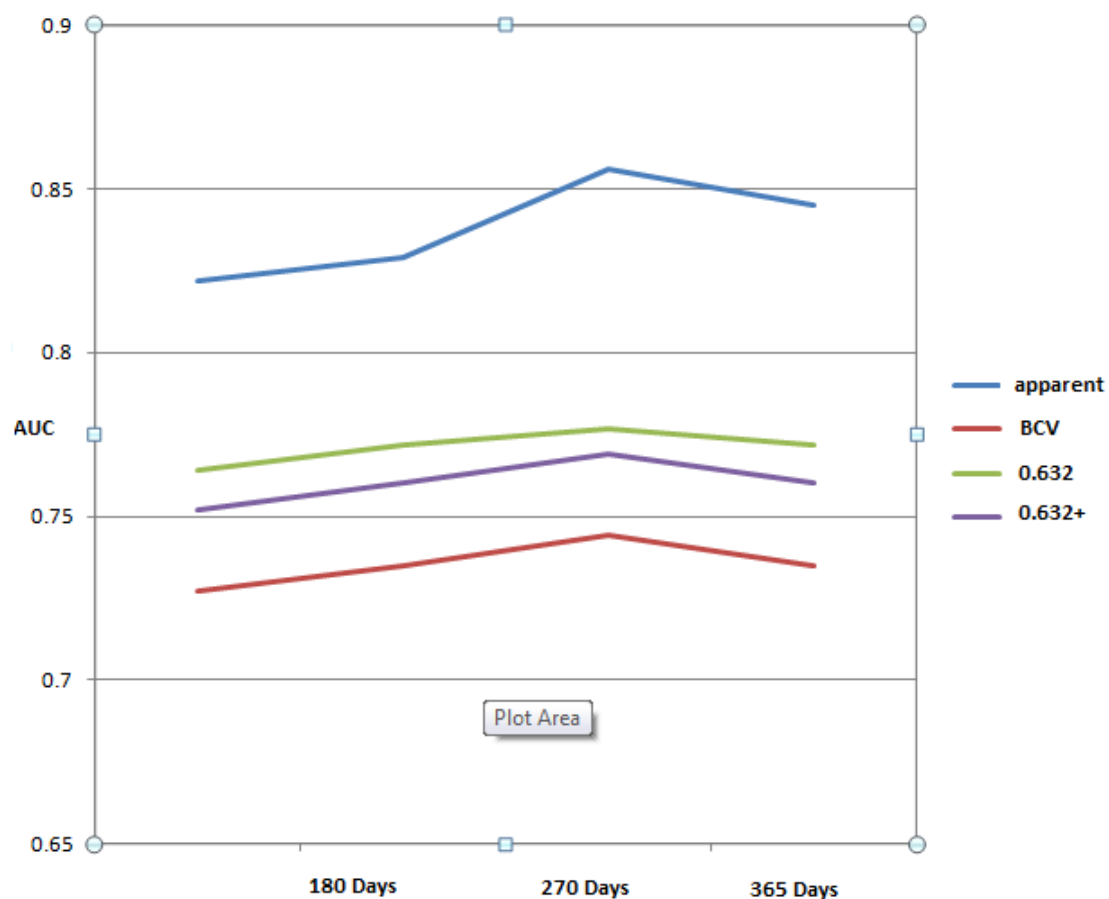


Figure 7.6: AUC according to the prognostic times and the different estimators using the ridge-Cox model regression for TBM/HIV dataset

---

## **7.7. Conclusion**

---

In this chapter we used the LASSO method variable selection method in order to estimate  $AUC(t)$ . The LASSO uses a penalty like ridge regression, except the penalty is the  $L_1$  norm of the coefficient vector, which causes the estimates of some coefficients to be exactly zero. This is in contrast to ridge regression which never sets coefficients to zero. The fact that the LASSO sets coefficients to zero can be a big advantage for the sake of interpretation - unlike ridge which tend to give higher values of  $AUC(t)$  but is difficult in terms of interpretation. It is mentioned in [50] that the use of the penalized LASSO-Cox regression appears to be the best approach as it balances between prediction and interpretation. However our proposed method of using ridge-Cox regression models appeared to perform similarly to LASSO-Cox regression in terms of  $AUC(t)$  estimations. Our simulations showed that two level bootstrap methods performed better with the 632 estimator in terms of confidence intervals. The application on TBM/HIV data showed that CD4 and TBM could significantly predict mortality. More specifically CD4 counts could strongly distinguish between patients who may die before six months and those who may survive thereafter.

# **The role of ambiguous nucleotides as biomarkers of recent HIV infection in rural KwaZulu-Natal, South Africa**

---

## **8.1. Introduction**

---

HIV infection is a global health problem as thousands of people are newly infected every year, therefore HIV infection has become one of the most common health problems in the world. However, it is a challenge to determine the difference between a recent infection and a chronic infection. It is well known that using ART for the management of HIV infected patients has been associated with reduction in morbidity and mortality. Estimation of the HIV incidence in populations is important for developing specific prevention strategies. HIV incidence is classically estimated by prospective cohort studies which are expensive and time consuming. In recent years, several methods based on viral sequences have been developed to identify recent HIV infection. Importantly, Kouyos et al. [76] showed that the proportion of ambiguous nucleotides is correlated with the time elapsed between HIV infection and sampling for genotyping. In a study in Switzerland, the ambiguous nucleotide method performed well using a



threshold of 0.5% as the cut off giving a high sensitivity of 86.8%, a reasonably high specificity of 70% and a high negative predictive value of 98.7%. However, it has yet to be determined if this method of measuring ambiguous nucleotides within a patient between sampling times could be used to determine recent infections in South Africa.

The aim of this chapter is to evaluate the use of the proportion of ambiguous base cells in HIV population sequences as a biomarker for use in an HIV incidence assay. To achieve this, we used samples from ART-naïve study patients from a South African HIV Study. We are interested in distinguishing HIV recent infection ( $\leq 36$  months) from long-term HIV infection ( $> 36$  months) using genetic data from a high risk HIV region in the province of KwaZulu-Natal in South Africa.

We analysed the proportion of ambiguous nucleotides as biomarkers to distinguish between HIV status. Samples from treatment naïve participants from three rounds of an annual population based HIV surveillance programme in rural KwaZulu-Natal were genotyped for drug resistance. The sample types included capillary blood microtubes in 2010 and dried blood spots (DBS) in 2011 and 2012. Using the genetic data available, the proportion of ambiguous nucleotides was calculated. The receiver operator characteristic analysis (ROC) was used to evaluate the diagnostic performance of ambiguities and determine the best cut-off values to identify recent infection. The chi-squared test was used to test for difference in the proportion of participants with ambiguities between recent and chronic infected patients.

In this chapter, we discuss some methods for evaluation the ambiguities as a biomarker in Section 8.2. Results and interpretation of results are described in Section 8.3.

---

## **8.2. Methods**

---

### **8.2.1 Data description**

The study [85, 86] used samples collected from a population-based HIV surveillance conducted in 2011 and 2012 in KwaZulu-Natal, South Africa. The Africa Centre for Health and Population Studies (Africa Centre) has conducted a longitudinal, population-based HIV surveillance programme in the rural district of uMkhanyakude in northern KwaZulu-Natal since 2003. Adult (15-49 years) HIV prevalence in 2011 was 29% [154] and crude HIV incidence was 2.63 per 100 person years. There has been rapid expansion of ART coverage in the area since 2004 with an estimated 37% of all HIV-infected adults on ART in July 2011 [136]. HIV treatment and care is delivered through a decentralized primary health care programme in accordance with the national Department of Health guidelines. HIV-1 viral load tests were done on all dried blood spot (DBS) samples that tested positive for HIV-1 during the 2011 and 2012 surveillance rounds. Only samples from treatment naive participants with viral loads greater than 10,000 RNA copies/ml were genotyped. For participants with more than one sample during the study period, only the earliest sample was used for analysis because the DBS genotyping protocol has low amplification rates at viral loads  $< 10,000$  RNA copies/ml. The HIV-1 RNA was extracted using an automated platform the NucliSense EasyMag - BioMerieux with an elution volume of 50ml for the viral load determination. The same RNA extract was used for HIV-1 drug resistance genotyping within six hours of extraction. Previously published sequences from 2010 were also used in this analysis. The previously described SATuRN/Life Technologies genotyping system was used for the genotyping [85, 86]. Briefly, the extracted RNA was reverse transcribed using the Superscript III first strand synthesis kit (Life Technologies, Foster City, CA) followed by nested PCR using Platinum Taq polymerase (Life Technologies, Foster

City, CA). Successful PCR amplification was assessed using 1% agarose gel (Bioline, Taunton, Massachusetts) electrophoresis run at a 100V and 400mA for 40 minutes. The PCR products were cleaned up using the PureLink QUICK PCR Purification Kit (Life Technologies, Foster City, CA) and sequenced using the Big Dye Terminator kit ver3.1 (Life Technologies, Foster, City) and a set of four bidirectional primers. Capillary sequencing electrophoresis was done on 3130XL Genetic Analyser (Life Technologies, Foster, CA). The sequences covering all the 99 protease codons and the first 300 reverse transcriptase codons were assembled using Geneious Pro genetic analyser [34]. The quality of the sequences was assessed using the HIV-1 Quality Analysis Tool [4] and the Calibrated Population Resistance (CPR) tool [54]. HIV-1 subtyping was performed using the REGA HIV-1 Subtyping Tool ver 3.0 [102]. Phylogenetics was used to rule out contamination among the samples.

The following table shows the encoding for the four bases (A, C, T, G) and for ambiguous positions in the DNA sequence.

A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	G/T (keto)	S	G/C (strong)	Y	T/C (pyrimidine)
M	A/C (amino)	W	A/T (weak)	R	G/A (purine)
B	G/T/C	D	G/A/T	H	A/C/T
V	G/C/A				

With the above, every patient has a sequence of nucleotides, which is either A, C, T or G and ambiguities, which is either W, S, M, K, R, Y, B, D, H, V or N. Thus there was a need to count all the (A,C,T,G) sequences in the DNA sequences and the ambiguities. We compiled a special program in JAVA for this purpose.

The estimated date of infection was calculated as the midpoint between the last negative test date and the first positive test date. The duration of infection (months) was determined by calculating the time between the estimated date of infection and sample date. Participants with a duration of infection up to 36 months were classified as recent infections and non-recent otherwise.

### **8.2.2 Statistical methods**

Our main goal is to evaluate the usefulness of the proportion sequences of ambiguities as a biomarker to predict recent infected patients. For this purpose the generalised linear model (GLM) function in the R software with infection status, recent versus chronic as the outcome and ambiguities as an independent variable was used. Both crude and adjusted analyses, using sex, age, viral load and resistance status were included as covariates, first in a univariate analysis and then a full multivariate analysis.

The relationship between the proportion of ambiguities (the dependent variable) and the duration of infection was investigated using GLM. We performed a bootstrap analysis with 1,000 replicates to obtain 95% confidence intervals in order to assess the ability to accurately discriminate between the recently and chronically infected patients. The proportions of samples with any ambiguity were compared in terms of recent and chronically infected patients, using the Chi-squared test for trend, to see if there was any trend. The trend analysis was confirmed using logistic regression. The model was fitted using the generalised linear model function using ambiguity status, having any ambiguity (W, S, M, K, R, Y, B, D, H, V, N) versus having no ambiguity as the outcome and infection status as the independent variable.

The level of statistical significance was set at 0.05. We sought to identify a cutoff in the pro-

portion of ambiguous sites to classify a patient's infection status into recent (infected for  $\leq 3$  year) or chronic (infected for  $> 3$  year). Optimal cut-off values of the proportion of ambiguous sites were established to distinguish recent from long-term infections. The classification performance of two categorizations of the proportion of ambiguous sites was evaluated with receiver operating characteristic (ROC) analyses. Sensitivity, specificity, positive predictive value (PPV) and AUC were calculated for ambiguities.

---

### **8.3. Results and discussion**

---

Table 8.1 summarizes the HIV-1 variables in recent and chronic patients. The median was used to explain continuous variables, and the fractions were used to explain categorical variables. The median for age was 24 and 31 in recent and chronic patients respectively.

Most participants were females comprising of 80.6% of the sample. The proportion of females in long-term infected patients is 84% while the male proportion was only 16%. We also found that 5% of patients are drug resistant and 34% of all patients have at least one ambiguity. The proportion of recently infected patients having ambiguities was 28% whereas 63% of chronically infected patients did not have any ambiguity. We also obtained proportions of recently versus chronically infected at different times as in Table 8.1. In this table, logvl is the log of viral load, res is resistance and amb is ambiguity.

Univariate and multivariate analyses were performed to identify significant differences between recently infected patients and longtime infected patients. The results of these analyses are shown in Table 8.2.

In univariate analysis the proportion of ambiguities appeared to be significant ( $p$ -value = 0.03) in addition to age and gender. However in multivariate analysis the proportion of ambiguities

Table 8.1: Summary of the HIV-1 variables in recent and chronic patients

Variables		Recent $N = 179$		Chronic $N = 372$	
Age Min - Max (Median)		16 - 77 (24)		16 - 87 (31)	
Gender	Male	46/179	26%	61/372	16%
	Female	133/179	74%	311/372	84%
Logvl (Median)		5.058		4.991	
resistance	res	10/179	6%	19/372	5%
	non-res	169/179	94%	353/372	95%
ambiguities	amb	50/179	28%	139/372	37%
	non-amb	129/179	72%	233/372	63%
Sampling frame	2010	44/179	25%	23/372	6%
	2011	92/179	51%	252/372	68%
	2012	43/179	24%	97/372	26%

was not significant whereas age and gender remained significant.

We calculated the area under the curves for various variables, and the results are shown in Table 8.3. Standard error and confidence intervals were also calculated using the bootstrap method.

The  $p$ -value obtained from linear regression between the proportion of ambiguous and the duration time of infection was 0.107. The proportions of samples with any ambiguity were compared for recent and chronic HIV infected subjects, using the Chi-squared test for trend, to see if there was any trend in ambiguities. The median value of ambiguous nucleotides fractions

Table 8.2: Univariate and multivariate analysis for genetic HIV data

Models	Variables	Estimations	SE	Z-value	P-value
Univariate analysis	ambiguities	0.431	0.198	2.18	0.0295 *
	age	0.063	0.010	6.197	5.76e-10 ***
	sex	-0.567	0.221	-2.566	0.0103 *
	logvl	-0.172	0.161	-1.069	0.285
	resistance	-0.0947	0.4017	-0.236	0.814
Multivariate analysis	ambiguities	0.2801	0.235	1.192	0.233
	age	0.070	0.012	5.844	5.09e-09 ***
	sex	-0.799	0.257	-3.105	0.0019 **
	logvl	-0.113	0.174	-0.650	0.516
	resistance	-0.321	0.428	-0.750	0.454

Table 8.3: AUC estimations for some variable in HIV genetic data

Variables	AUC	SE	CI
sex	0.547	0.019	0.510 - 0.585
resistance	0.502	0.011	0.483 - 0.522
ambiguities	0.547	0.021	0.506 - 0.587
age	0.709	0.023	0.662 - 0.755
logvl	0.527	0.041	0.447 - 0.604

for recent and long-term infections were 27.9% and 37.4%, respectively. The trend analysis was confirmed using logistic regression and the  $p$ -value for trend test is 0.029. We found that

the fraction of ambiguous nucleotides varied between recent and long-term infections, with the fraction increasing with long-term infections. Then we attempted to find the optimal cut-off value of the fraction of ambiguous nucleotides to determine the difference between recent infections and long-term infections. The optimal cut-off value for determination of early HIV-1 infection was 0.5. That is, if the fraction of ambiguous sites from a sequence was not larger than 0.5, then the sequence was from a recently-infected patient, otherwise the sequence was from a long-term infected patient. The sensitivity and specificity were 0.374 and 0.721. The area under the ROC curve (AUC) was 0.547 (95% *CI*, 0.508 – 0.590, *p*-value = 0.015).

Figure 8.1 is the AUC for the proportion of ambiguities and its confidence interval. The optimal cut-off is included with its specificity and sensitivity.

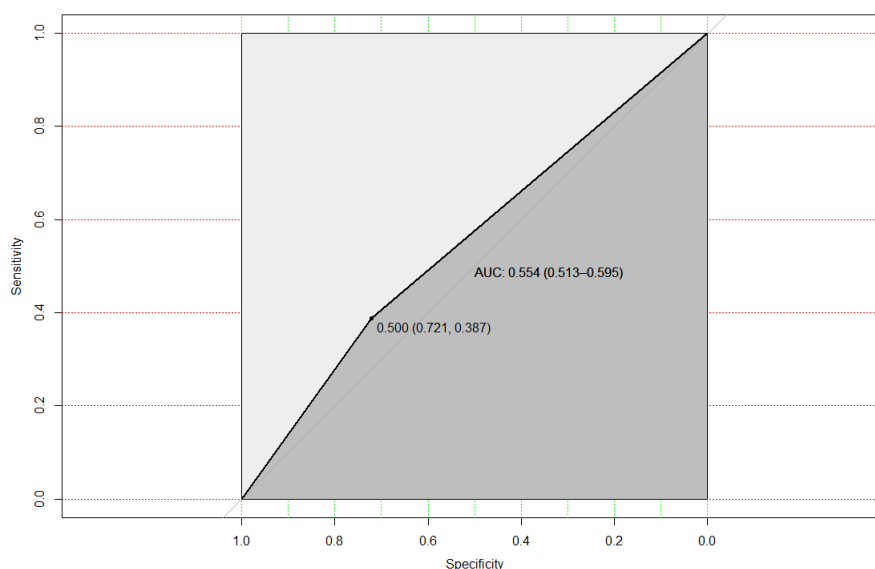


Figure 8.1: Ambiguous AUC estimation using optimal cut-off from HIV genetic dataset

Figure 8.2 explains the proportion of ambiguities for individuals which is spread from 0 to



2.0%.

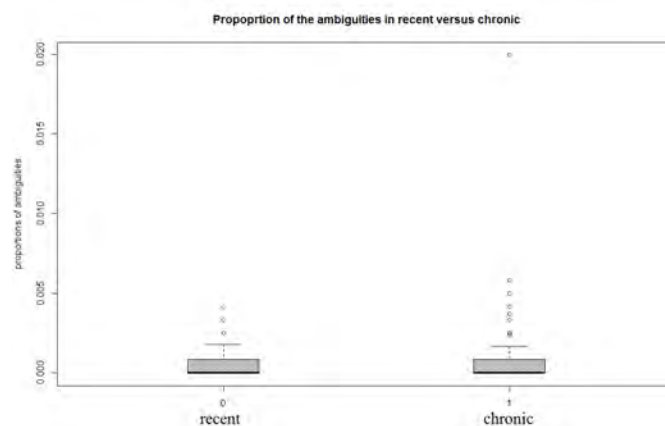


Figure 8.2: Proportion of ambiguities for individuals from HIV genetic dataset

We also calculated the AUC for combination of the ambiguities and other predictors (e.g. sex and age) which is equal to 0.719. This indicates that the combination did not give improvement compared to age alone. In conclusion, the proportion of ambiguous nucleotides may not be a useful marker to distinguish recent from long-term HIV-1 infections. This new genotypic tool cannot be used alone and must be interpreted with clinical and serological data. This method has not performed well in the South African population.

## Conclusion and future work

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 in 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. The first part of this work discussed the terminology of the receiver operating characteristic curve in classic and time dependent scenarios. We also introduced missing data and some of the imputation strategies, then we evaluated many strategies in order to predict disease outcome using simulation studies and applications to interesting real data sets.

In literature many standard approaches for cross-validation suffer from extensive bias or variance when the AUC is used as performance measure. In Chapter 5 we recommend the use of bootstrapping LOOCV for performance estimation, as it avoids many of these problems. The bootstrapping LOOCV estimator is firstly easy to understand and the interpretation of the resulting AUC is straightforward and secondly, LOOCV obtained nearly unbiased estimates. Our proposed methods also involves the variable selection in each bootstrap iteration. It is

possible that the LDA would give higher AUC estimates (the differences were small), but logistic regression would presumably be more robust if LDAs distributional assumptions are violated. The LDA estimator is only valid with a normal distribution data set - although in practice, the two approaches do usually give similar results.

In addition the LDA estimator is only valid with normal distribution. We demonstrated in Chapter 5 that the bootstrap LOOCV estimator of ROC using stepwise logistic regression is useful to estimate the predictive accuracy of prognostic signature. The simulation reveals that the bootstrap cross-validation method is unbiased and outperforms the cross-validation method. We also applied the proposed method to predict TB-IRIS in TB patients, which constitutes an example of an application in medical decision making.

Further extensions of this work will include combining markers measured over time using time to event outcomes including longitudinal data measurements with censored observations to add new insight into the problem. It was noted that more simulation studies are required to investigate other models based on the use of 632 and 632+ resampling methods in order to further understand and improve the methods. Model choice based on predictive criteria methods, which can be viewed as minimizing posterior predictive loss [78, 98] may add some improvements.

In Chapter 6, three imputation methods revealed similar results in the estimation of the time dependent  $AUC_{bcv}(t)$ . The difference between AUC estimates obtained using these three imputation methods were not statistically significant. This conclusion can also be made with application to PBC data (Section 6.7). Practically the MCAR assumption is hard to justify thus we recommend the use of multiple imputation method. We recall that the multiple imputation is based on MAR assumption and allows for the appropriate evaluation of imputation uncertainty. Cox model revealed  $AUC(t)$  estimations similar the true  $AUC(t)$  value, where

logistic regression tends to underestimate the  $AUC(t)$ .

The introduction of a longitudinal component to the analysis added complexities but will be a good extension to the current approach. Considering other types of variable selection methods (for example LASSO) will improve the estimations. In this current work we noted that LOOCV can be time consuming, thus other types of cross-validation combined with advanced bootstrapping methods such as 632+ may better improve the results.

In Chapter 7, a retrospective study of Tuberculous Meningitis in a high HIV prevalence setting at GF Jooste Hospital in Cape Town is used to describe the application of cross-validation, 632 and 632+ bootstrapping. These methods were used together with a penalized Cox model using LASSO variable selection algorithm to estimate the TBM-IRIS scores. We also proposed two level bootstrapping techniques to estimate variances; the proposed method was evaluated through simulation studies. Our simulation results show that the two level bootstrapping method could easily estimate the variances. This method performed better with the 632 estimator compared to BCV and 632+ estimators in terms of coverage probabilities and the proportion of times that the lower limit of confident intervals exclude 0.5, for various scenarios. However this method is time consuming. Our proposed method of using ridge-Cox regression to estimate  $AUC(t)$  obtained similar results to LASSO method, however more simulation studies are required in order to do an appropriate investigation. The ridge-Cox regression model with application to TBM/HIV appeared to be overoptimistic in terms of AUC estimations. Exploring other survival models and alternative AUC estimation approaches will be good extensions to current work. The application show that the CD4 counts could strongly predict the TBM-IRIS patients and could distinguish between the patients who will die before 180 days and those who may survive after that time.

In Chapter 8 the suggested methods that could be applied to the problem of distinguishing

between recent and non-recent HIV infections. We evaluated the use of genotypic tool to distinguish HIV-1 recent infection from long-term HIV-1 infection in South Africa. We analysed the proportion of ambiguous nucleotides as a biomarker to distinguish between HIV-1 status. The aim of the chapter was to test the hypothesis that the method could be used to discriminate between recent and non-recent infections for HIV. Our findings show that the proportion of ambiguous nucleotides may not be a useful marker to distinguish recent from long-term HIV-1 infections. This new genotypic tool cannot be used alone and must be interpreted with clinical and serological data. This method has not performed well in the South African population. Considering the limitation in our study, more investigations are needed to confirm our findings.

In conclusion, we agree that evaluation of multiple biomarkers adds 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. In disease screening, the combination of multiple biomarkers often substantially improves the diagnostic accuracy over a single marker. This is particularly true for longitudinal biomarkers and may improve the diagnosis. We discussed many strategies to select and combine biomarkers in order to address the diagnosis problem. However evaluating these strategies in longitudinal biomarkers will be an interesting extension.

# Bibliography

- [1] S. S. Abdool Karim, K. Naidoo, A. Grobler, N. Padayatchi, C. Baxter and et al. *Timing of initiation of antiretroviral drugs during tuberculosis therapy*, N Engl J Med, **362**(8) (2010), 697 - 706.
- [2] 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.
- [3] M. Akritas, *Nearest neighbor estimation of a bivariate distribution under random censoring*, The Annals of Statistics, **22**(3) (1994), 1299 - 1327.
- [4] L. C. Alcantara, S. Cassol, P. Libin and et al. *A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences*, Nucleic Acids Res, **37** (2009), 634 - 642.
- [5] P. D. Allison, *Multiple imputation for missing data: A cautionary tale*, Sociological Methods and Research, **28** (2000), 301 - 309.
- [6] P. Andersen and R. Gill, *Cox's regression model for counting processes, a large sample study*, Annals of Statistics **10**(4) (1982), 1100 - 1120.

- 
- [7] R. Andridge and R. Little, *A review of hot deck imputation for survey non-response*, International Statistical Review **78**(1) (2010), 40 - 64.
- [8] K. Aoki, J. Misumi, T. Kimura, W. Zhao and T. Xie, *Evaluation of cut 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**(3) (1997), 143 - 151.
- [9] V. Asselman, F. Thienemann, D. J. Pepper, et al., *Central nervous system disorders after starting antiretroviral therapy in South Africa*, AIDS, **24**(18) (2010) 2871 - 2876.
- [10] 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**(11) (2004), 1138 - 1146.
- [11] M. J. Azur, E. A. Stuart, C. Frangakis and P. J. Leaf, *Multiple imputation by chained equations: What is it and how does it work?*, Int J Methods Psychiatr Res., **20**(1) (2011), 40 - 49.
- [12] T. L. Baily, C. Elkan, *Estimating the accuracy of learned concepts*, Proceedings of International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, **20** (1993), 895 - 900.
- [13] D. Bamber, *The area above the ordinal dominance graph and the area below the receiver operating characteristic graph*, Journal of Mathematical Psychology, **12**(4) (1975), 387 - 415.
- [14] 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**(18) (2005), 2873 - 2893.
-

- 
- [15] M. Bankier, J. M. Fillion, M. Luc and C. Nadeau, *Imputing Numeric and Qualitative Variables Simultaneously*. In: Proceedings of the Section on Survey Research Methods, American Statistical Association, 1994, 242 - 247.
- [16] C. Bekondi, C. Bernede, N. Passone, P. Minssart, C. Kamalo and et al., *Primary and opportunistic pathogens associated with meningitis in adults in Bangui Central African Republic, in relation to human immunodeficiency virus serostatus*, Int J Infect Dis, **10**(5) (2006), 387 - 395.
- [17] P. Blanche, A. Latouche and V. Viallon *Time-dependent AUC with right-censored data: a survey study*, Lecture Notes in Statistics, **215**, Springer, New York, 2013.
- [18] C. H. Brown, *Asymptotic comparison of missing data procedures for estimating factor loadings*, Psychometrika, **48**(2) (1983), 269 - 291.
- [19] T. M. Braun and T. A. Alonzo, *A modified sign test for comparing paired ROC curves*, Biostatistics, **9**(2) (2008), 364 - 372.
- [20] L. Breiman and P. Spector, *Submodel selection and evaluation in regression: the x-random case*, International Statistical Review, **60**(3) (1992), 291 - 319.
- [21] T. Cai, M.S. Pepe, Y. Zheng, T. Lumley, and N.S. Jenny. *The sensitivity and specificity of markers for event times*, Biostatistics **7**(2) (2006), 182 - 197.
- [22] G. Campbell, *General methodology I: advances in statistical methodology for the evaluation of diagnostic and laboratory tests*, Statistics in Medicine, **13**(5) (1994), 499 - 508.
- [23] J. Chen and J. Shao, *Nearest neighbour imputation for survey data*, Journal of Official Statistics, **16**(2) (2000), 113 - 131.
- [24] C. Chiang and H. Hung, *Non-parametric estimation for time-dependent AUC*, Journal of Statistical Planning and Inference Statistics, **140**(5) (2010), 1162 - 1174.
-



- 
- [25] J. B. Copas and P. Corbett, *Overestimation of the receiver operating characteristic curve for logistic regression*, *Biometrika*, **89**(2) (2002), 315 - 331.
- [26] M. G. Croda, J. E. Vidal, A. V. Hernandez, T. D. Molin, F. A. Gualberto and et al., *Tuberculous meningitis in HIV-infected patients in Brazil: clinical and laboratory characteristics and factors associated with mortality*, *Int J Infect Dis*, **14**(7) (2010), 586 - 591.
- [27] J. A. Crump, M. J. Tyrer, S. J. Lloyd-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**(4) (1998) 1008 - 1009.
- [28] G. L. Dean, S. D. Edwards, N. J. Ives, G. Matthews, E. F. Fox and et al., *Treatment of tuberculosis in HIV-infected persons in the era of highly active antiretroviral therapy*, *AIDS* **16**(1) (2002), 75 - 83.
- [29] 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**(3) (1988), 837 - 845.
- [30] L. E. Dodd, *Regression Methods for Areas and Partial Areas Under the ROC Curves*, PhD Thesis, University of Washington, 2001.
- [31] L. E. Dodd and M. S. Pepe, *Partial AUC estimation and regression*, *Biometrics*, **59**(3) (2003), 614 - 623.
- [32] 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**(462) (2003), 409 - 417.
-

- 
- [33] 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**(9) (1992), 723 - 731.
- [34] A. J. Drummond, B. Ashton, S. Buxton and et al. *Geneious 5.1*, 2010. Available through <http://www.geneious.com/>
- [35] T. Dunning and D. A. Freedman, *Modeling selection effects*. In W. Outhwaite and S. P. Turner (Eds.), The Sage handbook of social science methodology, 225 - 231, London, Sage.
- [36] C. K. Enders, *Applied Missing Data Analysis*, The Guilford Press, New York, London, 2010.
- [37] B. Efron, *Bootstrap Methods: Another look at the Jackknife*, The Annals of Statistics, **7**(1) (1979), 1 - 26.
- [38] B. Efron, *Estimating the error rate of prediction rule: improvement on cross-validation*, Journal of the American Statistical Association, **78**(382) (1983), 316 - 330.
- [39] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, **57**, Chapman and Hall, 1993.
- [40] J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
- [41] M. B. Elshareef, *Evaluation of Strategies to Combine Multiple Biomarkers in Diagnostic Testing*, MSc Thesis, University of KwaZulu-Natal, South Africa, 2013.
- [42] L. M. Esteban, G. Sanz and . Borque, *Linear combination of biomarkers to improve diagnostic accuracy in prostate cancer*, Monografias Matemticas Garca de Galdeano **38**(1) (2013), 75 - 84.

- 
- [43] R. Etzioni, C. Kooperberg, M. Pepe, R. Smith, and P. H. Gann, *Combining biomarkers to detect disease with application to prostate cancer*, Biostatistics **4**(4) (2003), 523 - 538.
- [44] Y. Fang, Q. Gengsheng and X. Huang, *Optimal combinations of diagnostic tests based on AUC*, Biometrics, **67**(2) (2011), 568 - 576.
- [45] J. Fan and J. Lv *A selective overview of variable selection in high-dimensional Feature Space*, Statistica Sinica, **20**(1) (2010), 101 - 148.
- [46] J. J. Faraway, *Linear Models with R*, Second Edition, Chapman and Hall, Boca Raton London NewYork Washington, D.C., 2014.
- [47] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters Journal, **27**(8) (2006), 861 - 874.
- [48] J. P. Fine and R. J. Bosch, *Risk assessment via a robust probit model, with application to toxicology*, Journal of American Statistical Association, **95**(450) (2000), 375 - 382.
- [49] R. Fluss, *Estimation of the ROC Curves and its Associated Indices Under Verification Bias*, PhD Thesis, University of Haifa, 2007.
- [50] Y. Foucher and R. Danger, *Time dependent ROC curves for the estimation of true prognostic capacity of microarray data*, Statistical Applications in Genetics and Molecular Biology **11**(6) (2012), 1515 - 1544.
- [51] W. J. Fu, R. J. Carroll and S. Wang, *Estimating misclassification error with small samples via bootstrap cross-validation*, Bioinformatics, **21**(9) (2005), 1979 - 1986.
- [52] M. H. Gail and S. B. Green, *A generalization of one-sided two sample Kolmogorov-Smirnov statistics for evaluating diagnostic tests*, Biometrics, **32**(3) (1976), 561 - 570.
-

- 
- [53] J. J. Geoman, *L1 penalized estimation in the Cox proportional hazards model*, Biom J., **52**(1) (2010), 70 - 84.
- [54] R. J. Gifford, T. F. Liu, S. Y. Rhee and et al, *The calibrated population resistance tool: standardized genotypic estimation of transmitted HIV-1 drug resistance*, Bioinformatics, **25**(9) (2009), 1197 - 1198.
- [55] M. Gonen, *Analyzing Receiver Operating Characteristic Curves with SAS*, SAS Institute Inc., Cary, NC, US, 2007
- [56] S. B. Gordon, A. L. Walsh, M. Chaponda, M. A. Gordon, D. Soko and et al. *Bacterial meningitis in Malawian adults: pneumococcal disease is common, severe, and seasonal*, Clin Infect Dis, **31**(1) (2000), 53 - 57.
- [57] J. W. Graham, *Missing data analysis: making it work in the real world*, Annu. Rev. Psychol., **60** (2009), 549 - 576
- [58] S. Greenland and W. D. Finkle, *A critical look at methods for handling missing covariates in epidemiologic regression analyses*, American Journal of Epidemiology, **142**(12) (1995), 1255 - 1264.
- [59] D. M. Green and J. A Swet, *Signal Detection Theory and Psychophysics*, Wiley, New York, 1966.
- [60] 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**(1) (2000), 23 - 41.
- [61] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under an ROC curve*, Radiology, **143**(1) (1982), 29 - 36.
-

- 
- [62] J. A. Hanley, *The robustness of the binormal assumptions used in fitting ROC curves*, Medical Decision Making, **8**(3) (1988), 197 - 203.
- [63] J. A. Hanley, *The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests*, Statistics in Medicine, **15**(14) (1996), 1575 - 1585.
- [64] W. W. Hauck, T. Hyslop and S. Anderson, *Generalized treatment effects for clinical trials*, Statistics in Medicine, **19**(7) (2000), 887 - 899.
- [65] Y. He, A. M. Zaslavsky, M. B. Landrum, D. P. Harrington and P. Catalano, *Multiple imputation in a large-scale complex survey: a practical guide*, **19**(6) (2009), 653 - 670.
- [66] P. J. Heagerty, T. Lumley and M. S. Pepe, *Time-dependent ROC curves for censored survival data and a diagnostic marker*, Biometrics, **56**(2) (2000), 337 - 344.
- [67] P. Heagerty and Y. Zheng, *Survival model predictive accuracy and ROC curves*, Biometrics **61**(1) (2005), 92 - 105.
- [68] W. Hoeffding, *A class of statistics with asymptotically normal distribution*, Annals of Mathematical Statistics, **19**(3) (1948), 293 - 325.
- [69] H. Hung and C.T. Chiang, *Estimation methods for time-dependent AUC models with survival data*, Canadian Journal of Statistics, **38**(1) (2010), 8 - 26.
- [70] A. K. Jam, R. C. Dubes and C. Chen, *Bootstrap techniques for error estimation*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), **9**(5) (1987), 628 - 633.
- [71] J. N. Jarvis, G. Meintjes, A. Williams, Y. Brown, T. Crede and et al., *Adult meningitis in a setting of high HIV and TB prevalence: findings from 4961 suspected cases*, BMC Infect Dis **10** (2010), 10 - 67
-

- 
- [72] Joseph A. Ludwig and John N. Weinstein, *Biomarkers in cancer staging, prognosis and treatment selection*, Nature Reviews Cancer **5**(11) (2005), 845 - 856.
- [73] G. Kalton and D. Kasprzyk, *The treatment of missing survey data*, Survey Methodology **12** (1986), 1 - 16.
- [74] M.G. Kenward, E. Lesaffre and G. Molenberghs, *An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random*, Biometrics, **50**(4) (1994), 945 - 953.
- [75] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Proceedings of the International Joint Conference on Artificial Intelligence (IJ-CAI), Morgan Kaufmann, (1995), 1137 - 1143.
- [76] R. D Kouyos, V. von Wyl, S. Yerly, J. Boni, P. Rieder and et al., *Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection*, Clin Infect Dis, **52**(4) (2011), 532 - 539.
- [77] B. J. LaFleur and R. A. Greevy, *Introduction to permutation and resampling-based hypothesis tests*, Journal of Clinical Child and Adolescent Psychology, **38**(2) (2009), 286 - 294.
- [78] P. W. Laud and J. G. Ibrahim, *Predictive model selection*, Journal of the Royal Statistical Society Ser. B **57**(1) (1995), 247 - 262.
- [79] R. J. Little, *Robust estimation of the mean and covariance matrix from data with missing values*, Applied Statistics, **37**(1) (1988), 23 - 38.
- [80] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley, New York, 1987.
-

- 
- [81] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2<sup>nd</sup> edition, New York, John Wiley, 2002.
- [82] 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 and STDS* **21**(4) (2007) 234 - 239.
- [83] C. E. Lunneborg, *Data Analysis by Resampling: Concepts and Applications*, Pacific Grove, CA: Duxbury, 2000
- [84] S. Ma and J. Huang, *Regularized ROC method for disease classification and biomarker selection with microarray data*, *Bioinformatics*, **21**(24) (2005), 4356 - 4362.
- [85] J. Manasa, D. Katzenstein, S. Cassol, M. L. Newell and T. de Oliveira, *Primary Drug Resistance in South Africa: Data from 10 Years of Surveys*, *AIDS Res Hum Retroviruses*, **28**(6) (2012), 558 - 565.
- [86] J. Manasa, S. Danaviah, S. Pillay and et al. *An affordable HIV-1 drug resistance monitoring method for resource limited settings*, *J Vis Exp*, **30**(85) (2014), doi: 10.3791/51242.
- [87] J. Manasa, S. Danaviah, R. Lessells, M. Elshareef, E. Wilkinson, S. Pillay, H. Mthiyane, H. Mwambi, D. Pillay and T. de Oliveira, *HIV-1 drug resistance in adults participating in a population based HIV surveillance in rural KwaZulu-Natal South Africa*, submitted.
- [88] 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**(6) (2006) 357 - 363.
- [89] S. Marais, J. Dominique, D. J. Pepper, C. Schutz, R. J. Wilkinson and G. Meintjes, *Presentation and outcome of tuberculous meningitis in a high HIV prevalence setting*, *PLoS One*, **6**(5) (2011), doi: 10.1371/journal.pone.0020077
-

- 
- [90] S. Marais, G. Meintjes, D. J. Pepper, L. E. Dodd, C. Schutz<sup>1</sup>, Z. Ismail, K. A. Wilkinson and R. J. Wilkinson, *Frequency, severity and prediction of tuberculous meningitis immune reconstitution inflammatory syndrome*, Clinical Infectious Diseases, **56**(3) (2012), 450 - 460.
- [91] R. J. Marshal, *The predictive value of simple rules for combining two diagnostic tests*, Biometrics, **45**(4) (1989), 1213 - 1222.
- [92] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2<sup>nd</sup> edition, Chapman and Hall, Boca Raton, London, New York, Washington, D.C., 1989.
- [93] M. W. McIntosh and M.S. Pepe, *Combining several screening tests: Optimality of the risk score*, Biometrics **58**(3) (2002) 657 - 664.
- [94] P. E. McKnight, K. M. McKnight, S. Sidani and A. J. Figueredo, *Missing data: A gentle introduction*, The Guilford Press, New York, 2007
- [95] G. Meintjes , S. Lawn , F. Scano, et al., *Tuberculosis-associated immune reconstitution inflammatory syndrome: case definitions for use in resource-limited settings*, Lancet Infect Dis, **8**(8) (2008), 516 - 523.
- [96] X. L. Meng *Multiple imputation inferences with uncongenial sources of input*, Statistical Science, **9**(4) (1994), 538 - 573. .
- [97] C. E. Metz, *Some practical issues of experimental design and data analysis in radiological ROC studies*, Investigative Radiology, **24**(3) (1989), 234 - 245.
- [98] M. C. Meyer and P. W. Laud *Predictive Variable Selection in Generalized Linear Models*, Journal of the American Statistical Association, **97**(459) (2002), 859-871.
-



- 
- [99] A. Moise, B. Cliement and M. Raissis, *A test for crossing receiver operating characteristic (ROC) curves*, Communications in Statistics Theory and Methods, **17**(6) (1988), 1985 - 2003.
- [100] Hu. Nan, *Regression methods of time-dependent ROC curve for evaluating the prognosis capacity of biomarkers*, PhD Thesis, University of Washington, 2010.
- [101] N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner and D. H. Bent, *SPSS: Statistical Package for the Social Sciences*, 2<sup>nd</sup> edition, McGrawHill, New York, 1975.
- [102] T. de Oliveira, K. Deforche, S. Cassol and et al., *An automated genotyping system for analysis of HIV-1 and other microbial sequences* Bioinformatics, **21**(19) (2005), 3797 - 3800.
- [103] M. S. Pepe and M. L. Thompson, *Combining diagnostic test results to increase accuracy*, Biostatistics, **1**(2) (2000), 123 - 140.
- [104] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Sciences Series, **31**, Cambridge University Press, 2003.
- [105] M. S. Pepe and T. Cai, *The analysis of placement values for evaluating discriminatory measures*, Biometrics, **60**(2) (2004), 528 - 535.
- [106] M. S. Pepe, T. Cai and G. Longton, *Combining predictors for classification using the area under the receiver operating characteristic curve*, Biometrics, **62**(1) (2006), 221 - 229.
- [107] D. J. Pepper, S. Marais, G. Maartens and et al, *Neurologic manifestations of paradoxical tuberculosis associated immune reconstitution inflammatory syndrome: a case series*, Clin. Infect. Dis. **48**(11) (2009), 96 - 107.
-

- 
- [108] A. Perez, R. Dennis, J. Gil, M. Rondon and A. Lopez, *Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia*, Stat. Med. **21**(24) (2000), 3885 - 3896.
- [109] A. Perperoglou, *Cox models with dynamic ridge penalties on time varying effects of the covariates*, Statistics in Medicine, **33**(1) (2014), 170 - 180.
- [110] 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.
- [111] T. E. Raghunathan, J. M. Lepkowski, J. van Hoewyk and P. Solenberger, *A multivariate technique for multiply imputing missing values using a series of regression models*, Survey Methodology, **27**(1) (2001), 85 - 96.
- [112] T. E. Raghunathan, P. W. Solenberger and J. Van Hoewyk, *IVEware: Imputation and Variance Estimation Software User Guide*, Michigan, University of Michigan, 2002.
- [113] D. B. Rubin, *Inference and missing data*, Biometrika, **63**(3) (1976), 581 - 592.
- [114] D. B. Rubin, *Statistical matching using file concatenation with adjusted weights and multiple imputations*, Journal of Business and Economic Statistics, **4**(1) (1986), 87 - 94.
- [115] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, J. Wiley and Sons, New York, 1987.
- [116] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London, 1997.
- [117] J. L. Schafer, *Multiple imputation: A primer*, Statistical Methods in Medical Research, **8**(1) (1999), 3 - 15.
-

- 
- [118] J. L. Schafer and J. W. Graham, *Missing data: Our view of the state of the art*, Psychological Methods, **7**(2) (2002), 147 - 177.
- [119] D. E. Shapiro, *The interpretation of diagnostic tests*, Statistical Methods in Medical Research, **8**(2) (1999), 113 - 134.
- [120] J. Siddique and T. R. Belin, *Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data*, **53**(2) (2008), 405 - 415.
- [121] M. Sill, T. Hielscher, N. Becker and M. Zucknick *Extended Inference with Lasso and Elastic-Net Regularized Cox and Generalized Linear Models*, Journal of Statistical Software, **62**(5) (2014), 10.18637/jss.v062.i05.
- [122] 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.
- [123] H. Skalská and V. Freylich, *Web-Bootstrap estimate of area under ROC curve*, Austrian Journal of Statistics, **35**(2-3) (2006), 325 - 330.
- [124] I. Sohn and C. Sung, *Predictive modeling using a somatic mutational profile in ovarian high grade serous carcinoma*, PLoS ONE **8**(1) (2013), 540 - 589.
- [125] H. H. Song, *Analysis of correlated ROC areas in diagnostic testing*, Biometrics, **53**(1) (1997), 370 - 382.
- [126] 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.
- [127] S. A. Spector, *Mother-to-infant transmission of HIV-1: the placenta fights back*, J Clin Invest. **107**(3) (2001), 267 - 269.
-

- 
- [128] K. Strimbu and M.D. Jorge A. Tavel, *What are Biomarkers?*, HIV AIDS, **5**(6) (2010), 463 - 466.
- [129] E. A. Stuart, M. Azur, C. E. Frangakis and P. J. Leaf, *Practical imputation with large datasets: A case study of the children's mental health initiative*, American Journal of Epidemiology, **169**(9) (2009), 1133 - 1139.
- [130] J. Q. Su and J. S. Liu, *Linear combination of multiple diagnostic markers*, Journal of the American Statistical Association, **88**(424) (1993), 1350 - 1355.
- [131] K. Subsai, S. Kanoksri, C. Siwaporn and L. Helen, *Neurological complications in AIDS patients: the 1-year retrospective study in Chiang Mai University, Thailand*, Eur J Neurol, **11**(11) (2004), 755 - 759.
- [132] J. A. Swet and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press, New York, 1982.
- [133] J. A. Swet, *Indices of discrimination or diagnostic accuracy: Their ROCs and implied models*, Psychological Bulletin, **99**(1) (1986), 100 - 117.
- [134] J. A. Swet, *Measuring the accuracy of diagnostic systems*, Science, **240**(4857) (1988), 1285 - 1293.
- [135] J. A. Swet, R. M. Dawes and J. Monahan, *Better decisions through science*, Scientific American, **283**(4) (2000), 82 - 87.
- [136] F. Tanser, T. Barnighausen, E. Grapsa, J. Zaidi and M. L. Newell, *High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa*, Science, **339**(6122) (2013), 966 - 971.
- [137] T. Therneau and P. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York, 2000.
-

- 
- [138] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society. Series B (Methodological), **58**(1) (1996), 267 - 288.
- [139] R. Tibshirani, *The LASSO method for variable selection in the Cox model*, Statistics in Medicine, **16**(4) (1997), 385 - 395.
- [140] M. E. Torok, T. T. Chau, P. P. Mai, N. D. Phong, N. T. Dung and et al., *Clinical and microbiological features of HIV-associated tuberculous meningitis in Vietnamese adults*, PLoS One, **3** (2008), e1772, DOI: 10.1371/journal.pone.0001772.
- [141] A. N. Tosteson and C. B. Begg *A general regression methodology for ROC curve estimation*, Medical Decision Making, **8**(3) (1988), 204 - 215.
- [142] 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, **21**(8) (2007), 527 - 532.
- [143] S. van Buuren, H. C. Boshuizen and D. L. Knook, *Multiple imputation of missing blood pressure covariates in survival analysis*, Statistics in Medicine, **18**(6) (1999), 681 - 694.
- [144] S. van Buuren, *Multiple imputation of discrete and continuous data by fully conditional specification*, Statistical Methods in Medical Research, **16**(3) (2007), 219 - 242.
- [145] H. C. van Houwelingen , T. Bruinsma , A. A. Hart, L. J. van'tVeer, L. F. Wessels, *Cross-validated Cox regression on microarray gene expression data*, Statistics in Medicine, **25**(18) (2006), 3201 - 3216.
- [146] E. S. Venkatraman and C. B. Begg, *A distribution free procedure for comparing receiver operating characteristic curves from a paired experiment*, Biometrika, **83**(4) (1996), 835 - 848.

- 
- [147] E. S. Venkatraman, *A permutation test to compare receiver operating characteristic curves*, Biometrics, **56**(4) (2000), 1134 - 1138.
- [148] 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**(3) (2003), 177 - 178.
- [149] F. Wang-Clow, M. Lange, N. M. Laird and J. H. Ware, *A simulation study of estimators for rates of change in longitudinal studies with attrition*, Statistics in Medicine, **14**(3) (1995), 283 - 297.
- [150] 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**(3) (1989), 585 - 592.
- [151] X. Xue, M. Kim and R. Shore, *Cox regression analysis in presence of collinearity: an application to assessment of health risks associated with occupational radiation exposure* Lifetime Data Analysis, **13**(3) (2007), 333 - 350.
- [152] W. Yu and T. Park, *Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the ROC curve*, Computational Statistics and Data Analysis, **88**(1) (2015), 15 - 27.
- [153] Z. Yuan and D. Ghosh, *Combining multiple biomarker models in logistic regression*, Biometrics **64**(2) (2008) 431 - 439.
- [154] J. Zaidi, E. Grapsa, F. Tanser, M. L. Newell and T. Barnighausen, *Dramatic increase in HIV prevalence after scale-up of antiretroviral treatment* AIDS, **27**(14) (2013), 2301 - 2305.

- [155] 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.
- [156] X. H. Zhou, N. A. Obuchowski and D. K. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley, New York, 2002.
- [157] K. H. Zou, *Receiver operating characteristic (ROC) literature research*, Harvard 2002.