APPLICATION OF NOVEL STATISTICAL METHODS FOR BIOMARKER SELECTION TO HIV INFECTION DATA

Bosny J PIERRE-LOUIS

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of DrPH in the Department Biostatistics of the UNC Gillings School of Global Public Health.

Chapel Hill
2010

Approved by:

C. M. Suchindran, PhD

Pai-Lien Chen, PhD

Charles S. Morrison, PhD

Stephen R. Cole. PhD

Pranab K. Sen. PhD

ABSTRACT

Bosny J Pierre-Louis: APPLICATION OF NOVEL STATISTICAL METHODS FOR BIOMARKER SELECTION TO HIV INFECTION DATA

(Under the Direction of Drs. C.M. Suchindran and Pai-Lien Chen)

The past decade has seen an explosion in the availability and use of biomarker data as a result of innovative discoveries and recent development of new biological and molecular techniques. Biomarkers are essential for at least four key purposes in biomedical research and public health practice: they are used for disease detection, diagnosis, prognosis, to identify patients who are most likely to benefit from selected therapies, and to guide clinical decision making. Determining the predictive and diagnostic value of these biomarkers, singly or in combination, is essential to their being used effectively, and this has spurred the development of new statistical methodologies to assess the relationship between biomarkers and clinical outcomes. One active area of research is the development of variable importance measures, a class of estimators that could reliably capture the effect of a specific biomarker on a clinical outcome. The central question addressed in this dissertation is the following: Given a large set of biomarkers that potentially predict a clinical outcome, how can one make a determination as to which ones are the most important? In the first paper, we estimate a targeted variable importance measure through Van der

Laan's theory of targeted maximum likelihood estimation in the point treatment setting and use the same objective function to compute an alternative measure of marginal variable importance based on weights from a flexible propensity score model. Covariate-adjusted targeted variable importance measures are compared to estimates from this alternative methodology and to incremental value estimates from partial ROC curves. In the second paper, we extend the applicability of the TMLE methodology to analyze longitudinal repeated measures data. It addresses the gap caused by the absence of a generally accepted approach for generating a longitudinal variable importance index by proposing an estimator involving both TMLE and computation of the area under or above the LOESS curve. A graphical method is proposed for visual assessment of the longevity of a biomarker in terms of its predictive power, information that could be used to determine when repeated measures of a biomarker should be taken. Finally, in the third paper we take right censoring in the outcome variable into consideration and achieve biomarker selection in the presence of confounding and potential informative censoring through the use of stabilized weights in a time-dependent Cox proportional hazards model. A dataset from the Hormonal Contraception and HIV Genital Shedding and Disease Progression Study that includes longitudinal HIV infection data on a sample of 306 HIV-infected adult women from Uganda and Zimbabwe was used to develop and evaluate the methods discussed in the three papers. This study collected information on a number of biomarkers related to HIV infection, including plasma viral load, HIV subtype, CD4 and CD8 lymphocyte counts, hemoglobin level, and *herpes simplex virus* 2 (HSV-2). The relationships of these biomarkers with changes in CD4 cell counts were considered in three different contexts: cross-sectional,

longitudinal and survival. In short, baseline CD4 cell counts, HIV subtype, and HSV-2 were found to be important biomarkers for the outcome variable studied.

ACKNOWLEDGEMENT

This dissertation would not have been possible without the help and support of numerous people. I am particularly pleased to thank my wife Marjorie, my son Neil, and my daughter Maelie for their unconditional love and support. I hope this conveys to them my expression of gratitude!

I would like to take this opportunity to express my admiration, respect, and gratitude to my academic and dissertation advisor, Dr. C.M. Suchindran. Without his help and guidance, it would have been almost impossible to complete the DrPH program and to write my dissertation. Over the years, his support, encouragement, and guidance have been invaluable. Dr Suchi has always been available to help and has served as a constant source of fresh ideas throughout the dissertation process. He graciously reviewed countless versions of the manuscript, and his suggestions have helped improve vastly the quality of my work.

My heartfelt gratitude goes also to Dr. Pai-Lien Chen. Over the span of the past 10 years, Dr Chen has served as a trusted mentor who has always shown great interest in both my professional and academic growth. He has provided me with insightful references and ideas critical for this dissertation. I credit him with sparking my interest towards novel statistical methods appropriate for the analysis of observational data, a key theme of this dissertation. Dr Chen has been kind

enough to review numerous versions of my manuscript, and he has always provided excellent suggestions for improvement.

I would also like to communicate my deepest thanks to Dr Stephen Cole for his innovative ideas and for providing me with unparalleled opportunity to learn from his expertise in causal inference and counterfactual methodology. Dr Cole has not only proposed insightful revisions to the manuscript, he has also steered me towards key pertinent references and work upon which my dissertation has drawn largely.

I am deeply appreciative of the tremendous support of Dr Charles Morrison and Family Health International (FHI) for granting me permission to use the Hormonal Contraception and HIV Genital Shedding and Disease Progression (GS) study dataset. Dr Morrison's input throughout the process has been instrumental for the successful completion of my dissertation. His superior knowledge of substantive issues concerning biomarkers in HIV research provided a unique perspective to the research problem and helped immensely in the interpretation of the results. I would also like to thank Cynthia Kwok for creating the analysis datasets and for having always been available to answer my questions. I am grateful to the women of Uganda and Zimbabwe who took part in the study. Without their participation in this study, I could not have access to such a relevant and rich dataset to illustrate the methods covered in this dissertation.

I'd like to send my deepest thanks to Dr. P.K. Sen whose judicious suggestions have helped strengthen the statistical methodology used in this dissertation. Dr Sen was kind enough to recommend revisions that enhance greatly the quality of the manuscript and maintain scientific rigor.

Lastly, I would like to acknowledge countless other people who have supported me in various respects during the completion of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

1. Introduction/Background

Fueled by recent advances in modern biology and technology, biomarkers have become a popular research topic in clinical investigations. As biological tools that can be used to monitor the presence or absence of disease, disease progression, the effect of a treatment, and the toxicity of a drug, biomarkers are important to the pharmaceutical industry, to federal regulatory agencies such as the Food and Drug Administration (FDA), and to public health researchers. In the pharmaceutical industry, the growing need for biomarker data collection stems from, among many benefits, their usefulness in driving decision-making, particularly at early phases of clinical trials. By facilitating the identification of positive responders and non-responders to therapeutics, biomarkers have provided an impetus for the development of targeted therapies and personalized medicines. By the same token, they have the potential to improve the late phase trial success rate through better decision making earlier in the process of drug development. This targeted strategy has the potential to improve efficiency in the drug development process while at the same time maximizing patient safety and efficacy.

Another practical advantage of biomarkers is that they have the ability to reduce the need for hard clinical endpoints. Overall, biomarkers lend themselves to earlier and easier measurements than clinical outcomes, are less subject to competing risks,

and can reduce clinical trials sample size requirements. In pharmaceutical research, this could translate into more cost-effective trials, shorter study duration, improved compliance, and the opportunity to bring new therapies to the market more quickly. In the public health arena, biomarker data can be used to measure the prevalence of certain health conditions, to identify disease risk factors, and to evaluate the impact of interventions. They could play a pivotal role in disease prevention efforts by providing ways to detect diseases in the preclinical stage when it may be possible to achieve a positive reversal in the outcome. In vaccine trials, biomarkers reduce reliance on costly and lengthy efficacy studies by facilitating identification of serologic tests that predict protection from given conditions or by making possible earlier assessment of the safety and efficacy of candidate vaccines (Hogrefe, 2005). The numerous benefits of biomarkers in clinical research explain why the watershed FDA–National Institutes of Health (NIH) consensus conference biomarkers held in 1999, the 2006 FDA Critical Path Initiative, the European Medicines Agency (EMEA) Road Map to 2010, and various stakeholders from the pharmaceutical industry and from patient advocacy groups, have all called for a key role of biomarkers in the drug development process. Collaborative efforts to advance biomarkers research are exemplified by the work of the Biomarkers Consortium (http://www.biomarkersconsortium.org), a public-private partnership, launched in 2006 with the goal to "identify and qualify new biological markers to speed the detection, diagnosis, and treatment of disease" and by the innovative approach to cancer biomarker development taken by the National Cancer Institute's Early Detection Research Network (http://edrn.nci.nih.gov/).

In spite of the popularity of biomarkers, their vast spectrum of uses and utility, and the ever-increasing availability of biomarker data, there are still concerns about either a lack of progress in biomarker discovery or a gap between the pace of biomarker discovery and the development of novel statistical methods to evaluate their different performance characteristics. Such factors could serve as an impediment to effective treatment/cure for diseases such as diabetes, obesity, endocrine, digestive, and metabolic conditions, among others. The prerequisite for clinical use of biomarkers in the diagnosis, treatment, and prognosis of these conditions is an adequate biomarker discovery process. Key statistical questions asked during this process are whether the candidate biomarkers reliably predict the outcome of interest, whether the observed association between candidates and outcome is not confounded by extraneous factors, and which candidates display the best performance characteristics. Answers to these questions require the development and application of novel statistical methods for biomarker selection and validation.

In this dissertation, the research question is how to quantify the impact of a biomarker on a given outcome (e.g. CD4 cell count) and use this measure of impact as a tool for ranking biomarkers. Answers to this question are of great practical importance to public health. By allowing researchers to take a pool of biomarkers and make a reliable determination of which ones are the most important, such measures could lead to faster decision making in epidemiologic studies and clinical trials. For instance, an effective and efficient selection of the best subset of biomarker amongst a set could help direct further biological research in early phases

of clinical trials by limiting the focus to the pool of most promising ones only. From a scientific standpoint, this could guide statistical research in the area of biomarker validation by making available a list of good candidates for surrogacy status determination. Such a list could also be useful in the quest for the best combinations of biomarkers.

1.1. Critical Evaluation of Existing Knowledge

1.1.1. Biomarker and HIV Review

According to the Biomarkers Definition Working Group, a task force convened by the National Institutes of Health (NIH), a biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Clinical Pharmacology and Therapeutics 2001, 69: 89–95). This broad definition encompasses all diagnostic tests, imaging technologies, and any other objective measures of a person's health status (Desai et al, 2006). Despite the fact that over the past decade biomarkers have been a driving force behind medical practice innovation, concerns over the availability of sound and effective statistical methods necessary to evaluate their characteristics have impeded their efficient application (Lasserre et al, 2007). Given that biomarker data can be used for disease detection, diagnosis, prognosis, or for identification of patients who are most likely to benefit from selected therapies (Alaiya et al, 2005), advances in biomarker discovery could pave the way towards a predictive, preventative, and personalized approach to medicine (Weston and Hood, 2004). Thus, it is imperative to develop

improved statistical methods that can help harness the vast potential of biomarkers in clinical practice.

A biomarker that is intended to substitute for a clinical endpoint is a surrogate endpoint. Temple (1995) defines a surrogate endpoint as a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives. Changes induced by a therapy on a surrogate endpoint are expected to reflect changes in a clinically meaningful endpoint.

The process of conferring surrogacy status to a biomarker entails a rigorous qualification process. One element of this process is statistical validation. The goal, in validating surrogate markers, is to establish that they can reliably predict clinical benefit, harm, or futility of a new therapy. A qualification process map for the purpose of classifying biomarkers as exploratory, probable valid, or known valid has been established by the FDA (Goodsaid and Frueh, 2007).

Prentice (1989) was the first to provide a statistical definition of surrogacy. According to this operational definition, a surrogate marker is a response variable for which the conditional distribution of the outcome given the surrogate marker alone is the same as the conditional distribution of the clinical outcome given the surrogate marker and the treatment. Borrowing notation from Molenberghs and colleagues (2009), let T and S be random variables denoting true and surrogate endpoints respectively, and let Z be an indicator variable for treatment. To establish Prentice's criterion, the following conditions are necessary (Molenberghs et al., 2009):

1. Treatment Z has a significant impact on the surrogate endpoint S.
2. Treatment Z has a significant impact on the true endpoint T.
3. The surrogate endpoint S has a significant impact on the true endpoint T.
4. The surrogate endpoint S captures the entire effect of the treatment Z upon the true endpoint T.

The current consensus in statistical literature is that Prentice's criteria are neither sufficient nor necessary and are difficult to achieve (Molenberghs et al., 2004, 2009; Qu and Case, 2006). Also, a serious drawback of Prentice's criterion, identified by Frangakis and Rubin (2002) is its lack of causal interpretation. Prentice's criterion reflects net effects, a combination of the causal effect of a treatment and systematic differences between compared groups arising from possible selection bias. Frangakis and Rubin (2002) proposed instead a surrogate validation model based on counterfactuals. They introduced a new definition of surrogate endpoint, the "principal surrogate" based on principal stratification and principal effects. In this framework, principal effects with respect to a post-treatment variable (e.g. biomarker) are evaluated within principal strata defined as "cross-classification of subjects defined by the joint potential values of that post-treatment variable under each of the treatments being compared" (Frangakis and Rubin, 2002). Principal strata are not affected by treatment and, therefore, play a role akin to baseline covariates. Comparisons within principal strata yield causal interpretations as these comparisons are made between comparable groups of subjects.

Using the principal stratification proposed by Frangakis and Rubin, Gilbert and Hudgens (2008) introduced an estimand for evaluating a principal surrogate, the

"causal effect predictiveness surface". This quantity measures how well causal treatment effects of the biomarker predict causal treatment effects of the clinical endpoint (Gilbert and Hudgens, 2008).

For an exhaustive review of statistical approaches to biomarker validation, see Weir and Walley (2006), Lasserre (2007), Buyse et al (1998, 2000), Molenberghs et al. (2009), Frangakis and Rubin (2002). Biomarker validation is beyond the scope of this research.

In spite of their advantages, biomarkers do have limitations that warrant a cautionary note. They can be prone to measurement errors, storage problems, high costs of measurements, and can be fraught with ethical issues (Mayeux, 2004). Extreme caution should be exercised when biomarkers are used to replace true clinical endpoints. According to existing literature, earlier attempts to use biomarkers as surrogates have led to some flawed and sometimes harmful conclusions (Fleming and Demets, 1996). A well publicized failed attempt to use surrogate endpoints has been the FDA approval of the drugs encainide, flecainide, and moricizine, predicated on the notion that arrhythmia suppression by these drugs would lead to overall survival. The use of electrocardiographic findings as surrogates for sudden death was subsequently undercut by the Cardiac Arrhythmia Suppression Trial (CAST) that found higher mortality and non fatal cardiac arrest events in patients taking these drugs as compared to placebo (NEJM, 1989, 321:406-12; Echt et al., 1991).

Notwithstanding these concerns, biomarkers continue to play an increasingly important role in drug development. As a driving force for pharmaceutical

innovation and personalized medicine (pharmacogenomics), biomarkers have the potential to lead to faster decision making since they can lend themselves to earlier and easier measurements than related clinical outcomes and could be less subject to competing risks. This could translate into more cost-effective trials, shorter drug study duration, improved compliance, and the opportunity to bring new therapies to the market more quickly, as evidenced by the benefits to the whole drug life cycle brought by the accelerated approval of the drug Herceptin (trastuzumab) by the FDA in 1998. Manufactured by Genentech/Roche, this drug targets the 25-30% of breast cancer patients who have the genetic alteration of the human epidermal growth factor receptor2 protein HER2. As a result of accelerated approval for this molecular targeted therapy, Roche saved an estimated $35 million in clinical trial costs, collected $2.5 billion of income, while 120,000 patients gained access to the drug earlier than they would normally do (Thomson Scientific White Paper, March 2008). Note that this drug has been approved together with a molecular diagnostic that could determine (based on the expression of genetic biomarker HER-2) whether a patient might benefit from the drug.

An area of research where the use of biomarkers can foster innovative advances is HIV/AIDS - the focus for the application of statistical methods used in this research. In fact, based on existing regulation known as the subpart H Approval that allows the use of surrogate endpoints for serious or life-threatening illnesses, the FDA has used scientific evidence from biomarkers to grant approval for several anti-HIV drugs, including Didanosine, Nevirapine, Lopinavir, and Efavirenz (Desai et al, 2006). Over time, with the increasing complexity of HIV infection research and the quest for

a safe and effective treatment, the use of biomarkers related to HIV infection remains an active area of research. In an article published in April 2009 in the Journal of Acquired Human deficiency Syndrome (JAIDS), MacLachlan et al. touted the multiple beneficial roles that biomarkers could play in HIV vaccines. This paper emphasized how biomarkers such as the BED enzyme immunoassay (BED-EIA) and the nucleic acid amplification testing (NAAT) could be used to estimate prevalence and incidence of HIV, information that could then be used for the determination of sample size and study populations for HIV efficacy trials. It also demonstrated how sexually transmitted infection (STI) biomarkers could be used to generate valuable information on HIV infectiousness, transmissibility, and disease progression (MacLachlan et al., 2009).

In HIV infection studies, HIV RNA copies and CD4 cell counts are biomarkers routinely used to assess response to treatment or to monitor the progression of disease (Ellenberg, 1991; Lagakos and Hoth, 1992; Machado et al., 1990; Fleming, 1994; Chen et al., 2007; Brown et al., 2009). As a measure of a patient's immune capacity, CD4 cell count is considered as a standard method for determining eligibility for highly active antiretroviral therapy (HAART) and HIV disease progression. Using a cohort of 489 Kenyan pregnant women, Brown et al (2009) found that CD4 could be a useful predictor of mortality. This is in line with prior conclusions from both individual studies (Kawado et al, 2006; Planella et al. 1998; Liotta et al. 2004) and meta-analysis (Cross Continents Collaboration for Kids (3Cs4kids) Analysis and Writing Committee, 2008). However, it is not always possible to measure CD4 cell count, especially in resources-deprived settings. Both

viral load and CD4 cell count measurements require highly skilled personnel and costly maintenance of sophisticated pieces of equipment. These costs can be prohibitive in resource-poor countries, thereby limiting access to these tests for those who need them the most. A literature search reveals a few studies that examined the usefulness of less expensive biomarkers as potential surrogates for a CD4 cell count. The results, however, have not always been conclusive. Some studies have found HIV-1 RNA to be the best predictor of long term CD4 cell count responses and disease progression (Mellors et al., 1997; Fiscus et al., 1998). Others have suggested that total lymphocyte count (TLC) is a good predictor of low CD4 cell counts (Montaner et al, 1992; Blatt et al. 1993; Martin et al, 1995; Shapiro et al., 1998). At least one study concludes that TLC is not a good predictor of CD4 cell count and, therefore, should not be used in the clinical care of HIV/AIDS patients (Van der Ryst et al., 1998). For a more complete list of references covering the relationship between TLC and CD4 cell count, see Chen et al (2007). Overall, no studies have attempted to provide a single scalar measure of the marginal importance of biomarkers on CD4 cell count, to the best of our knowledge. The closest attempt was made in Brown et al. (2009) who provided screening performance measures for several biomarkers separately. However, the clinical outcome considered was mortality, and no attempt was made to generate and compare results from different statistical methodologies. What is lacking is a unified list of biomarkers that can predict of CD4 cell counts based on sound statistical methodologies. This dissertation is intended to fill these gaps.

1.1.2. Review of statistical methods for biomarker selection

A survey of statistical methods used for biomarker selection reveals that both standard and novel statistical methods have been employed to address the challenges of biomarker selection. A non-exhaustive list of methods used to this end include univariate testing; classical multivariable regression techniques such as ordinary least squares and logistic regression; the Receiver-Operating Characteristic (ROC) curve; non-linear models and machine learning techniques such as classification and regression tree, Bagging, Boosting, random forest, and pattern recognition techniques; and marginal structural models for causal inference. We present a summary of some of these methods below.

1.1.2.1. Univariable Screening

In the univariate setting, a series of separate tests of the null hypothesis of no difference in the distribution of each biomarker across groups (e.g. diseased and no diseased) are performed, and screening is made based on the p-values associated with those tests. For comparison of biomarkers measured on a continuous scale, t-tests or variants thereof (Wilcoxon test, Welch test) are used (Dudoit el al., 2002; Guoan et al., 2002; Tusher et al., 2001; Cui and Churchill, 2003). For binary biomarker variables, chi-square tests of Fischer exact tests are often conducted.

From a modeling standpoint, simple linear regression is often used to model the relationship between the outcome and each biomarker separately. In this setting, parameter estimation is done via the method of least squares. Simple logistic regression is a common choice for binary outcomes. The coefficient for each biomarker of interest is interpreted as its importance measure. P-values, based on

univariate analysis of group mean differences for each biomarker are often adjusted for multiplicity (Tuglus, 2008).

1.1.2.2. Multivariable Screening

Multivariable screening, linear or non-linear, provides an analytical framework where all biomarkers can be evaluated simultaneously with or without covariate adjustment. Classical multivariable regression techniques such as ordinary least squares and logistic regression are two widely used linear models. An example of widely popular methods based on ordinary least squares estimation is the analysis of covariance (ANCOVA) where potential confounders are also included as predictors in a regression model (Cook and Campbell, 1979).

In more recent developments, focus has been placed on non-linear models and machine learning techniques to improve biomarker prediction, classification and selection. Among those, classification and regression tree (Breiman et al., 1984), Bagging (Breiman, 1996), Boosting (Freund and Schapire, 1997), random forest (Breiman, 2001), and pattern recognition techniques (such as support vector machines, neural networks and Markov models) (Vapnik, 1998; Burges ,1998) have been successfully applied to high dimensional genomic and proteomic data (Wu et al., 2003; Qu et al., 2002). In their 2003 paper, Wu et al. reviewed and compared the performance of several multivariate methods used for classification and selection of biomarkers. These include both classical discriminant methods such as linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbor classifier, as well as machine learning techniques like bagging and boosting classification trees, support vector machine and random forest. Wu's application of these techniques to

an ovarian cancer Mass spectrometry dataset has shown that, among all these methods, random forest has had better performance in terms of feature selection and sample classification. One remarkable advantage of random forest classifier is the fact that it reports an importance measure for each variable (VIM). These VIM represent internal estimates of the decrease in the classifier's overall accuracy if that particular variable was not used in building the classifier. In this regard, variables with larger importance measures can be deemed to have more classification power (Datta and Depadilla, 2006). Tuglus et al (2008), however, outlines the following pitfalls associated with prediction algorithms such as random forest: there is no guarantee that all biomarkers from a set will receive a measure of importance, there is no formal inference, and therefore, no p-values. Finally, variable importance measures obtained through such prediction algorithms tend to lack interpretability (Tuglus et al, 2008).

1.1.2.3. Causal Inference Framework

To facilitate the discussion, we let the observed data be represented by O= (A, W, Y) where A represents a set of binary biomarker variables, W a vector of covariates, and Y the clinical outcome of interest. Thus, we define the observed data as $X = (Y, W)$, the counterfactual outcomes of interest as Y_a, and the full data as $X^{FULL} = (X_a, a \in A)$, where $a \in \{0,1\}$. Under the general assumption of no unmeasured confounders, known as ignorability of the treatment assignment mechanism (Rubin, 1978), the causal inference framework seeks to answer the following question: what is the causal effect of a given biomarker A on the clinical outcome Y?

One possible measure of impact of a biomarker on the outcome, in the causal inference framework, is the marginal effect at the population level, known also as the average treatment effect. It is estimated as: $E(Y_1) - E(Y_0)$, where Y_1 is the outcome the subject would have had if s/he received treatment, and Y_0 is the outcome the subject would have had if s/he received control. However, in the observed data, instead of (Y_0, Y_1), only one outcome is possible for each subject such that $Y=Y_1A + Y_0(1-A)$,where $A_i=1$ and $Y_i=Y_{1i}$ if subject i was exposed to treatment; and $A_i=0$ and $Y_i=Y_{0i}$ if the subject was not exposed (control). The use of counterfactuals allows one to cast the problem as a missing data issue and thus opens the way to finding an approximation for the potential outcome. Popularized by Rubin (1978, 2004, 2005), the counterfactual paradigm relies on one key assumption: Each subject in the sample has potential outcomes in two states, the one in which the subject is observed and the one in which the subject is not observed. Thus, each subject has in theory two counterfactuals Y_0 and Y_1 (Winship and Morgan, 1999). This potential outcomes framework allows one to estimate the unobservable difference for each subject between outcomes under both conditions. For more detailed technical discussions and applications of counterfactuals, see Robins et al (2000), Greenland and Brumback (2002), Petersen et al. (2006), Gelman and Meng (2004), Holland (1986), Rosenbaum (2002), Rubin (2005), D'Agostino (1998), Sobel (1995), Morgan and Winship (2007), West, Biesanz, and Pitts (2000), Winship and Morgan (1999); and Winship and Sobel (2004).

In randomized studies, treatment assignment is independent of a subject's potential outcomes. Therefore, the difference of the sample averages $E\{Y_1|A=1\}$-

E $\{Y_0|A=0\}$ equals E $\{Y|A=1\}$- E $\{Y|A=0\}$ and is an unbiased estimate of the population average casual effect , $E(Y_1) - E(Y_0)$. In observational studies, exposure is not controlled, thus treatment received may not be independent of potential outcomes. In this case, E $\{Y_1|A=1\}$- E $\{Y_0|A=0\}$ may not be an unbiased estimate of the average treatment causal effect. One way to account for this dependency is to find all important covariates (W) related to both potential outcome and treatment exposure and use them in the estimation of the population average causal effect. The covariates W are chosen such that the potential outcomes Y_0 and Y_1 are independent of A|W. In other words, if W contains all confounders, then among subjects sharing the same W, the potential outcomes (Y_0, Y_1) and A are independent conditional on W (As would be the case in a blocked experiment where the treatment or biomarker A would be randomized within the levels of W). In this case, E{E(Y|A=1, W)}= E{E(Y_1|1, W)}= E{E(Y_1|W)}=E(Y_1) and similarly, E{E(Y|A=0, W)}= E(Y_0), and E(Y_1) – E(Y_0) would be an unbiased estimator of the average causal effect. For a deeper insight into how to choose W, refer to Cole and Hernán (2008), Schafer and Kang (2008), Robins (2001), Hernán et al (2002), Brookhart et al (2006).

Robins has developed a class of models known as marginal structural models (MSM) whose aim is to "replicate the findings of a randomized controlled trial using observational data". (Petersen et al., 2006). These models allow one to estimate the average effect of a treatment or biomarker. Below, we will list the assumptions behind MSMs and we will review three MSM estimators: G-computation, inverse probability of treatment weighting (IPTW), and double robust (DR) estimator (Robins et al., 1998, 2000; Hernán et al., 2001). We will also provide an overview of

a new double robust method known as targeted maximum likelihood estimation. For simplicity, we assume a point-treatment study.

The causal inference methods lie on the following assumptions:

a. Consistency: The data for a subject is simply one of the counterfactual outcomes from the full data. The observed data is $O = (A, X_a)$.

b. Randomization: A $\perp Y_a|$ w, $\forall$ A. In other words, there are no unmeasured confounders for A, which means within strata of W, A is randomized.

c. Experimental Treatment Assignment Assumption (ETA): $P\,(A = a|W) > 0$, for all W.

Of the three above assumptions, only the ETA is verifiable. For a more complete description of these assumptions and their practical applications, we refer to Cole and Hernán (2008) and Cole and Frangakis (2009). Based on these assumptions, the likelihood of the data can be written as $L(O) = P(Y|A,W)P(A|W)$, where p could be a logistic function. While some of the methods described below make assumptions only about $P(A|W)$, others use $P(Y|A,W)$, and some make assumptions about both conditional distributions (i.e. both $P(A|W)$ and $P(Y|A,W)$).

1.1.2.3.1. G-computation

The G-computation (Robins, 1986, 2000) is a method used to estimate counterfactuals means. It assumes that $P(Y|A,W)$ is correctly estimated. It relies on the following general formula: Pr (Y_a=1) = $\int$ Pr$(Y = 1|A = a, W = w)dP(w)$.

For a dichotomous outcome Y, Pr (Y_a=1) = E(Y_a) = $\int E(Y = 1|A = a, W = w)dP(w)$, (E denotes Expectation). The G-Computation estimate of the counterfactual mean in the simple context of point-treatment is then $\hat{E}[Y_a]$= $\sum_1^n \frac{1}{n}\hat{E}$[Y|A=a, W=$w_i$]. This estimate can then be used to derive causal parameters such as risk difference, relative risk, and odds ratio. To estimate the causal risk difference, for instance, one has to first postulate a model for the outcome regression E (Y|A,W), fit the model, and then average the resulting estimates E(Y|A=1, W)- E(Y|A=0, W) over all observed W. Assuming a logistic function, Log $\frac{p(Y=1|A,W)}{p(Y=0|A,W)} = \beta A + \gamma W$, the risk difference is θ = E (Y|A=1, W) - E (Y|A=0,W)= $\frac{e^{(\beta+\gamma W)}}{1+e^{(\beta+\gamma W)}}$ - $\frac{e^{(\gamma W)}}{1+e^{(\gamma W)}}$. Then, the parameter of interest for the average causal effect θ is estimated by plugging the maximum likelihood for the parameters β and γ into the above equation and then averaging over all observed W_i:

$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\{\frac{e^{(\hat{\beta}+\hat{\gamma}W_i)}}{1+e^{(\hat{\beta}+\hat{\gamma}W_i)}} - \frac{e^{(\hat{\gamma}W_i)}}{1+e^{(\hat{\gamma}W_i)}}\}$.

1.1.2.3.2. Inverse Probability Weighting

The inverse probability treatment weighting (IPTW) approach relies on the correct specification of postulated propensity score model for a good performance. We will first present a brief review of the propensity score method followed by a summary of the IPTW methodology.

1.1.2.3.2.1. Propensity Score

Assuming a logistic regression model, A|W~ bin (1, $p_{A|W}$), $\log(\frac{p(Ai=1\,|Wi)}{p(Ai=0\,|Wi)}) = \alpha_0 + \alpha_1 Wi$, the propensity score is defined as $e(W) = P(A = 1|W) = \frac{e^{(\alpha_0 + \alpha_1 W)}}{1 + e^{(\alpha_0 + \alpha_1 W)}}$. In short, it is the conditional probability of assignment to exposure A given a vector of observed covariates (Rosenbaum and Rubin, 1983; Mansson et al., 2007). As can be seen from this equation, it is dependent on the random variables W and has its own probability distribution.

Because observed differences in observational data may reflect underlying differences between groups, it is critical to mitigate bias resulting from the imbalance in covariate distributions. The use of propensity score does improve comparability of exposure groups with regards to measured covariates, and thus decreases bias. Given $e(W)$, W and A are conditionally independent, which balances measured covariates across exposure groups (Rosenbaum and Rubin, 1983;Mansson et al., 2007). In other words, groups with similar distributions of $e(W)$ should have similar distributions of W . Other properties of the propensity scores are as follows:

a) If it is sufficient to adjust for covariates W, then it is sufficient to adjust for their propensity scores $e(W)$ (Joffe and Rosenbaum, 1999);
b) Estimated propensity scores do a better job at removing bias than true propensity scores , because the estimated propensity scores remove both systematic and chance imbalances, while the true propensity score removes only systematic imbalances (Joffe and Rosenbaum, 1999; Cepeda et al., 2003).

Propensity scores have gained in popularity and have been widely used in statistical literature to control for baseline differences. These adjustment methods range from matching (Rosenbaum and Rubin, 1985; Rubin and Thomas, 1996) to regression adjustment (D'Agostino, 1998; Rubin and Thomas, 2000) to weighting (Robins,1997;Robins et al., 2000; Hirano and Imbens, 2001; Sato and Matsuyama, 2003).

The selection of the propensity score model is often achieved through logistic regression. The logistic model $\log(\frac{p(Ai=1\,|Wi)}{p(Ai=0\,|Wi)}) = \alpha_0 + \alpha_1 Wi$ does assume a linear relationship between the response and the covariates. This assumption, however, is not always tenable, especially when some of the covariates are continuous. Moreover, Kang and Schaffer (2007) have shown that logistic regression might not always be a good way to estimate response propensities, and have advocated for the use of more robust procedures, especially in the presence of outliers. One alternative approach to logistic regression has been a modeling framework that estimates a flexible function of the covariates while relaxing the linearity assumption. One such model is the generalized additive model (GAM) for binary dependent variable. Pioneered by Hastie and Tibshirani (1990), the GAM assumes that the mean of the dependent variable depends on an additive predictor through a nonlinear link function, and allows the response probability distribution to be any member of the exponential family of distributions, including logistic model for binary data. It allows for a more flexible relationship between continuous covariates and response by using smoothing techniques (Hastie and Tibshirani, 1990).

While the logistic regression models the logit of the response probability with the linear form $\log(\frac{p(A=1\,|W)}{p(A=0\,|W)}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j W_j$, the logistic additive model replaces this linear predictor with an additive one of the form: $\log(\frac{p(A=1\,|W)}{p(A=0\,|W)}) = \alpha_0 + \sum_{j=1}^{p} f_j\,(W_j)$, where $f_1(.), f_2(.),\ldots,f_p(.)$, are smooth functions that define the additive component. In this setting, the predicted probabilities, or propensity scores, are given by $p(A=1\,|W) = \frac{e^{(\alpha_0 + \sum_{j=1}^{p} f_j(W_j),}}{1+ e^{(\alpha_0 + \sum_{j=1}^{p} f_j(W_j))}}$ and $p(A=0\,|W) = 1 - p(A=1\,|W) = \frac{1}{1+ e^{(\alpha_0 + \sum_{j=1}^{p} f_j(W_j))}}$.

The benefits of using GAM over logistic regression in estimating propensity scores have been demonstrated by Woo and colleagues (2008). Using both simulated and genuine data, they showed how GAMs outperformed logistic regression in improving covariance balance, particularly for higher moments of the covariate distributions.

In the case of continuous A, a flexible parametric approach, proposed by Irano and Imbens (2004) can be used to compute a generalized propensity score. First, one postulates a normal distribution of the continuous biomarker A given the covariates, i.e. $A_i|W_i \sim N\,(\alpha_0 + \beta'_1 W_i\,, \sigma^2)$. Then the parameters $\alpha_0,\ \beta_1, \sigma^2$ are estimated by the least squares regression or by maximum likelihood estimation (MLE). The generalized propensity scores are estimated by plugging in these parameter estimates into the normal density: $\widehat{R}_i = \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} \exp\,(-(\frac{A_i-(\widehat{\alpha}_0+\widehat{\beta}_1 W_i)}{2\widehat{\sigma}^2}))$.

1.1.2.3.2.2. IPTW Methodology

From the likelihood of the data $L(O) = P(Y|A,W)P(A|W)$, the inverse weighting method uses the treatment assignment distribution $P(A|W)$ to create a pseudo-population in which the treatment assignment is no longer confounded (Robins, 1998). Rather than using the difference of simple averages $\mathrm{E}(Y_1) - \mathrm{E}(Y_0)$, this method estimates θ , the average causal effect, by the difference of inverse propensity score weighted averages, i.e. $\hat{\theta}_{IPTW} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{e(W_i,\widehat{\alpha_1})} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-A_i)Y_i}{1-e(W_i,\widehat{\alpha_1})}$, where $e(W, \alpha_1)$ is the postulated propensity score model for the true propensity score.

The first step in the implementation of the IPTW approach is to fit a multivariate regression of the probability of the biomarker A given the covariates W. This model, referred to as the treatment mechanism (Petersen et al., 2006), is used to compute propensity scores. Then, each subject in the sample is assigned a weight *SW* equal to the inverse probability of being in a certain group A_j given their observed covariates W. The higher the probability of a subject being in A_j, the lower the weight assigned to this subject; and vice versa. The end result is a pseudo dataset where the treatment is randomized. Finally, a regression model of the outcome Y on the biomarker A_j with observations weighted by *SW* is fitted to estimate the IPTW parameter. For an in-depth look at the IPTW methodology, see Robins (1986, 1998), Robins et al. (2000), Greenland and Brownback (2002).

This method of weight estimation may suffer from a shortcoming whenever W is strongly associated with A, especially in non-saturated models. In such occurrences, the weights might have large variability. Studies by Kang and Schaffer (2007) have

shown how unstable propensity score weights could result in a poor performance of the IPTW estimator. A solution proposed by Robins has been to use stabilized weights instead. Instead of using 1 as the numerator in the weight computation, it is recommended to use the sample proportion of subjects having $A_j = a$ (where a={0,1} is the set of all potential values of a given biomarker A_j). Denote the stabilized weights by $q_n^*(A,W)$, where $q_n^*(A,W) = \frac{p(Ai=a)}{p(Ai=a\,|Wi)}$. In practice, stabilized weights can be computed following the same basic steps taken by Cole and Hernán (2004). These steps are:

1. Estimate the propensity score model with covariates and obtain the predicted values.
2. Estimate the propensity score model without covariates by fitting an intercept-only model and generate the predicted values.
3. Obtain the ratio of the estimates obtained in (a) and (b). This gives the stabilized weights for each subject: estimates in (b) over estimates in a.

For continuous A, the estimated stabilized weights are computed as a ratio of densities (i.e, $SWi = \frac{f_{ai}}{f_{ai}|Wi}$) as proposed in Robins et al (2000). f_{ai} is the marginal density of the continuous biomarker A, and $f_{ai}|Wi$ is the conditional density of the biomarker A given the set of covariates W. One way to estimate the numerator f_{ai} is to specify a normal distribution (i.e. $A_i \sim N(\alpha_0, \sigma^{*2})$, and then plug the mean $\hat{\alpha}_0$ and the empirical variance $\hat{\sigma}^{*2}$ of the biomarker A values into the normal density. The denominator is estimated based on the generalized propensity score method of Hirano and Imbens (2004) described in the propensity score sub-section above.

1.1.2.3.3. Doubly Robust (DR) Estimator

The DR estimation method combines both the G-computation and the IPTW approaches, by incorporating both $P(Y|A,W)$ and $P(A|W)$. Under this method, the average causal effect is estimated as:

$$\hat{\theta}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{e(W_i,\widehat{\beta})} - \frac{\{A_i - e(W_i,\widehat{\beta})\}}{e(W_i,\widehat{\beta})} m_1(W_i,\widehat{\alpha_1})\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-A_i)Y_i}{1-e(W_i,\widehat{\beta})} + \frac{\{A_i - e(W_i,\widehat{\beta})\}}{1-e(W_i,\widehat{\beta})} m_0(W_i,\widehat{\alpha_0})\right]$$

$$= \hat{\mu}_{1,DR} - \hat{\mu}_{0,DR}$$

In the above equation, $e(W,\beta)$ is the postulated model for the true propensity score (from logistic regression); $m_0(W, a_0)$ and $m_1(W, a_1)$ are the postulated regression models for the true relationship between the vector of covariates and the outcome within each level of treatment A (i.e. E(Y|A=0, W) and E(Y|A=1, W)).

While the G-Computation relies on the consistent estimation of P(Y|A,W) and the IPTW assumes that P(A|W) is correctly specified, the DR method produces consistent estimates as long as either one of the two models (propensity score or outcome regression model) is correctly specified (Robins and Rotnizky, 2005). Because of this property, referred to as double robustness, this approach presents a notable advantage, especially in situations where it might be easier to correctly specify the relationship of the biomarker with covariates or in situations where it might be easier to model the relationship between the clinical outcome and the biomarker and covariates. Moreover, this approach presents a clear alternative whenever concerns linger about the correct specification of either the IPTW model (P(A|W)) or the G-computation model (P(Y|A,W)). For an extensive review of the DR estimator, the reader is directed to Robins and Rotnitzki (1995, 2001), Van der Laan and Robins (2003), Lunceford and Davidian (2004), Carpenter et al. (2006),

Davidian et al. (2005), and to Kang and Schafer (2007). A SAS macro for doubly robust estimation and a companion book chapter by Funk et al. (2010) are available at http://www.unc.edu/~mfunk/dr/.

1.1.2.3.4. Targeted Maximum Likelihood Estimation

In a seminal paper published in 2006, Mark Van der Laan pioneered a new way to establish a ranking of biomarkers. Considered as free from standard model assumptions, this method known as targeted maximum likelihood is employed, among other purpose, to generate a marginal variable importance measure that captures the impact of each biomarker on an outcome (Van der Laan and Rubin, 2006). For a formal and theoretical discussion, refer to Van der Laan and Rubin (2006), Van der Laan, (2005); for empirical examples or applications in biomarker selection, see Bembom et al. (2006, 2008), Tuglus and Van der Laan (2008). It has been shown through simulations studies that this method has good statistical properties (adequate bias-variance tradeoff, efficiency, consistency, robustness) and that that the variable importance measure obtained under the TMLE can be, under certain conditions, a doubly efficient and robust measure that accommodates both low and high dimensional data (Van der Laan and Rubin, 2006; Bembom et al, 2008; Tuglus and Van der Laan, 2008).

The TMLE methodology achieves double robustness by applying both the G-computation and the IPTW models simultaneously. The parameter of interest, which measures the true marginal importance of each biomarker A with regards to the outcome Y, is $\Psi = E_w$[E(Y=1| A_J=1, W) - E(Y=1| A_J=0, W)] (for each biomarker,

indexed by j). In practice, implementation of the targeted maximum likelihood estimation (TMLE) to generate a targeted measure of variable importance (TVIM) involves adding a covariate *h(A,W)* (stretching function) to an initial regression model denoted by Q_0(A,W), and then averaging the regression over the covariates for fixed value of A (Van der Laan and Rubin, 2006). In the repeated measures setting, targeted maximum likelihood estimation of the variable importance measure takes time into account. More specifically, for a time-varying outcome Y(t), the parameter of interest is given by: Ψ(t) ≡ Ew [E (Y(t)/Aj =1, Wj) - E (Y(t)/A j=0, Wj) for discrete A. Incorporating time in the objective function allows one to estimate the impact of each biomarker on the time trajectory (Bembon et al., 2006).

Simulation studies and application of the TMLE to real data have generated promising results. Analyses performed by Tuglus et al. (2008) have shown TMLE has generated a list of differentially expressed genes in patients with acute lymphoblastic leukemia and acute myeloid leukemia, that shows greater biological plausibility than results obtained from univariate least squares regression, penalized least squares regression (Least Absolute Shrinkage and Selection Operator (LASSO) regression), and random forest. Likewise, Bembon et al. (2008) applied the TMLE methodology to data from the Stanford Drug Resistance Database to generate measures of impact for a set of candidate genetic mutations with regards to their importance in conferring resistance to the protease inhibitor drug Lopinavir. In this analysis, the ranking of genetic mutations based on the TMLE methodology was in best agreement with current medical knowledge, as compared to results from univariate least squares regression and G-computation.

1.1.2.3.5. Classifier Performance Assessed by ROC Curve

As a tool for assessing diagnostic accuracy, the Receiver-Operating Characteristic (ROC) curve has received a great deal of attention in the statistics literature (Begg, 1991; Hanley, 1989; Faraggi and Reier, 2002). Originally used in the signal detection theory developed in the 1950s (Green et al., 1996), this technique has been, over the years, extended to a variety of research fields including radiology (Obuchowski, 2005; Hanley, 1998; laboratory testing (Obuchowski et al.,2004; Zweig and Campbell, 1993), epidemiology (Pepe, 2000, 2003; Pepe and Janes, 2008; Baker, 2003; Pencina and D'Agostino, 2004), bioinformatics and machine learning (Li and Fine,2008; Provost and Faucett, 2001; Lasko et al., 2005, Kjetil, 2009), and countless other clinical disciplines (Zheng et al., 2006; Zou et al.,2007; Musial et al., 2003, Cheun et al.,2001).

Let X be a binary test result and D the outcome (disease or not). Condition on disease status, two basic measures of performance for a binary test are sensitivity and specificity (Pepe, 1983). Sensitivity is defined the fraction of subjects with disease that a diagnostic test correctly identifies as positive (true positive fraction (TPF), i.e. $P[Y = 1|D = 1]$) while specificity is the fraction of subjects without the disease that the test correctly identifies as negative (true negative faction, i.e. $P[Y = 0|D = 0]$). The quantity 1 – specificity is called false positive fraction (FPF) and represents the probability of a positive test given that disease is not present (i.e. $P[Y = 1|D = 0]$).

The ROC curve generalizes the notions of FPF and TPF to continuous tests X. Assuming that a test is classified as positive if Y is above a threshold c , then TPF(c) =

P[Y >c| D = 1] and FPF(c) = P[Y >c| D = 0], and ROC(.) = {(FPF(c), TPF(c)), c ϵ $(-\infty, \infty)$} (Pepe, 2000). The ROC curve is a plot of the test sensitivity (TPF) along the y axis versus its 1-specificity (FPF) along the x axis, for all possible threshold values c (Heagerty at al., 2000). In mathematical terms, it can be written as ROC_t = $F_D(F_{\bar{D}}^{-1}(t))$, t ∈ (0,1) where $F_D$and $F_{\bar{D}}$ are survivor functions for the test result in the diseased (cases) and non-diseased (controls) populations. It is a monotone increasing function in (0,1) that can be estimated both parametrically and non-parametrically (Hanley et al., 1982a & 1982b; Zou et al., 1996; Hanley, 1988; Metz, 1978). More complete assessments of the performance, advantages or disadvantages of either parametric or non parametric estimation methods of the ROC curve are available in Hajian-Tilaki et al. (1997), Goddard (1989), and in Faraggi and Reiser (2000).

1.1.2.3.5.1. Area Under the ROC Curve and Partial Area Under the ROC Curve

The most commonly used summary ROC Index is the area under the ROC curve (AUC). In general, AUC=$\int_0^1 ROC(t)dt$, AUC ∈ [0,1]. It is interpreted as the probability that the test result from a randomly chosen diseased individual is more indicative of disease than that from a randomly chosen non diseased individual (Pepe, 2000). As a rule of thumb, a more convex ROC translates in a better classifier, and so does a higher AUC. At AUC=1, the discrimination /accuracy of the classifier is deemed perfect. An AUC=0.5 reflects pure noise conditions, thus an uninformative test (Kjetil, 2009; Pepe , 2003). Both parametric estimation through

the binormal model (Swets and Pickett, 1982) and nonparametric estimation through the Wilcoxon Mann-Whitney U-statistic (Bamber, 1975; Hanley and McNeil, 1982a & 1982b) are available for the AUC statistic.

One notable pitfall often attributed to the AUC measure is its lack of clinical relevancy (Dodd and Pepe, 2003). As Obuchowski (2005) points out, the ROC curve extends well beyond the clinically relevant area of potential clinical interpretation. More often than not, the interest lies in only a fraction of the ROC space that corresponds to clinically relevant values of test specificity and sensitivity. For instance, in using CD4 cell count level as a determining factor for starting HAART therapy, a low false positive rate may be desirable, thereby making the lower tail of the ROC curve the region of interest. This is one of the cases where it may not make sense clinically to look at the whole ROC curve. In response to situations like this, a measure with greater clinical appeal, that considers only regions of interest in the ROC space, has been developed. It is called partial AUC (McClish, 1989; Thompson and Zucchini, 1989) and it focuses on a limited range of false positive rates. The partial AUC does also provide notable benefits when two ROC curves cross. Analyses by Zhang et al. (2002) and by Fawcett (2006) show that partial ROC analysis tends to provide more information and helps better with clinical decision making in cases of two crossed ROC curves. For false positive rates (FPR) $\in$ (0, t) for some $t < 1$, the partial area under the curve is defined as pAUC=$\int_0^t ROC(t)dt$ and can be estimated both parametrically and non-parametrically. For partial AUC estimation methods, inference, and statistical properties of the estimator, see McClish (1989), Wieand et

al. (1989), Zhang et al. (2002), Pepe (2003), Dodd and Pepe (2003), Janes et al. (2005), Cai and Dodd (2008).

1.1.2.3.5.2. Incorporation of Covariates Information in ROC Curves

While the ROC curve and the AUC have been extensively studied in statistical literature and are widely used to assess classifier performance, there has been until relatively recently a gap in the search for standards methods of incorporating covariate information in ROC curves. This topic has now received increased attention, and for practical reasons: Covariate adjustment can help eliminate potential confounding (Huang and Pepe, 2009). Janes and colleagues (2006, 2007, 2008) argue that without covariate adjustment, ROC curves can be differentially biased, which can lead to faulty marker comparisons.

A discussion of different uses of covariates in ROC analysis can be found in Janes and Pepe (2007, 2008a, 2008b, 2008c) and in Janes, Longton, and Pepe (2008). These authors have made a clear distinction between covariate adjustment and other related uses of covariates. Adjustment is recommended when the covariate W affects the distribution of the marker among controls. A relevant measure of classification accuracy in this case would be the covariate-adjusted ROC Curve (AROC), a stratified measure of ROC performance (Janes and Pepe, 2006; Janes and pepe, 2007; Janes, Longton, and Pepe, 2008). Procedures for deriving the *A*ROC and other related metrics of biomarker comparison such as the area under the adjusted ROC curve (*A*AUC) are provided in Janes, Longton, and Pepe (2008).

In situations where the covariates W affect the separation between case and control distributions (i.e. affect discrimination), Janes and colleagues recommend ROC regression as a way to dealing with covariates. Two examples of covariates that fall in this category are disease severity and specimen storage time. In the ROC regression setting, the end result is an estimate of the ROC curve as a function of covariates (covariate-specific ROC curve).

The third approach to incorporating covariate information in ROC analysis is the incremental value estimation. This approach takes hold when the covariates W are a set of risk factors or other baseline predictors (Janes, Longton, and Pepe, 2008). For these factors that contribute to discrimination, Janes et al. (2008) recommend combining the biomarker and covariate information and determining the incremental value of the biomarker beyond and above the covariates. Typically, two models containing both the covariates, but with and without the biomarker are fitted. ROC curves comparison is then made for the linear predictions from the two models. A similar approach has been implemented in the analysis of the Atherosclerosis Risk in Communities Study to determine, among a panel of 19 novel biomarkers, those with the biggest increase in AUC for CHD prediction above and beyond a set of covariates (Folsom and Chambless, 2006). Additional examples of implementation of this method to assess discrimination of biomarkers above and beyond that of classic cardiovascular disease risk factors can be found in Danesh et al (2004), Koenig et al (2004), Pepe et al (2004), Ricker et al (2002), Rutter et al (2004), Shlipak et al (2005), Van der Meer et al (2003), and Wilson et al (2005).

1.1.3. Remaining Statistical Challenges

The methods summarized above do have their merits, but they are not are not exempt from shortcomings. Despite all the advances in the statistical analysis of biomarker data, some critical challenges remain.

a. The t-test does not control for confounding and tends to lack robustness, especially in high throughput data (Tuglus et al., 2008; Yu et al., 2006). While classical multivariable regression methods provide an analytical framework where all biomarkers can be evaluated simultaneously with or without covariate adjustment, they tend to be unstable when multicollinearity exists. Furthermore, they are prone to model misspecification, may not even be feasible when the data are high dimensional, and ultimately, may lead to biased estimates of the variable importance measures (Tuglus, 2008). As for the most commonly used summary Index of the ROC curve, the Area Under the ROC curve (AUC), it can lack clinical relevancy (Dodd and Pepe, 2003; Obuchowski, 2005). Finally, non-linear models and machine learning techniques, while considered as valuable tools for biomarker selection and classification in high throughput datasets, tend to be complex to non sophisticated users and are generally computationally intensive (Levy et al., 2005).
b. There is no unifying framework for biomarker selection. Different ranking features applied to the same data often generate different rankings of biomarkers. For instance, Dutkowski and Gambi (2007) used a proteomic mass spectrometry dataset to evaluate several feature selection methods and

ended up with different feature lists. This confirms earlier findings by Levy et al. (2005). In all methods, the goal and expectation should be to generate a reliable list of the top-ranked candidates that are significantly associated with the outcome of interest. Having different lists can be counterproductive and constitutes an impediment to the efficient use of biomarkers. Thus, the question of creating, through methods with good statistical properties, unified biomarker lists for further biologic examination or for subsequent statistical assessment of surrogacy, remains an open one. To increase the use and utility of biomarkers in drug development and public health research, these statistical issues need to be addressed.

c. Methods based on the causal inference framework, and more specifically the targeted maximum likelihood estimation for variable importance measure, are promising approaches that are worth exploring. However, the TMLE, as a novel method, has not been widely applied in public health and pharmaceutical research. To our knowledge, there has been no systematic comparison between this method and other statistical approaches in their ability to select and rank biomarkers in different settings: cross-sectional, longitudinal, and time to events. Furthermore, there have been no established guidelines for data collection and assessment based on this method.

This dissertation aims at filling these research gaps by addressing the core issue of biomarker selection in the presence of covariates. It is worth noting that most of the methods mentioned above are used for biomarker screening where the goal may

be to find all biomarkers that are statistically significant. This research does go one step further by applying methodologies aimed at identifying, from a set of candidate biomarkers, the ones that are the most important in terms of their contributions to a clinical outcome. In short, the aims here are beyond biomarker screening and encompass variable importance assessment and ranking.

1.1.4. Multiplicity Considerations in Biomarker Research

In biomarker studies, large number of hypothesis tests are often conducted to identify candidates associated with an outcome. This gives rise to a multiple hypothesis testing problem. Suppose that m independent tests are conducted, the probability of at least one false positive result is 1 - $(1-\alpha)^m$ and converges to 1 as m increases. For instance, this probability jumps from 0.226 to 0.994 if the number of tests m increases from 5 to 100. In microarray gene expression experiments, for instance, thousands of genes are often examined. With this high number of simultaneous hypotheses tests (one for each gene), the probability of obtaining at least one false positive result is near certainty. In such cases, one needs to adjust for multiple testing.

The goal of multiple testing is to minimize the type I error while maximizing power. Traditional multiple comparison procedures such as the Bonferroni correction impose a penalty for multiple testing. This penalty can, however, be too stringent (Devlin et al., 2003). As the number of test increases, traditional adjustment methods such as the Bonferroni procedure become powerless. A competing approach to multiple testing, that is more powerful, more liberal, and that is now widely used, is based instead on controlling the false discovery rate (FDR). With this

new approach that allows a reasonable number of false discoveries, the goal shifts from controlling the family-wise error rate to keeping in check the expected proportion of false discoveries (Benjamini and Hochberg, 1995, 2000).

Suppose that m is the total number of null hypothesis tested (H_{01}, H_{02},..., H_{0m}). Furthermore, let p_1, p_2,..., p_m be the p-values obtained from those tests. Denote by F the number of false positives (i.e. type I error), by T the number of true positives, and by S the total number of rejections (i.e. F + T) and define Q= $\begin{cases} \frac{F}{S} \; when \; S > 0 \\ 0 \; otherwise \end{cases}$. The false discovery rate (FDR) is E(Q), or expectation of proportion of type I errors among all rejections. The Benjamini and Hochberg method controls the FDR at level $\frac{m_0}{m}$q ≤ q and works as follows:

a. Order the m p-values $p_{(1)}$, $p_{(2)}$,..., $p_{(m)}$from smallest to largest, and order the corresponding hypotheses: $H_{0(1)}$, $H_{0(2)}$,..., $H_{0(m)}$.
b. Set a threshold value for rejection by finding the largest integer i such that $p_i \leq i\,q/m$, i.e. k= max{ i: $p_{(i)} \leq \frac{i}{m}q$}. If no such integer i exists, then no hypothesis is rejected.
c. Reject any hypothesis with a p-value ≤ p_i.

The Benjamini and Hochberg FDR-controlling step-up procedure works on the assumption of independence. It further assumes that the true null hypotheses p-values are uniform (0,1) random variables under the null hypotheses. It has been, however, demonstrated in literature that the Benjamini and Hochberg procedure does control the FDR under some dependency structures (namely positive

dependency) and covers many problems of general interest (Benjamini and Yuketieli, 2001). For a greater range of dependency problems, a simple modification of the Benjamini and Hochberg procedure has been proposed by Benjamini and Yekutieli (2001) to control the FDR. Known as the Benjamini and Yekutieli FDR controlling procedure, this method sets the threshold for rejection at: k= max{ i: $p_{(i)} \leq \frac{i}{m} q^*$}, where $q^* = \frac{q}{\sum_{i=1}^{m} \frac{1}{i}}$. It always controls the FDR at level less than or equal to $\frac{m_0}{m}$q (Benjamini and Yekutieli, 2001).

For in-depth discussions and comparisons of multiple comparison procedures, including FDR, see Benjamini and Hochberg (1995, 2000), Yekutieli and Benjamini (1999), Benjamini and Yekutieli (2001), Genovese and Wasserman (2002, 2003), Storey (2002, 2003), Storey and Tibshirani (2001, 2003), Finner and Roters (2002), Dmitrienko et al. (2005), Dudoit, Shaffer, and Boldrick (2003), Westfall et al. (1999), Brown and Russell (1997), and Pollard, Dudoit, and Van der Laan (2004).

1.1.5. Specific Aims of the Research

The past decade has seen an explosion in the availability and use of biomarkers data as a result of innovative discoveries in areas such as combinatorial chemistry, mass spectrometry, high throughput screening, DNA microarrays, and proteomics. This has been accompanied by a growing emergence of biomarkers as a topic of clinical research. Publications identified by the keyword “biomarker” in Pub Med from 1999 through 2008 have increased dramatically and reached a peak of 37,000 in 2007 (Wagner, 2009). With that explosion of data come new challenges. A good

fraction of biomarkers data are high-dimensional and do not lend themselves to standard statistical methods. Also, due to the large number of candidate biomarkers for a given condition, selection of the ones with maximal impact has become a critical issue. An effective biomarker discovery streamlining process could help save time and previous resources by directing researchers' focus on the best candidates. On the other hand, choosing the wrong candidates could lead to incorrect decision-making about potentially effective agents. Major advances in biomarker discovery underscore the need for novel statistical methods, especially in the area of biomarker selection. Even though several statistical methods have been proposed as solutions to the biomarker selection problem, there are still some major hurdles. So far, there is no unified approach for biomarker selection. Different methods tend to generate different results. Some of the newest methods, although promising, have had limited use in clinical research and lack clearly-established guidelines for sample size calculation and power analysis in studies involving biomarker selection. To increase the use and utility of biomarkers in drug development and public health research, it is imperative that these statistical issues be addressed.

The proposed dissertation project seeks to fill this gap by addressing the core issue of biomarker selection in the presence of covariates. Its primary aim is to examine the effectiveness of three novel statistical methods in identifying biomarkers with good performance characteristics. It also seeks to provide guidelines for data collection and assessment, such as sample size computation and power analysis, in studies involving biomarker selection where these methods are

employed. It uses real and simulated data to apply these innovative statistical methods to the concrete issue of biomarker selection in the context of HIV infection.

The specific objectives of this research are:

1. To estimate a marginal variable importance measure (VIM) separately for each biomarker that determines the clinical outcome CD4 cell count using the following estimation methods: Targeted Maximum Likelihood, Binary Regression based on flexible propensity score estimation, Incremental Value Estimation based on partial Receiver Operating Characteristic (ROC) Curve methodology, and weighted Cox proportional hazards model.
2. To use variable importance measures computed in (1) to make inference about the importance of each biomarker. This would help determine whether there exist, among all the biomarkers considered, more affordable alternatives that can accurately predict CD4 cell count.
3. To develop an index that represents a longitudinal measure of the importance of each biomarker over time.
4. To establish guidelines for future data collection and assessment, based on results from simulations.

These specific aims have been addressed in three separate papers. The content of each paper is presented in the next three (3) chapters.

CHAPTER 2

Comparison of the effectiveness of three novel statistical methods for biomarker selection with application to an HIV infection dataset.

2.1. Introduction

The importance of biomarkers both in the drug development process and in public health practice is well established. Fueled by recent advances in modern biology and technology, biomarkers have become a popular research topic in clinical investigations. Publications identified by the keyword "biomarker" in PubMed from 1999 through 2008 have increased dramatically and reached a peak of 37,000 in 2007 (Wagner, 2009). With that explosion of data come new challenges. An overarching aim is to find ways to use this wealth of biomarker information to help guide clinical decision making. Therefore, the development of improved statistical methods that can adequately explain the relationship between biomarkers and an outcome, is of great interest. One domain that is evolving in this regard concerns the quest for variable importance measures, a class of estimators that could reliably capture the effect of a specific biomarker on a clinical outcome. Such estimators are used in the identification, among many candidate biomarkers, of the best subset that is significantly associated with an outcome of interest. This could help reduce waste and time by directing biologists' focus to top performing biomarkers, or by allowing practitioners to direct resources towards the most promising candidate biomarkers.

This paper addresses the core issue of biomarker selection in the presence of covariates, especially when the goal is to identify biomarkers with good performance characteristics from among a large number of candidates. This research is the first of its kind to compare the performance of three novel statistical methods of biomarker selection and then use these estimators to address a major public health issue: the relationship of CD4 cell counts with other biomarkers in HIV-infected patients. The contribution of this paper is enhanced by the fact that it evaluates the impact of finite sample size on the performance of these estimators.

The outline of the paper is as follows. Section 2.2 provides an appraisal of current statistical methods for biomarker selection. Section 2.3 discusses three novel methods for biomarker selection, while section 2.4 applies these methods to an HIV infection dataset. Section 2.5 presents a Monte Carlo simulation to examine the behavior of the three methods under different sample sizes. Section 2.6 concludes and provides suggestions for further research.

2.2. Current Statistical Methods for Biomarker Selection

A survey of statistical methods used for biomarker selection reveals that both standard and novel statistical methods have been employed to address the challenges of biomarker selection. The panoply of methods used in this regard includes the t-test (Dudoit el al., 2002); classical multivariable regression techniques such as ordinary least squares and logistic regression; the Receiver-Operating Characteristic (ROC) curve (Pepe, 2000, 2003); and non-linear models and machine learning techniques such as classification and regression trees (Breiman et al., 1984), bagging (Breiman, 1996), boosting (Freund and Schapire,

1997), random forest (Breiman, 2001), and pattern recognition techniques (Vapnik, 1998; Burges ,1998).

All these methods suffer from shortcomings. The t-test does not control for confounding and tends to lack robustness, especially in high throughput data (Tuglus et al., 2008; Yu et al., 2006). While they provide an analytical framework where multiple biomarkers can be evaluated simultaneously, classical multivariable regression methods tend to be unstable when multicollinearity exists, are prone to model misspecification, and may not even be feasible when the data are high dimensional. The most commonly used summary index of the ROC curve, the Area Under the ROC curve (AUC) lacks clinical relevancy (Dodd and Pepe, 2003; Obuchowski, 2005). Finally, non-linear models and machine learning techniques, while considered as valuable tools for biomarker selection and classification in high dimensional data, tend to be complex to non-sophisticated users and are generally computationally intensive (Levy et al., 2005). Furthermore, there seems to be no unifying framework for biomarker selection. Different ranking features applied to the same data often generate contrasting rankings of biomarkers. For instance, Dutkowski and Gambi (2007) used a proteomic mass spectrometry dataset to evaluate several feature selection methods and ended up with different feature lists. This confirms earlier findings by Levy et al. (2005). In all methods, the goal and expectation should be to generate a reliable list of the top-ranked candidates that are associated with the outcome of interest. Having different lists can be counterproductive and constitutes an impediment to the efficient use of biomarkers. Thus, the question of creating, through methods with good statistical properties,

unified biomarker lists for further biologic examination or for subsequent statistical assessment of surrogacy, remains an open one. To increase the use and utility of biomarkers in drug development and public health research, these statistical issues need to be addressed.

An additional framework under which biomarkers are often evaluated is the causal inference paradigm. Under the assumptions of consistency, randomization, and experimental treatment assignment (Cole and Hernán, 2008; Cole and Frangakis, 2009), causal models enable researchers to estimate, among other measures, an average effect of a biomarker, which under certain conditions could carry a causal interpretation. Three of the most commonly used estimation techniques under causal inference approaches are G-computation (Robins, 1986, 2000), inverse probability of treatment weighting (IPTW) (Robins et al., 2000; Cole and Hernan, 2008) and the double robust estimator (Van der Laan and Robins, 2003). Two of the methods used in this paper originate from this framework.

2.3. Materials and Methods

2.3.1. Study sample

The dataset used for application in this paper came from the Hormonal Contraception and HIV Genital Shedding and Disease Progression or GS Study. The GS Study is a prospective multicenter study of 306 HIV infected women aged 18 to 45 years old from Uganda and Zimbabwe. This study started in 2001 as an add-on to the HC-HIV study (Morrison et al., 2007) and was completed in the field in December 2009. Women who seroconverted during the course of the HC-HIV study were recruited for the GS study, based on procedures outlined in Morrison et al.

(2010). The study specific objectives are described in details elsewhere (Morrison et al., 2007, 2010), but one key research question is the effect of hormonal contraception on the biological parameters of the infectivity of women with primary and chronic HIV infection to their sex partners.

The GS consisted of a baseline visit and follow-up visits at 2, 4, 8 and 12 weeks following HIV seroconversion, and then every 12 weeks for up to 9 years. Women who developed severe HIV infection or who had successive CD4 cell counts at or below 200 cells per mm^3 were offered highly active antiretroviral therapy (HAART) and were seen twice a month for the first month, then monthly thereafter. In addition to baseline demographic characteristics, at each time point, information on various laboratory parameters, reproductive variables, contraceptive exposure, and recent sexual behavior was collected. Gathered laboratory data included HIV plasma viral load, HIV sub-type, CD4, CD8 and total lymphocyte counts, serum chemistries, lipid profiles, specimens for the detection of chlamydial, gonococcal, syphilis, *herpes simplex virus* 2 (HSV-2), and human papillomavirus (HPV) infections. The study also collected information on hormonal contraceptive use, HIV disease progression parameters, as well as virologic, immunologic, and clinical responses to HAART among hormonal and non-hormonal contraceptive method users. For a detailed description of the study population and procedures, the reader is directed to Morrison et al. (2010).

2.3.2. Data structure

To facilitate the discussion of the methods, we let the observed data be represented by O= (A, W, Y) where A represents a set of either binary or continuous

biomarker variables, W a vector of covariates, and Y the clinical outcome of interest (Y=1 for diseased and Y=0 for non-diseased). Define the observed data as $X = (Y, W)$, the counterfactual outcomes of interest as Y_a, and the full data as $X^{FULL} = (X_a, a \in A)$, where a=(0,1) (binary case) or a ϵ $\mathcal{A}$=$\{a_1, a_2, \ldots, a_n\}$ (continuous case).

The set of covariates (W) used in this analysis consist mainly of behavioral, reproductive, and demographic factors reported in the literature to be predictive of incident HIV infection or clinically associated with HIV disease progression (Van Der Pol et al., 2008). These covariates included age, country, commercial sex work status, number of coital acts in previous 3 months, condom use consistency, study subject's partner's sexual behavior and risk, frequency of nights away from home by study subject's partner, history of sexually transmitted infections (STI), presence of STI symptoms at enrollment, having more than one sex partner, and breastfeeding. The vector of biomarkers A contained measures such as plasma viral load, HIV sub-type, HIV RNA load, hemoglobin level, CD4, CD8 and lymphocyte counts, CD4/CD8 T cell ratio, CD4 percentage, and HSV-2 status. The latest biomarkers and covariates measurements available at 6 months from estimated date of HIV infection were used in this analysis. While HIV sub-type and HSV-2 status were binary variables, all other biomarkers used in this analysis were measured on a continuous scale.

In this analysis, we defined baseline as the latest biomarker measurement or covariate value available 6 months after estimated infection date. The estimated infection date refers to the mid-point between the last visit where a subject was HIV-uninfected in the HC-HIV study and the first visit where this subject was confirmed infected. Based on the timing implied by this definition of baseline, a number of

subjects might not have baseline data because of the length of time elapsed between the date they were notified of their HIV infection and their first GS-enrollment visit (where specimens for plasma and cervical viral loads were collected). For these subjects, the GS enrollment visit might have occurred more than 6 months after the estimated infection date. Thus, subjects with missing baseline biomarker information because of the timing of their first GS visit did not contribute data to the analysis.

The clinical outcome (Y) was a binary variable representing two successive drops of CD4 cell count at or below 350 cells/mm^3 in the first two years following the viral set point (i.e. 121 days following the estimated infection date). The choice of the threshold of 350 cells per mm^3 for the clinical outcome of CD4 was based on current guidelines for the use of antiretroviral agents in adults and adolescents infected with HIV-1 in the absence of an AIDS-defining illness (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2008). Also, evidence from the literature suggests that initiation of antiretroviral therapy before the CD4 cell count has fallen below 350 cells per mm^3 significantly improves survival, as compared to deferred therapy (Sax and Baden, 2009; Kitahata et al., 2009; When to Start Consortium, 2009).

2.3.3. Measures of Effect

We implemented three methods to derive a marginal variable importance measure for each biomarker used from the GS Study. These methods were: targeted maximum likelihood estimation (TMLE), propensity score weighting (PSW), and incremental value estimation for partial area under the ROC curve.

Under the TMLE and the PSW methodologies, the measure of effect for each binary biomarker was the difference in probabilities between those with and without the biomarker exposure of interest (A=1 vs A=0), averaged over the entire population:

$\Psi = E_w$[E(Y=1| A_J=1, W) - E(Y=1| A_J=0, W)].

In the continuous case, the parameter of interest was:

$\Psi = E_w$[E(Y=1| A_J=a, W) - E(Y=1| A_J=$\bar{a}$, W)], where $\bar{a}$ is the empirical mean of the biomarker A_J.

Under the incremental value scheme, the measure of variable importance was the incremental value, i.e. the amount of discriminatory accuracy of the biomarker A_J over and above the covariates (W). Essentially, we estimated the optimal difference in partial AUC between a model with only covariates and a model with both the biomarker A_J and covariates, based on the following non-parametric estimator proposed by Dodd and Pepe (2003):

$\text{pAUC}_{(t0,t1)} = \frac{1}{mn}\sum_1^m \sum_1^n I(S_1 > S_0, S_0 \epsilon (q_1, q_0))$, (where (q_1, q_0) are sample quantiles; m and n represent the sample sizes from non-diseased sample S_0 and from diseased sample S_1).

To generate standard errors for the estimates of the VIM measures obtained under each method, bootstrapping (Efron and Tibshirani, 1994) was implemented. We applied the following algorithm for bootstrap selection and for assessing multiple comparison:

1. From the original sample S, we drew B independent bootstrap samples $S_1, S_2 \ldots, S_B$, each of size n, assuming simple random sampling with replacement.
2. For each bootstrap sample, we then computed the sample quantity of interest $\hat{\theta}_b$ This resulted into B values of the statistic $\hat{\theta}_n$.
3. The bootstrap estimator of the parameter θ was given as the mean of the bootstrap estimates $\overline{\hat{\theta}_b}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$, with variance $\hat{V}(\hat{\theta}_b) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \overline{\hat{\theta}_b})^2$. The standard deviation of the distribution of the statistic $\hat{\theta}_b$ was then: $\sqrt{\hat{V}(\hat{\theta}_b)}$ $= \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \overline{\hat{\theta}_b})^2}$.
4. We subsequently assessed the strength of the evidence for a non-zero effect by conducting the following test of hypothesis: $H_0: \hat{\theta}_b = 0$ versus $H_A: \hat{\theta}_b \neq 0$, for each j (j=1,..., k).
5. We repeated these steps k times, that is once for each biomarker A_j, (J = 1,.....k to generate k estimates of θb and k p-values denoted as $P_1, P_2, \ldots, P_k$.
6. Finally, we adjusted the p-values for multiplicity testing to control the false discovery rate. The Benjamini and Yekutieli (2001) False Discovery Rate (FDR) controlling procedure was used to account for the dependence of the test statistics. Significance for each biomarker was assessed by comparing related adjusted p-value to the 0.05 alpha level. A lower the p-value denoted a better measure of importance.

   Below, we present an overview of the three procedures.

2.3.4. Targeted Maximum Likelihood for Variable Importance Measure

Developed by Van der Laan (2006), the targeted maximum likelihood estimation methodology is employed, among other purposes, to generate a marginal variable importance measure that captures the impact of a given biomarker on an outcome. For a formal and theoretical discussion, refer to Van der Laan and Rubin (2006), Van der Laan (2005); for empirical examples or applications in biomarker selection, see Bembom et al. (2008), Tuglus and Van der Laan (2008). Under certain conditions, the variable importance measure obtained under the TMLE can be a doubly efficient and robust measure (Van der Laan and Rubin, 2006).

We followed a three-step approach to implement this methodology. First, we modeled the conditional distribution of Y given A and W using the following logistic mean model: $Logit(\Pr(Y_{ij} = 1)) = \beta_0 + \beta_1 A_i + \sum_{j=2}^{9} \beta_{ij} W_{ij}$, where A_i represents the target biomarker variable, and W_{ij} the vector of covariates listed in section IV. From this outcome model, we generated fitted values denoted as $\widehat{Q_0}$ (A,W).

Next, we estimated the conditional distribution of the biomarker given the covariates. The predictors were the same set of covariates used in the outcome model. For binary biomarkers (A), the estimates of the probabilities $p(A_J=1|\ W)$ and $p(A_J=0|\ W)$, denoted by g_n^0 (A, W), A=0 or 1, were computed and then used to calculate a covariate h(A,W) as follows: $h(A,W) = \left(\frac{I(A=1)}{g_n^0(1,W)} - \frac{I(A=0)}{g_n^0(0,W)}\right)$, (where I denotes an indicator function). For continuous biomarker (A), we followed Tuglus et al. (2008) and defined the covariate h(A, W) as: $h(A,W) = A - E(A|W)$.

The next stage in applying the TMLE entailed updating the initial estimate $\widehat{Q_0}$(A,W) with the covariate h(A,W) by regressing the binary outcome Y on the covariate *h(A,W)* in an intercept-only model where the initial estimates $\widehat{Q_0}$ (A,W) were held fixed. Specifically, the updated estimate Q_1(A,W) is given by Q_1(A,W)=(A,W)= Q_0(A,W) + $(\varepsilon) * h(A,W)$. The regression coefficient ε is obtained through maximum likelihood estimation (MLE) using $\widehat{Q_0}$ (A,W) as offset. From this regression, we obtained the maximum likelihood estimate $\hat{\epsilon}$ for the covariate h(A,W) and used $\hat{\epsilon}$ to update the initial estimate of Q_0(A,W) such that $\widehat{Q_1}$(A,W)= $\widehat{Q_0}$ (A,W) + $(\hat{\epsilon}) * h(A,W)$. Finally, we computed the targeted estimate of the marginal variable importance measure of interest, for each binary biomarker, by evaluating $\widehat{Q_1}$(A,W) at both A_J=1 and A_J=0 for each individual i, and then averaging over all i. For continuous biomarkers, $\widehat{Q_1}$(A,W) was evaluated at both A_J=a and A_J= $\bar{a}$. These steps were repeated for each biomarker.

2.3.5. Propensity Score Weighting for Variable Importance Measure

As in the TMLE, this method uses a counterfactual structure and the same two nuisance parameters: P(Y|A,W) and P(A|W). The difference, however, lies in the algorithm we executed to update the initial estimates of P(Y|A,W). While Van der Laan et al. (2006) add a covariate created from P(A|W) to the initial regression model, we used weights constructed from estimated propensity scores P(A|W) to incorporate the relationship between A and W in the estimation of P(Y|A,W).

The generation of variable importance measures through the Propensity Score Weighting method was performed in three steps, as described below. In the first stage, we estimated the propensity scores. For binary biomarkers, we modeled P(A|W) using the generalized additive framework (Hastie and Tibshirani, 1990) for a more flexible relationship between continuous covariates and response (Woo et al., 2008). Assuming a logistic additive model of the form $\log(\frac{p(A=1\,|W)}{p(A=0\,|W)}) = \alpha_0 + \sum_{j=1}^{p} f_j\,(W_j)$, where $f_1(.), f_2(.),\ldots,f_p(.)$, are smooth functions that defined the additive component, we computed the following expressions of probability:

$$p(A=1\,|W) = \frac{e^{(\alpha_0 + \sum_{j=1}^{p} f_j(W_j),}}{1+ e^{(\alpha_0 + \sum_{j=1}^{p} f_j(W_j))}} \text{ and } p(A=0\,|W) = \frac{1}{1+ e^{(\alpha_0 + \sum_{j=1}^{p} f_j(W_j))}}.$$

For continuous A, a flexible parametric approach, proposed by Irano and Imbens (2004) was used to compute a generalized propensity score. First, we postulated a normal distribution of the continuous biomarker (A) given the covariates, i.e. $A_i|W_i \sim N\,(\beta_0 + \beta'_1 W_i\,, \sigma^2)$. Then the parameters β_0, β_1, σ^2 were estimated by least squares regression. Following Hirano and Imbens (2004), we estimated the generalized propensity scores by inserting the estimates of these parameters into the normal density: $\widehat{R_i} = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\,(-(\frac{A_i-(\hat{\beta}_0+\hat{\beta}_1 W_i)}{2\hat{\sigma}^2}))$.

In the second stage of this approach, we created the weights. Stabilized weights for binary biomarkers A were computed as $q_n^*\,(A,W) = \frac{p(Ai=a)}{p(Ai=a\,|Wi)}$. For continuous biomarker variables, the estimated stabilized weights were constructed as a ratio of densities (i.e, $SWi = \frac{f_{ai}}{f_{ai}|Wi}$) as proposed in Robins et al. (2000); f_{ai} is the marginal density of the continuous biomarker (A), and $f_{ai}|Wi$ is the conditional density of the

biomarker (A) given the set of covariates (W). To estimate the numerator f_{ai} we specified a normal distribution (i.e. $A_i \sim N(\alpha_0, \sigma^{*2})$, and then substituted the mean $\hat{\alpha}_0$ and the empirical variance $\hat{\sigma}^{*2}$ of the biomarker (A) values into the normal density. The denominator was estimated based on the generalized propensity score method of Hirano and Imbens (2004) described above.

Finally, the weights were incorporated into a logistic regression model to generate adjusted estimates. Separate estimates were created for each biomarker, and each logistic model contained binary CD4 as a dependent variable and the biomarker of interest and covariates as predictors. All covariates used in the model were believed to be related to the outcome.

2.3.6. Incremental Value Estimation for Partial Area Under the ROC Curve (pAUC)

This method used the partial Area under the Receiver Characteristic Operating (ROC) Curve (AUC) methodology (McClish, 1989) while incorporating covariates in the analysis. The goal was to assess the ability of each biomarker to discriminate above and beyond the set of covariates.

Briefly, this method compared ROC curves for each combination of a given biomarker A_J and the covariates (W) to the ROC curve for the covariates (W) alone (McIntosh and Pepe, 2002; Janes and Pepe 2008; Janes, Longton and Pepe, 2008). Each comparison assessed the amount of improvement in classification accuracy (incremental AUC) generated by adding the biomarker A_J to the covariates (W). The

basic assumption was that the covariates (W) contribute as well to discrimination between those with and without the outcome.

Because the full AUC measure tends to lack clinical relevancy (Dodd and Pepe, 2003; Obuchowski, 2005), we conducted this analysis at a false positive rate of 5% based on reports that the most common false positive rates of viral loads measurements often vary from 3% to 10% (Mendoza et al., 1998). In terms of implementation, this procedure was accomplished in two steps: First, for each biomarker A_J, we estimated p(Y=1|A, W) and p(Y=1|W). Using the predicted probabilities from the two fitted logistic models, we computed estimates of sensitivity and specificity over the restricted false positive range of 5%, and generated an index summary for the partial ROC Curve, referred to as pAUC. Finally, the measure of interest, the difference between the pAUC for the two models at hand, was computed.

2.4. Data Analysis

In the GS Study, the median age at enrollment was 27 years. Women from Zimbabwe accounted for 58.5% of the population while those from Uganda made up the remaining 41.5%. About 8% of all study subjects had at least two sex partners while 14% had a STI history, 8% were breastfeeding, and 45% displayed STI symptoms. These women averaged 11.2 (Standard Deviation [SD] = 15.5) sex acts per month, but only 35% of them reported consistent condom use. On average, the partners of these women spent 10 (SD=15.2) nights away from home, and 75% of those partners had been reported to have had sex with another woman in the three

months prior to enrollment in the study. Finally, 59% of the subjects' partners met the study definition for primary partner risk, a composite variable that included having a partner with HIV, urethral discharge, weight loss, nights spent away from home, or a history of sex with female sex workers.

The proportion of subjects with biomarker data available anytime from enrollment to 6 months after estimated infection date was as follows: 59% of the 306 women had baseline CD4, CD8, and CD4/CD8 T cell ratio. The distribution of baseline CD4 cell counts was different in the two study countries. On average, patients in Uganda had a higher level of baseline CD4 cell counts (650.72 [SD=237.35]) than those in Zimbabwe (532.97 [SD=207.97]). In addition, 79% of the study population had plasma viral load measurements while 56% had lymphocyte counts data. Finally, 58% of the 306 women had hemoglobin information, 98% had HSV data, and 97% had HIV sub-type information. This proportion included 57 Zimbabweans imputed as subtype C based on the fact that all Zimbabweans with available subtype information were subtype C. Overall, the majority of subjects were subtype C (59%), followed by subtype A (27%), and subtype D (11%). Covariate information was available for all 306 subjects. Subjects with missing baseline data for any biomarker were excluded from the analysis for that particular biomarker. In addition, 23 subjects were removed from the analysis sample because their baseline CD4 cell counts were below 350 cells/ mm^3.

For each biomarker and under each method, we computed a variable importance measure and used bootstrapping for inference. We conducted a separate hypothesis test of no effect for each biomarker. In each case, the resulting p-value

was adjusted for multiplicity based on the Benjamini & Yekutieli dependent false discovery rate. Results under both the TMLE and the PSW did not support this hypothesis, as a number of biomarkers were deemed to have had a significant impact on the outcome (Table 1). Based on the magnitude of the p-values, the most important biomarkers under the TMLE methodology, among the 11 biomarkers considered, were baseline CD4 cell count and CD4/CD8 T cell ratio. Under the PSW approach, the biomarkers selected as the most important ones were: HSV-2 status, CD4/CD8 T cell ratio, baseline CD4 cell count, and Plasma Viral Load. No biomarker was selected as important by the incremental value method.

2.5. Simulation Studies

We conducted simulation studies to evaluate the finite sample performance of the three proposed estimators. The goal was to test the ability of each of these estimators to identify "true" biomarker variables significantly related to an outcome. The simulated dataset consisted of a binary outcome Y, a 3-dimensional vector of continuous baseline covariates W=(W1, W2, W3), and a 5-dimensional vector of biomarkers A=(A1, A2, A3, A4, A5). These variables were generated based on the following setup:

a. The biomarkers were randomly assigned as: Pr(A1)=0.2, Pr(A2)=0.45, Pr(A3)=0.35, Pr(A4)=0.4, and Pr(A5)=0.5.
b. Each baseline covariates followed a normal distribution: W1 ~(5,2), W2 ~(6,1.5), W3~N(0,1).
c. The following model was postulated for the outcome:

$P(Y = 1|A, W) = g^{-1}(0.1 + 1.05A1 + 4.5A2 + 4A3 + 0.0002A4 + 0.1028A5 - 1.5W1 + 1.2W2 - 1.8W3)$, With $g^{-1}(.)$ the inverse logit function, Aj the jth biomarker, and Wk the kth baseline covariate. A larger coefficient for Aj did correspond to a larger effect on the outcome Y. In this context, A2 and A3 exerted a larger effect, while A1 had a moderate effect.

Predicted probabilities generated from this model were compared against a uniform (0,1) random variable to create a binary outcome variable. For any record where the random variate was less than the predicted probability, a value of 1 was assigned to the outcome variable; else the outcome was 0.

d. For each biomarker (A) and under each method, the simulation was run 5000 times on increasing sample sizes of N=100, 150, 200, 250, and 1000.

2.5.1. Simulation Results

The simulation results (Tables 2-4) showed that all three estimators performed better with increasing sample sizes. At N < 200, both the TMLE and the Propensity Score Weighting method picked up a single biomarker (A2) as significant. These two methods, although lacking power at that sample size level, did outperform the incremental value approach, which failed to detect any significant result. At N ≥ 200, all three methods correctly picked up biomarkers A2 and A3 as significant. This improvement in performance over increasing sample sizes is in accordance with previous simulation studies that assessed the finite sample properties of causal inference estimators such as G-Computation, IPTW, and Double robust estimators (Neugebauer and Van der Laan, 2005). As for the ROC methodology, simulations by

Janes and Pepe (2006) did show a trend towards an increasing power of both full and partial of ROC curves in the presence of covariates as the sample size increases. Finally, it is worth noting that at N ≥ 1000 (results not shown), the weighted logistic method displayed slightly increased power as the proportion of rejected tests was higher than that of the other methods (three significant biomarkers detected by this approach as opposed to two biomarkers picked up by the other methods). Based on its coefficient, the biomarker A3 could be considered to be moderately associated with the outcome Y, and the PSW method was powerful enough to detect this meaningful association.

2.6. Discussion and Conclusion

We compared three methods for estimating biomarker variable importance measures. Our simulation results suggest that PSW works well in small sample sizes (say N> 100), but may be anticonservative when sample size is large (say, N> 1000). These results further indicate that TMLE could be a robust method that performs reasonably well in moderate to large sample size (say N≥150). Finally, the incremental value approach displays an unsatisfactory ability in detecting significant biomarkers when the sample size is less than 200, but works satisfactorily with sample sizes exceeding 200.

From a public health perspective, this research is relevant for the following reasons. It enabled us to identify from the GS study potentially useful candidate biomarkers based on their true importance with regards to the clinical outcome of CD4 cell count. As a measure of a patient's immune capacity, CD4 cell count has been considered as a standard method for determining eligibility for HAART and

HIV disease progression (Ellenberg, 1991; Fleming, 1994). However, CD4 count measurements are expensive and could be prohibitive in resource-poor countries. If, as our results indicate, baseline CD4 cell counts are highly predictive of future CD4 cell count level, one strategy to both contain costs and save lives might be to obtain one initial CD4 measure when infection is discovered, rather than having to do it repeatedly over time. Furthermore, the list of biomarkers identified as important predictors of CD4 cell counts in this research, while in line with current medical knowledge, could also pave the way for the use of simpler and relatively less expensive biomarkers (e.g. CD4/CD8 T cell ratio) that are highly predictive of CD4 cell decline and disease progression. This could support decision-making with regards to HAART initiation or could help monitor patients' immune status during therapy without having to make additional expensive CD4 measurements.

If is worth pointing out that the use of baseline CD4 to predict future levels of CD4 cell counts could raise the issue of circularity. However, from a clinical standpoint, this could be a reasonable exercise because evaluation of baseline CD4 cell counts could help identify patients at risk for CD4 cells depletion so that they could be monitored more closely and started on HAART, when necessary. This could potentially help save lives, time, and money, especially in resources-deprived countries where the costs to measure CD4 are often prohibitive.

This study does have limitations. The two best performing methods (TMLE, PSW) rely on the assumption of no unmeasured confounders, i.e. within strata of covariates (W), the target biomarker (A) is randomized. The thinking is that if the vector of covariates (W) contains all confounders, then among subjects sharing the

same W, the potential outcomes (Y_0, Y_1) and the biomarker (A) would be independent conditional on W (as would be the case in a blocked experiment where the treatment or biomarker (A) would be randomized within the levels of W) (Cole and Hernán, 2008; Schafer and Kang, 2008). Given that the GS Study was observational in nature, exposure was not controlled, thus "treatment" (i.e. biomarker) received might not be independent of potential outcomes. In this case, the difference E $\{Y_1|A=a\}$ - E $\{Y_0|A=0\}$ might not be an unbiased estimate of the average treatment effect, as would be the case in a randomized study. To account for this dependency, we tried to find all important covariates (W) believed to be related to both potential outcomes and exposure (A) based on the literature or expert knowledge, and included them in the estimation of the population average effect. There is no direct way to verify whether there remained any putative confounders that were not part of the vector of covariates (W) used in this study.

Sample size could be another limitation of this study. Over half of the biomarkers under consideration had 200 or fewer non-missing observations, with the minimum being 138. As shown in our simulation studies, sample size does affect the ability of all three methods to detect "true" significant biomarkers; all three showed a decreased ability in pinpointing significant biomarkers at smaller sample sizes (e.g. N=100). Hence, sample size constraints may have hindered us from detecting additional significant biomarkers in the GS dataset. The implication for applications is that, with small sample sizes, these methods may not achieve adequate power to detect the effect of a given biomarker. Thus, strong consideration should be given to sample size and power issues when designing biomarker studies using these analytic

methods. The sample size cut-points we identified through our simulations could serve as a starting point towards establishing sample size requirement guidelines for future data collection and assessment in biomarker studies where these selection methods are used.

In summary, this study shows a promising practical application of both targeted maximum likelihood estimation and propensity score weighting to biomarker selection from observational studies. Nonetheless, the list of significant biomarkers obtained varies with sample sizes, as demonstrated by our simulations. For more conclusive results, future investigations, especially those involving the incremental value approach, should employ a much larger sample size. Furthermore, we recommend that repeated measures analysis of longitudinal data be conducted to capture the trends and various dimensions of the CD4 count clinical outcome in its relationships with other HIV infection biomarkers.

CHAPTER 3

Application of Longitudinal Targeted Variable Importance Measures (LTVIM) to Biomarker Selection from an HIV Infection Dataset

3.1. Introduction

The past decade has seen an explosion in the availability and use of biomarkers data as a result of innovative discoveries and recent development of new biological and molecular techniques. Biomarkers are essential for at least four key purposes in biomedical research and public health practice: They are used for disease detection, diagnosis, prognosis, to identify patients who are most likely to benefit from selected therapies, and to guide clinical decision making. Determining the predictive and diagnostic value of these biomarkers, singly and in combination, is essential to their being used effectively, and this has spurred the development of new statistical methodology to assess the relationship between biomarkers and clinical outcomes. One such method aims at identifying biomarkers with good performance characteristics by computing a marginal variable importance measure (VIM) for each biomarker from a set, using the theory of targeted maximum likelihood estimation (TMLE) (Van der Laan, 2006).

Much of the application of the targeted variable importance measure (TVIM) in literature deals with the simple case of point-treatment. There is currently no unified approach to addressing VIM in the context of repeated measures data, even

though numerous studies dealing with biomarkers collect longitudinal data. In this paper, we propose a novel approach to computing longitudinal VIM, when the interest lies in obtaining an estimate of the impact of a biomarker on the time course of a given clinical outcome. This method extends the VIM computation based on the TMLE methodology to repeated measures longitudinal data, while taking advantage of the flexibility of a nonparametric smoothing technique. The end result is an index that represents the strength of evidence for the importance of a biomarker with regards to a clinical outcome measured over time. Inference for this estimator is readily available through bootstrapping.

3.2. Statement of the problem

Often in biomarker studies, researchers are faced with the task of evaluating the impact of multiple biomarkers on a given outcome. Given a large set of biomarkers that potentially predict a clinical outcome, how can one make a determination as to which ones are the most important? Answers to this question are of great practical importance to public health practice and pharmaceutical research. Statistical methodologies that allow researchers to take a pool of biomarkers and make a reliable appraisal as to which ones are the most important, could lead to faster decision making in epidemiologic studies and clinical trials. For instance, an effective and efficient selection of the best subset of biomarkers amongst a set could help direct further biological research in early phases of clinical trials by limiting the focus to the pool of most promising ones. From a scientific standpoint, this could guide statistical research in the area of biomarker validation by making

available a list of good candidates for surrogacy status determination. Such a list could also be useful in the quest for the best combinations of biomarkers.

In this study, we address the issue of biomarker selection in the context of longitudinal repeated measures data. The outline of the paper is as follows. Sections 3.3 and 3.4 review the theory of targeted maximum likelihood estimation and provide the implementation steps for two estimators of interest. Section 3.5 describes an HIV infection dataset used to illustrate these estimators while section 3.6 reports on the analysis and results. Section 3.7 covers a Monte Carlo simulation to study the behavior of these two estimators under different numbers of repeated measures. Finally, section 3.8 provides a discussion of the results and ideas for further research.

3.3. Targeted Maximum Likelihood for Variable Importance Measure

To facilitate the discussion of the methods, we let the observed data be represented by O= (A, W, Y(t)) where A represents a set of either binary or continuous biomarker variables, W a vector of covariates, and Y(t), a time-varying clinical outcome. Define the observed data as $X(t) = (Y(t), W)$, the counterfactual outcomes of interest as Y_a, and the full data as $X^{FULL} = (X_a, \text{a} \in \text{A})$, where a=(0,1) (binary case) or A ϵ $\mathcal{A}$=$\{a_{1t}, a_{2t}, \dots, a_{nt}\}$ (continuous case). For simplicity, we assume that both A and W are static.

In order to select the most important biomarkers (A) affecting the outcome (Y(t)), we used a tool named targeted maximum likelihood estimation (Van der Laan

and Rubin, 2006). The TMLE is a versatile method suitable for parameter estimation in semi-parametric and nonparametric models from either randomized or observational studies. It could be used in a variety of settings: cross-sectional, repeated measures longitudinal, and time-to-events. In this study, we applied this technique to generate a marginal variable importance measure that captures the impact of each biomarker on the clinical outcome.

Our choice of this estimation method is motivated by its attractive statistical properties: adequate bias-variance tradeoff, efficiency, consistency, robustness. Theoretical and simulation studies have shown that the TMLE provides higher precision and reliability than other methods such as random forests, neural networks, least angle regression, univariate regression (Tuglus et al., 2008). Furthermore, the TMLE estimator could be, under certain conditions, a doubly robust measure that could also be free from standard regression model assumptions. By using information from two conditional distributions, namely the outcome model (e.g. E[Y(t)|A,W]), E denotes expectation), and the treatment mechanism (e.g (E[A|W]), the TMLE produces consistent estimates as long as one of these two nuisance parameters is correctly estimated (Bembom et al, 2008; Rosenblum and van der Laan, 2010). Because of this property, referred to as double robustness, the TMLE presents a notable advantage, especially in situations where it might be easier to correctly specify the relationship of the biomarker with covariates or in situations where it may be easier to model the relationship between the clinical outcome and the biomarker and covariates. Overall, this approach presents a clear advantage over

conventional estimation methods whenever concerns linger about the correct specification of either E[Y(t)|A,W]) or E[A|W].

For a formal and theoretical discussion of the TMLE, the interested reader is directed to Van der Laan and Rubin (2006), Van der Laan et al. (2009). Empirical examples or applications of the TMLE in biomarker selection are given in Bembom et al. (2008), Tuglus and Van der Laan (2008), Moore and Van der Laan (2007), and in Rosenblaum and Van der Laan (2010).

3.3.1. Measure of Effect

In this study, we pick as a meaningful measure of effect, for each biomarker, the targeted marginal mean. It is defined as the difference in probabilities between those with and without the biomarker exposure of interest (A=1 vs A=0), averaged over the entire population, at time t, i.e Ψ(t)= E_w[E(Y(t)| A_J=1, W) - E(Y(t)| A_J=0, W)]. In the continuous case, the parameter of interest was given by Ψ(t)= E_w[E(Y(t)| A_J=a, W) - E(Y(t)| A_J=$\bar{A}$, W)], ($\bar{A}$ refers to the empirical means of the biomarker measurements).

The TMLE is done in the context of potential outcomes or counterfactual framework. Essentially, each subject in the sample is assumed to have potential outcomes in two states, the one in which the subject is observed and the one in which the subject is not observed (Winship and Morgan, 1999). This potential outcomes framework allows one to estimate the unobservable difference for each subject between outcomes under both conditions. For more detailed technical discussions

and applications of counterfactuals, see Robins et al (2000), Winship and Morgan (1999); Winship and Sobel (2004).

3.3.2. TMLE Implementation

We followed a three-step approach to implement the TMLE methodology. First, we applied the generalized estimating equations (GEE) framework (Liang and Zieger, 1986; Zeger and Liang, 1986) to model the conditional distribution of Y(t) given A and W, i.e. $P(Y(t)|A,W)$. Note that, in this first step, the risk of model misspecification could be mitigated by using a variety or combination of techniques, including data adaptive procedures, to arrive at a suitable functional form. For example, when Y(t) is binary, we fit the model g[E(Y(t)|W, A)] $= \beta_0 + \beta_1 A + \sum_1^k \beta_K wk$, where g [.] is a logistic function. Predicted probabilities from this outcome model were denoted by $\widehat{Q_0}$ (A,W). Next, we estimated the conditional distribution of the biomarker given the covariates, $E(A|W)$. For binary biomarkers A, the estimates of the probabilities $p(A_J$=1| W) and $p(A_J$=0| W), denoted by g_n^0 (A, W), were used to calculate a specific covariate h(A,W) as follows: $h(A,W) = (\frac{I(A=1)}{g_n^0\,(1,W)} - \frac{I(A=0)}{g_n^0\,(0,W)})$. For continuous biomarker (A), we followed Tuglus et al. (2008) and defined the covariate h(A, W) as: $h(A,W) = A - E(A|W)$.

The next stage in applying the TMLE entailed updating the initial estimate $\widehat{Q_0}$(A,W) with the covariate h(A,W) by regressing the binary outcome Y on the covariate *h(A,W)* in an intercept-only model where the initial estimates $\widehat{Q_0}$ (A,W) were held fixed. Specifically, the updated estimate Q_1(A,W) is given by

Q_1(A,W)=(A,W)= Q_0(A,W) + (ε) $* h(A,W)$. The regression coefficient ε is obtained through maximum likelihood estimation (MLE) using $\widehat{Q_0}$ (A,W) as offset. From this regression, we obtained the maximum likelihood estimate $\hat{\epsilon}$ for the covariate h(A,W) and used $\hat{\epsilon}$ to update the initial estimate of Q_0(A,W) such that $\widehat{Q_1}$(A,W)= $\widehat{Q_0}$ (A,W) + ($\hat{\epsilon}$) $* h(A,W)$. Finally, we computed the targeted estimate of the marginal variable importance measure of interest, for each binary biomarker, by evaluating $\widehat{Q_1}$(A,W) at both A_J=1 and A_J=0 for each individual i, and then averaging over all i. For continuous biomarkers, $\widehat{Q_1}$(A,W) was evaluated at both A_J=a and A_J= $\bar{A}$. These steps were repeated for each biomarker, and the end result was an estimate of the variable importance of each biomarker at each time point, i.e. Ψ(t) at time t=0, 1, 2,...n.

3.4. Longitudinal Summary Index Measures

Our goal is to provide a summary measure over time instead of a visit by visit analysis. In the next section, we consider two estimators for computing a scalar value denoting the importance of a given biomarker over time. We refer to this index as longitudinal targeted variable importance measure (LTVIM).

3.4.1. Time Slope from Regression through the Origin

Assuming that the targeted variable importance measure (TVIM), as applied above, is close to 0 at time 0, this approach consists in fitting a no-intercept regression model with the TVIM as dependent variable and time as independent variable (Bembom et al, 2006). Since the data are made of a sequence of VIM data points at successive times, the regression errors may not be independent of each

other through time. To account for autocorrelation in the residual series, we fit a model where the errors are assumed to follow the first order autoregressive process. This model postulates an autocorrelation that diminishes rapidly as the distance between the times points increases. It takes the form: $TVIM_t = \alpha_1 T + \mathcal{V}_t$, where the random residual $\mathcal{V}_t = \varepsilon_t$-$\varphi_1 \mathcal{V}_{t-1}$, $\varepsilon_t \sim IN\ (0, \sigma^2)$, and $0 < \varphi_1 < 1$. In this model, ε_t (a residual called white noise) is assumed to be uncorrelated with other residual components, and φ_1 determines the sign and strength of the autocorrelation. The use of the subscript t emphasizes the fact that the data are taken over time.

Estimation of the parameters of the model (α_1 and φ_1) is performed through maximum likelihood, and the measure of impact for each biomarker is given by $\hat{\alpha}_{1,}$. A biomarker with a positive impact over time would have a positive coefficient α_1 while a negative sign of the coefficient would translate a negative effect. Inference can be made by using bootstrapping to construct standard error for the test of the null hypothesis: $\alpha_j = 0$ for each biomarker.

3.4.2. Area under the LOESS Curve.

Consistent estimate of the slope parameter in the auto-regressive model above relies on proper specification of the deterministic component (i.e. E($TVIM_t$), E denotes expectation), and of the residual component $\mathcal{V}_t$. Furthermore the white noise residual ε_t is assumed to satisfy all the classical assumptions (normality, independence, homoscedasticity). Those assumptions may not always be tenable. For instance, the true relationship between TVIM and time may be curvilinear. In

such case, having a straight line forced through the origin may not always provide a good approximation to the mean response E(TVIM).

One way to relax the assumptions about how the TVIM changes as a function of time would be to implement data-adaptive techniques. One such approach proposed in Bembom et al (2006) for fitting the regression is the Deletion/Substitution/Addition (D/S/A) algorithm (Sinisi, 2004). In this paper, we propose an alternative approach based on nonparametric regression. It consists in plotting the estimated variable importance, Ψ(t), as a function of time, using the locally weighted scatterplot smoothing technique (LOESS) (Cleveland, 1979; Cleveland and Devlin, 1988). In short, this method works by moving a window along the time-axis, computes a fitted value at each position in the window and then connects the predicted values to generate the LOESS curve. This method is highly flexible as it requires no specification of a parametric model. The only inputs needed are a smoothing parameter (usually between 0 and 1) and the degree of the local polynomial (usually 1 or 2) to be fitted to the data.

Using the predicted values from the LOESS function, we compute, as our longitudinal summary measure of variable importance, the area enclosed by the LOESS curve and the time axis. Based on the composite Simpson's rule for numerical integration, the index measure of interest is given by:

$\psi = \int_a^b f(x)dx = \frac{h}{3}\left(f(a) + f(b)\right) + \frac{2h}{3}\sum_{k=1}^{m-1} f(x_{2k}) + \frac{4h}{3}\sum_{k=1}^{m} f(x_{2k-1})$, where the time interval [a, b] is subdivided into 2m subintervals $\{[x_{k-1,}\, x_k\,]\}_{k=1}^{2m}$ of equal width $h = \frac{b-a}{2m}$. Because the TVIM could be negative, we take as the final measure of

impact the absolute value of the area under the LOESS curve, which means that a TVIM of -0.5 is equivalent to a TVIM of 0.5. What matters here is the magnitude of the effect, not its direction.

3.4.3. Bootstrapping Algorithm and Multiple Comparison

To generate standard errors for the estimates of the LTVIM measures obtained under each method, bootstrapping (Efron and Tibshirani, 1994) was implemented. We applied the following algorithm for bootstrap selection and for assessing multiple comparison:

1. From the original sample S, we drew B independent bootstrap samples $S_1, S_2 \ldots, S_B$, each of size n, assuming simple random sampling with replacement.
2. For each bootstrap sample, we then computed the sample quantity of interest $\hat{\theta}_b$ (Under the Regression through the origin method, $\hat{\theta}_b$ refers to the estimated time slope; in the LOESS method, $\hat{\theta}_b$ is the estimated area under the LOESS curve). This resulted into B values of the statistic $\hat{\theta}_n$.
3. The bootstrap estimator of the parameter θ was given as the mean of the bootstrap estimates $\overline{\hat{\theta}_b}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$, with variance $\hat{V}(\hat{\theta}_b) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \overline{\hat{\theta}_b})^2$. The standard deviation of the distribution of the statistic $\hat{\theta}_b$ was then: $\sqrt{\hat{V}(\hat{\theta}_b)} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \overline{\hat{\theta}_b})^2}$.

4. We subsequently assessed the strength of the evidence for a non-zero effect by conducting the following test of hypothesis: $H_0: \hat{\theta}_b = 0$ versus $H_A: \hat{\theta}_b \neq 0$, for each j (j=1,..., k). In the LOESS-based method, we conducted a one-tailed test of hypothesis: $H_0: \hat{\theta}_b = 0$ versus $H_A: \hat{\theta}_b > 0$.
5. We repeated these steps k times, that is once for each biomarker A_j, (J = 1,.....k to generate k estimates of θb and k p-values denoted as $P_1, P_2, \ldots, P_k$.
6. Finally, we adjusted the p-values for multiplicity testing to control the false discovery rate. The Benjamini and Yekutieli (2001) False Discovery Rate (FDR) controlling procedure was used to account for the dependence of the test statistics. Significance for each biomarker was assessed by comparing related adjusted p-value to the 0.05 alpha level.

3.5. Genital Shedding and HIV Infection (GS) Data Description

The dataset used for application in this paper came from the Hormonal Contraception and HIV Genital Shedding and Disease Progression or GS Study. The GS Study is a prospective multicenter study of 306 HIV infected women aged 18 to 45 years old from Uganda and Zimbabwe. This study started in 2001 as an add-on to the HC-HIV study (Morrison et al., 2007) and was completed in the field in December 2009. Women who seroconverted during the course of the HC-HIV study were recruited for the GS study, based on procedures outlined in Morrison et al. (2010). The study specific objectives are described in details elsewhere (Morrison et al., 2007, 2010), but one key research question is the effect of hormonal

contraception on the biological parameters of the infectivity of women with primary and chronic HIV infection to their sex partners.

The GS Study consisted of a baseline visit and follow-up visits at 2, 4, 8 and 12 following HIV seroconversion, and then every 12 weeks for up to 9 years. Women who developed severe HIV infection or who had successive CD4 cell counts at or below 200 cells per mm^3 were offered highly active antiretroviral therapy (HAART) and were seen twice a month for the first month, then monthly thereafter. In addition to baseline demographic characteristics, information on various laboratory parameters, reproductive variables, contraceptive exposure, and recent sexual behavior was collected at each study visit. Gathered laboratory data included HIV plasma viral load, HIV sub-type, CD4, CD8 and total lymphocyte counts, serum chemistries, lipid profile, and specimens for the detection of chlamydial, gonococcal, syphilis, *herpes simplex virus 2* (HSV-2), and Human papillomavirus (HPV) infections. The study also collected information on hormonal contraceptive use, HIV disease progression parameters, as well as virologic, immunologic, and clinical response to HAART among hormonal and non-hormonal contraceptive method users. For a detailed description of the study population and procedures, the reader is directed to Morrison et al. (2010).

3.5.1. Outcome Definition

In this analysis, the binary clinical outcome, Y(t), was defined as a drop in CD4 cell counts below 350 cells/mm^3 at time t. The choice of the threshold of 350 cells per mm^3 for the clinical outcome of CD4 was based on current guidelines for the use of antiretroviral agents in adults and adolescents infected with HIV-1 in the

absence of an AIDS-defining illness (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2008; Hammer et al., 2008). Also, evidence from the literature suggests that initiation of antiretroviral therapy before the CD4 cell counts falls below 350 cells per mm^3 significantly improves survival, as compared to deferred therapy (Sax and Baden, 2009; Kitahata et al., 2009; When to Start Consortium, 2009).

3.5.2. Biomarker Variables and Covariates

The set of covariates (W) used in this analysis consist mainly of behavioral, reproductive, and demographic factors reported in the literature to be predictive of incident HIV infection or clinically associated with HIV disease progression (Van Der Pol et al., 2008). These covariates included age, country, number of coital acts in previous 3 months, condom use consistency, study subject's partner's sexual behavior and risk, frequency of nights away from home by study subject's partner, history of sexually transmitted infections (STI), presence of STI symptoms at enrollment, having more than one sex partner, and breastfeeding. The vector of biomarkers A contained measures such as plasma viral load, HIV sub-type, hemoglobin level, CD4, CD8 and lymphocyte counts, CD4/CD8 T cell ratio, CD4 percentage, HSV-2 status. While HIV sub-type and HSV-2 status were binary variables, all other biomarkers used in this analysis were measured on a continuous scale.

3.6. Data Analysis

We carried out a set of separate analyses to estimate the effect of each biomarker on the mean outcome over time, adjusting for covariates. First, we implemented the TMLE methodology to produce a TVIM at each time point. As described above, we generated two nuisance parameters, namely E[Y(t)|A,W] and E(A|W). To model E(Y(t)|A,W), we specified the following logistic mean model: $Logit(\Pr(Y_{ij}=1)) = \beta_0 + \beta_1 A_i + \beta_2 Time_{ij} + \beta_3(A_i)(Time_{ij}) + \sum_{j=4}^{11}\beta_j W_{ij}$, where A_i represents the target biomarker variable, $Time_{ij}$ the measurement occasion for subject i at time j, and W_{ij}, the vector of covariates listed in section 4. We further specified the variance as $\mathrm{Var}(Y_{ij})=V(\mu_{ij})\phi = \mu_{ij}(1-\mu_{ij})$, where $\phi = 1$ and $\mu_{ij} = \frac{\exp(\boldsymbol{X_{ij}\beta})}{1+\exp(\boldsymbol{X_{ij}\beta})}$. Finally, we assumed that the correlation between measurements Y_{ij} and Y_{ik} taken on subject i at times t_{ij} and t_{ik}had an exchangeable structure.

From this model, we extracted the predicted values, which were subsequently updated with a covariate constructed from another nuisance parameter, E(A|W). In our application, the systematic component (η_i) that described the effect of the covariates (W) on the expected value of the biomarker (A) was given by: $\eta_i = g(\mu_i) = W'_i\beta = \beta_0 + \sum_{j=1}^{8}\beta_{ij}W_{ij}$, where g(.) was assumed to be a logit link for binary biomarkers ($0 < \mu_i < 1$), and an identity link for continuous biomarkers. The W_{ij}'s are all listed in section 4.

Once the TVIM was obtained, it was regressed over time to obtain the longitudinal measure of importance based on the regression through the origin method. We also took a second approach to analyzing the data by using smoothing

to highlight trends and patterns in the data. For each biomarker, we fit a smooth curve to the pair (TVIM, TIME) to produce a smooth estimate of the function, which is then used to compute the area under the LOESS curve based on a smoothing parameter of 2/3 and a local polynomial of degree 1. To mitigate the influence of outliers, we used the "symmetric" family option of the LOESS function from the STATS package in R (version 2.7.2). This option combines the local fitting with a robustness step that downweights the relatively large residuals from the fitted curve. This robust smoothing allows for a better extraction of signal from noise.

3.6.1. Results

The GS study sample included 306 subjects followed up to 9 years. Of those, 23 subjects were excluded from the analysis because their baseline CD4 cell counts were below 350 cells/mm^3. Any subject who was given HAART therapy was censored at the time of HAART initiation and contributed data to the analysis up to the most recent date preceding HAART initiation.

For each biomarker and under each method, we computed a variable importance measure and used bootstrapping for inference. We conducted a separate hypothesis test of no effect for each biomarker. In each case, the resulting p-value was adjusted for multiplicity based on the Benjamini & Yekutieli dependent false discovery rate. Results obtained under the two methods under consideration are reported in Table 5. Among the 11 biomarkers, the ones selected as the most important based on the magnitude of the p-values, are baseline CD4 cell counts, HIV

subtype, and HSV-2 status. This list was consistent across the two analytic methods highlighted in this paper.

A useful piece of information that could be extracted from a visual inspection of the LOESS plots in Figures 2 & 3 is an assessment of the longevity of the biomarkers in terms of their predictive power. For instance the predictive power of baseline CD4 cell counts reached its peak around visit 15 and then the LOESS curve displayed a steady decline, suggesting a slowing of the impact of this biomarker on the clinical outcome over time. For a number of biomarkers (i.e. subtype A and D, CD4 percent, CD4/CD8 T-cell ratio, HSV-2), the downward pattern started early and continued over the entire study period. Finally, there was no detectable relationship between VIM and time for CD8 cell counts, hemoglobin, HIV RNA, lymphocyte counts, and Hemoglobin level.. For these biomarkers, the LOESS smoother relating VIM trend to time was mostly a flat line aroud 0 (figure 2).

3.7. Simulation

We conducted a simulation to assess the performance of the two proposed estimators. The ultimate goal was to test the ability of each of the two estimators to identify "true" biomarker variables significantly related to a given outcome, based on different scenarios for the number of data points (i.e. number of repeated measures in the study design). Because we were primarily interested in evaluating how the two methods performed in selecting biomarkers based on the relationship between a continuous TVIM variable and time, we created a simulated dataset that contained

only these two variables. Thus, the scope of this simulation was limited in testing and quantifying the effect of the independent variable time on VIM.

We postulated the following regression model to describe the relationship between time and VIM: $Y = \beta_0 + \beta_1 * Time + \beta_2 * Time^2 + \varepsilon$, Where $\varepsilon \sim N(0, 0.32)$. We assigned the following values to the regression coefficients: β_0=4.17, β_1=0.06, and β_2=-0.002. The rationale for the choice of a model with curvature was two-fold. First, most real-life applications do not involve a straight line going through the origin, and may entail some degree of curvature. Second, such a model would allow us to assess the robustness of the first order autoregressive model with respect to violation of the linearity assumption.

Simulated data points were generated according to the number of measurements taken over the span of the study, and there was one VIM observation per measurement. Thus, if 10 measurements were taken on each subject, then the number of observations used in the regression model would be 10, regardless of the number of subjects, because the VIM, as defined above, is a summary measure of the importance of a given biomarker at each time point. We then evaluated the impact of time on VIM based on the following scenarios for the number of measurements: 5, 10, 20, and 25. In each case, we computed the two estimators, based on 5000 bootstrapped estimates. The results are presented in Table 6.

3.7.1. Simulation Results

At n=5, both methods failed to detect a significant relationship between time and TVIM. The LOESS-based method performs satisfactorily for sample sizes >=10,

while the autoregressive model detects significant relationship between time and VIM only when n >=20. These results indicate that the intensity of the effect of the biomarker, as measured by decreasing p-values, is better captured with increasing number of repeated measurements.

3.8. Discussion and Conclusion

In this paper, we highlighted two methods to generate an index measure that captures the impact of a biomarker on a clinical outcome, in the longitudinal repeated measures setting. Both approaches constitute an extension of the theory of targeted maximum likelihood estimation. While the time slope method is simple and easy to implement, it is not free from standard regression models assumptions. Simulation results show that its performance could be constrained when the number of repeated measures is less than 20.

In the LOESS method, no strong global assumptions are needed about the conditional mean of the VIM, and no specific functional form is assumed. Results of the analysis of the GS data as well as the simulation reported in this paper confirm that this method produces results that are at least as good as those from the time slope method. One improvement over the time slope approach is that the LOESS method provides insight into the longevity of the predictive power of a given biomarker. From the LOESS plots, one can readily assess which biomarkers predict the outcome early or late in the process, how far ahead in time a biomarker measure could predict a clinical outcome, and whether and when the effects of a given exposure start to wane. This could have practical implications: study investigators

could use this kind of information to decide when the optimal time to take repeated measurements of a biomarker would be. This could help reduce waste and improve compliance issues as repeated measurements would be taken only when they are needed.

From a public health perspective, this research is relevant for several reasons. It enabled us to identify from the GS study potentially useful candidate biomarkers based on their true importance with regards to the clinical outcome of CD4 cell count. Our results indicate that useful predictors of CD4 cell counts are: HIV subtype, HSV-2 status, and baseline CD4 cell counts. This list seems to be in line with current medical knowledge. For instance, studies by Kanki et al. (1999) and by Kaleebu et al (2000, 2002) have shown that the HIV-1 subtype an individual becomes infected with can be an important factor in the rate of disease progression. With respect to subtype, the results of this analysis should be interpreted in the African context, where the subtypes under study are the most prevalent. These results may not be applicable to Western developed countries where subtype B is largely dominant.

As a measure of a patient's immune capacity, CD4 cell count has been considered as a standard method for determining eligibility for HAART and HIV disease progression (Ellenberg, 1991; Fleming, 1994). However, CD4 count measurements are expensive and could be prohibitive in resource-poor countries. If, as our results indicate, baseline CD4 cell counts are highly predictive of future CD4 cell count level, one strategy to both contain costs and save lives might be to obtain one initial CD4 measure when infection is discovered, rather than having to do it

repeatedly over time. In the event that other repeated measures of CD4 cell counts are needed, one might be able to limit the frequency of these measurements by using the LOESS method outlined in this paper.

This study does have limitations. The TMLE method relies on the assumption of no unmeasured confounders, i.e. within strata of covariates (W), the target biomarker (A) is randomized. The thinking is that if the vector of covariates (W) contains all confounders, then among subjects sharing the same W, the potential outcomes Y_A and the biomarker (A) would be independent conditional on W (As would be the case in a blocked experiment where the treatment or biomarker (A) would be randomized within the levels of W) (Cole and Hernán, 2008; Schafer and Kang, 2008). Given that the GS study was observational, exposure was not controlled, thus "treatment" (i.e. biomarker) received might not be independent of potential outcomes. In this case, the difference E $\{Y_1|A=a\}$ - E $\{Y_0|A=0\}$ might not be an unbiased estimate of the average treatment effect, as would be the case in a randomized study. To account for this, we tried to find all important covariates (W), believed to be related to both potential outcomes and exposure (A) based on the literature or expert knowledge, and we included them in the estimation of the population average effect. There is no way to verify whether there remained any putative confounders that were not part of the vector of covariates (W) used in this study.

Another limitation is inherent to the LOESS procedure. If the data do not have a monotonic progression, the LOESS curve may not be an effective tool in differentiating signal from noise. If there are too many peaks and valleys, or up and

down patterns, the LOESS method may display inability to tease out true biomarker effects from noise. Moreover, LOESS works better on large and densely sample datasets, which could hinder its usefulness in studies where the number of measurements is relatively small. Our simulation study has shown that the two estimators have poor performance when the number of repeated measures is 5 or less. Because of this shortcoming, there may be little value in using these methods when the number of data points (i.e. repeated measures) is fewer than 10. They could be more suitable to longitudinal studies with extensive data collection over many time points. One extreme example could be the reported trial of a topical treatment for HIV-related peripheral neuropathy where patients were required to record pain four times a day for four weeks at baseline and at follow-up, for a total of 224 data points (Paice et al, 2000).

In summary, this study shows a promising practical application of the targeted maximum likelihood estimation to generate a longitudinal variable importance measure for each biomarker from a set. In this research, both biomarkers and covariates were chosen at 6 months from estimated infection date. One future area of research could be the inclusion of time-varying biomarkers and covariates. It is possible that some biomarkers would exhibit a significant effect on CD4 cell count only when they are allowed to vary over time. In such situations, the measure taken at baseline (in this application, at viral load set point) may not be a good predictor of the outcome over the long term. As a future direction, we would also recommend taking censoring into consideration, by applying the TMLE methodology in a time to events setting.

CHAPTER 4

Inverse Probability Weighting to Estimate Biomarker Variable Importance Measures from an Observational HIV Infection Dataset

4.1. Introduction

Biomarker identification related to many clinical and health outcomes is a focus of tremendous research activity on many levels, from basic laboratory studies through epidemiological investigations and late phases of clinical trials. Biomarker data are often used for disease detection, diagnosis, prognosis, to identify patients who are most likely to benefit from selected therapies, and to guide clinical decision making. Due to major advances in technology and in modern biology, a large number of biomarkers have been identified, and selection of the ones with maximal impact on clinical outcomes has become a critical issue. This has given rise to a quest for novel statistical methods that could adequately explain the relationship between biomarkers and outcomes of interest. This paper applies an innovative methodology, the inverse probability weighting, to the issue of biomarker identification in the presence of fixed covariates, in the time to events setting.

In biomarker studies involving longitudinal time to event analysis, a reasonable goal could be to estimate the importance of each biomarker from a set with respect to the time it takes for a clinical event to occur. A standard analytical approach to estimating biomarker variable importance in relation to survival

consists in fitting a Cox proportional hazards model with all measured covariates and then computing the measure of effect of interest. This approach, however, may produce biased estimates of the exposure effect if the censoring mechanisms are non-random (i.e. informative) (Robins, 1995). For instance, in the analysis of the effect of HIV infection biomarkers on CD4 cell counts, both loss to follow-up (depending on the reason) and death could lead to informative censoring, as they might be associated with the CD4 cell value at the time of the event. An example of such occurrence can be found in Touloumi et al (1999). In their comparison of CD4 cell count trends between subjects with low (L) and high (H) doses of didanosine (ddI) in patients with symptomatic HIV disease intolerant to zidovudine (AZT), using data from the Alpha Trial, Touloumi and his colleagues found that drop-outs due to death occurred more frequently in subjects with low CD4 cell counts. In this case, because the probability of drop-out was associated with the value of the previous CD4 cell count, the censoring mechanism was informative; for this reason, the use of standard analytic methods could lead to overestimation of exposure effects because subjects with worse CD4 count evolutions would have shorter follow-up times and hence would be weighted less in the estimations of the group rate of the average CD4 cell counts decline (Duvignac and Thiébaut, 2006).

One strategy for dealing with the bias arising from either confounding or informative censoring is through the use of inverse probability weighting, whereby a weight is attributed to the contribution of each subject *i* to the risk set at time t. This estimator exploits available auxiliary information to control for confounding through an exposure assignment process and adjusts for differential drop-out and

informative censoring via a censoring mechanism. The idea of weighting originates from the survey sampling field where subjects from a given sample are weighted by the inverse of their probability selection to ensure adequate representation of the population in which samples were drawn (Kish, 1965, 1990). In observational studies, there is no random allocation of exposure, and thus, subjects with certain baseline prognostic factors may be either over-represented or under-represented in certain exposure groups. Because observed differences in observational data may reflect underlying differences between groups, it is critical to mitigate bias resulting from the imbalance in covariate distributions. Similar situation occurs in event history data with differential attrition and censoring. To deal with this bias, Robins (1999, 2000) proposed a weighting scheme referred to as inverse probability of treatment weighting (IPTW).

The basic idea in using the IPTW estimator is to create a re-weighted dataset in which a balance is achieved in the distribution of covariates between exposure groups, as would be the case in a randomized intervention. Implementation of this approach involves fitting a multivariate regression of the probability of the exposure given covariates. This model, referred to as the treatment mechanism, is used to compute propensity scores or the conditional probability of receiving one's own exposure given a vector of observed covariates (Rosenbaum and Rubin, 1983). Letting p_i be the probability of a subject receiving the exposure actually received conditional on observed covariates (i.e. propensity scores), unstabilized weights (denoted by v_i) are given by $\frac{1}{p_i}$ for exposed subjects, and by $\frac{1}{1-p_i}$ for unexposed subjects. Thus, subjects with low probability of exposure are assigned relatively

larger weights while those with common exposure status given covariates are attributed lower weights. According to Robins and colleagues (2000), weighting creates a pseudo-population made of v_i copies of each subject i. For instance, a subject with $v_i = 10$ would contribute 10 copies of himself to the pseudo-population to make up for subjects with similar characteristics that have not been observed. In this pseudo-population, exposure is no longer confounded by covariates because outcome and exposure are conditionally independent given the covariates. Therefore, analysis performed on this re-weighted population could generate unbiased estimates of the exposure effect (Robins et al, 2000).

Unstabilized weights may suffer from a shortcoming whenever the set of covariates used are strongly associated with the exposure, especially in non-saturated models. In such occurrences, the weights might have large variability, which could result, according to Robins (2000), in a few subjects having extremely large values of the weights v_i; these subjects could dominate the weighted analysis because of the large number of copies of themselves they contribute to the pseudo-population relative to the contribution of other subjects. Studies by Kang and Schaffer (2007) have shown how unstable weights could result in a poor performance of the IPTW estimator. A solution proposed by Robins (2000) has been to use stabilized weights (designated by sw_i) instead, where some function of the exposure is used as the numerator of the weights instead of 1 (Robins et al, 2000). A common practice has been to use the sample proportion of subjects with the exposure (A) of interest (i.e. p(A=a)) to stabilize the weights, as was done in Cole and Hernán (2004). Thus, if an exposure (A) and a set of covariates (W) were unconfounded, the numerator and the

denominator of the stabilized weights would be the same (i.e. p(A=a) =P(A=a|W)), resulting in a sw_i of 1. All subjects with $sw_i = 1$ would then contribute the same weight to the risk set.

An indication of well-behaved weights is a mean of 1 (or in the neighborhood of 1) with a relatively small variability. A mean of 1 for the weights is often viewed as a necessary condition for correct model specification (Hernán and Robins, 2006; Cole and Hernán, 2008); however, near-perfect weights (with mean of 1 and a small range) may have no effect on the parameters of the Cox regression model, as they fail to control for confounding (Cole and Hernán, 2008). Conversely, weights with mean far from 1 or with extreme values may be symptomatic of violation of the assumptions used in the estimation of the weights.

The concept of weighting has been extended to studies involving censoring (Robins et al, 2000; Hernan et al, 2000) through the inverse probability of treatment and censoring (IPTC) estimator. The basic idea remains the same: exploit available auxiliary information through a weighting scheme to control for selection bias arising from differential drop-out and informative censoring. The censoring weights are estimated in an analogous way to the exposure weights except that a censoring indicator is now used as the dependent variable while exposure is an added as an additional regressor in the multivariate regression model. The censoring weight is the inverse of the probability of a patient *i* remaining uncensored up to time t. As in the case of the exposure weights, it is a good practice to stabilize the censoring weights as well. Finally, each subject's contribution to the risk set at time t is weighted by the product of the exposure and censoring weights (Robins et

al, 2000; Hernan et al, 2000). Application of a Cox proportional hazards model - or the equivalent pooled logistic regression model - to this weighted population generates consistent estimates of the exposure effect (Robins et al, 2000; Hernán et al., 2000, 2001) because the weights adjust for all measured confounders, which allows one to fit an association model on a dataset where selection bias has been removed.

The outline of the paper is as follows. Sections 4.2 reviews a weighted Cox proportional hazards model proposed for biomarker selection. Section 4.3 describes an HIV infection dataset used to illustrate the appropriateness of this modeling framework to the problem of biomarker selection using survival data. Section 4.4 reports on the analysis and results. Section 4.5 discusses the results while Section 4.6 concludes and provides ideas for future research.

4.2. Materials and Methods

4.2.1. Statistical Model

The goal of the analysis was to estimate a variable importance measure that reliably captures the marginal effect of a set of HIV infection biomarkers on the time to the clinical event of interest (as defined in section 4.3). To fix notation, let T_i denote the failure time of interest and C_i the censoring time for the ith subject in the sample. Furthermore, let $X_i = \min(T_i, C_i)$ be the observed time response variable, and $\delta_i = I(T_i < C_i)$ the censoring indicator such that δ_i=1 if the response was censored, and 0 otherwise. Finally, let's assume there are p exogenous covariates $W_i = (W_{i1}, W_{i2}, \ldots, W_{ip})$, and j biomarkers $A_i = (A_{i1}, A_{i2}, \ldots, A_{ij})$, recorded for each

subject i at baseline. The right censoring data structure is thus given by O =(W_i, A_{ij}, δ_i, X_i).

In the traditional framework of the Cox proportional hazards model (Cox, 1972), the regression model for the hazard as a function of a biomarker (A) and p exogenous covariates can be specified as the product of a baseline hazard and an exponential of the linear function of A and the W's. More specifically, the log hazard is given by log λ (t |A, W) = log $\lambda_0(t) + \beta_1 A + \sum_{j=2}^{p} \beta_j W_j$, whereas the hazard is defined as λ (t |A, W) = $\lambda_0(t) \exp\left(\beta_1 A + \sum_{j=2}^{p} \beta_j W_j\right)$, where $\lambda_0(t)$ is the baseline hazard. The conditional survival function is then S(t; A,W) = $\exp\{-\int_0^t \lambda_0(t) \exp\left(\beta_1 A + \sum_{j=2}^{p} \beta_j W_j\right) dt\}$. Typically, the measure of variable importance is the regression coefficient β_1.

As stated earlier, conventional estimate of $\beta_1$based on partial likelihood is biased. Therefore, to generate an estimate of the importance of each biomarker with regards to the failure time while taking into account the dependency between censoring process and survival and between covariates and exposure, we fitted a weighted Cox proportional hazards model . For this purpose, a 2-stage process was implemented: For each subject at each visit, we estimated a weight as a cumulative product of a treatment mechanism weight and a censoring weight. The process of weights creation is specified in section 3 below. These weights were then used in a time-dependent Cox proportional hazards model to generate adjusted estimates. Under certain assumptions, the inverse probability weighting estimator provides a valid test of the null hypothesis of no exposure effect while addressing the issues of

confounding and informative censoring (Robins et al, 2000; Robins, 1999; Robins, 2000). These assumptions are:

a. **Coarsening at random**: We assumed that the covariates used in this analysis were sufficient to adjust for both confounding and informative censoring. This implies that within strata of W, A is randomized, and that the probability of censoring is independent of the outcome a subject would have experienced in the absence of censoring.

b. **Experimental Treatment Assignment (ETA):** $0 < P\,(A = a|W) < 1$, that is, within all possible levels of the covariates (W), there are both exposed and unexposed subjects.

In addition to the above assumptions, consistency of inverse probability weighting estimators relies on correct estimation of the weights models. In this analysis, logistic regression was used to estimate the weights. However, if there are doubts about appropriate functional forms for the biomarker exposure model and the censoring model, data-adaptive techniques and cross-validation such as Deletion\Substitution \Addition (D\S\A) algorithm (Sinisi and van der Laan , 2004) or super learner (van der Laan et al., 2007; Polley and Van der Laan, 2010) could be considered. More detailed discussions on how to select the variables to be included in the exposure model can be found in Brookhart et al. (2006).

Of the three above assumptions, only the ETA is verifiable. For a more complete description of these assumptions and their practical applications, the interested reader is referred to Cole and Hernán (2008).

4.2.2. Measure of Effect and Parameter Estimation

Our goal was to estimate the effect of each biomarker (A) on survival T, while using covariate information to adjust for possible confounding, and informative drop-out. The parameter of interest was the log hazard ratio from the Weighted Cox proportional hazards model associated with each biomarker of interest.

Suppose A is a binary biomarker with values 1 and 0 denoting the presence of absence of a certain biological characteristic. Using a Cox regression model, we can estimate the importance of A with respect to T by taking the difference between the two following expressions:

$\log(hazard|A = 1) = log\lambda_0(t) + \beta_1 * 1 + \sum_{j=2}^{p} \beta_j W_j$

And

$\log(hazard|A = 0) = log\lambda_0(t) + \beta_1 * 0 + \sum_{j=2}^{p} \beta_j W_j.$

The difference in log hazards is simply β_1, the quantity of interest. It is estimated through the method of partial likelihood (Cox, 1972). Suppose there were m event times, and let the survival times (times to failure) be: $t_1 < t_2 < \ldots < t_k$ with corresponding "risk sets" $R_{t1}, R_{t2}, \ldots, R_{tk}$ (The risk set represents the set of subjects available for the event at time t_i). Furthermore, let R_j be the list of subjects at risk just before t_j. In the standard procedure, parameter estimates are obtained using the unweighted "partial likelihood" for : $L_p(\beta) = \prod_{i=1}^{m} \frac{\exp(W^*{}_i'\beta)}{\sum_{j\epsilon R(t_i)} \exp(W^*{}_j'\beta)} \delta_i$

where, the vector W^* contains both the biomarker (A) and the covariates (W). Taking the log yields:

$\text{Log } L_p(\beta) = \sum_{i=1}^{m} \delta_i [W^*{}_i \beta - log \sum_{j\epsilon R(t_i)} \exp(W^*{}_j \beta)\}].$

We then computed the derivative of this function with regards to β and set the ensuing score function to 0. The MLE of $\beta(i.e.\hat{\beta})$ was obtained as a solution to the system of the equations:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{m} \delta_i [W^*_i - \frac{\sum_{j\epsilon R(t_i)} W^*_j \exp(W^*_j \beta)}{\sum_{j\epsilon R(t_i)} \exp(W^*_j \beta)}] = 0$$

In the weighted Cox regression setting, the above Cox partial likelihood score for the parameter β should be modified to incorporate the weight for each subject in the risk set at time t_i. Suppose v_i is the product of the censoring weights and the exposure weights for each subject, the above score equation would become:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{m} \delta_i v_i [W^*_i - \frac{\sum_{j\epsilon R(t_i)} v_j W^*_j \exp(W^*_j \beta)}{\sum_{j\epsilon R(t_i)} v_j \exp(W^*_j \beta)}] = 0.$$

4.2.3. Inverse Probability Weighting Implementation

Two models were used in the estimation of the weights: the biomarker exposure model and the censoring model. While weights from the biomarker exposure model allowed one to adjust for confounding, those from the censoring model adjusted for possible informative censoring. In the weighted Cox regression model, a cumulative product of those two weights for each subject at each time point was used.

4.2.3.1. Biomarker Exposure Model

The first task in creating the weights consisted in estimating the biomarker exposure model, P(A|W). For binary biomarkers, the selection of this model was achieved through logistic regression. Assuming a model of the form

$\log(\frac{p(Ai=1\,|Wi)}{p(Ai=0\,|Wi)}) = \alpha_0 + \alpha_1 Wi$, we computed the following expressions of probability: $p(Aij = 1\,|Wi) = \frac{e^{(\alpha_0 + \alpha_1 Wi)}}{1+ e^{(\alpha_0 + \alpha_1 Wi)}}$ and $p(Aij = 0\,|Wi) = \frac{1}{1+ e^{(\alpha_0 + \alpha_1 Wi)}}$.

For continuous biomarkers, a flexible parametric approach, proposed by Hirano and Imbens (2004) was implemented to compute a generalized propensity score required to create the weights. First, we postulated a normal distribution of the continuous biomarker (A) given the covariates, i.e. $A_i|W_i \sim N\,(\beta_0 + \beta'_1 W_i\,, \sigma^2)$. Then the parameters β_0, β_1, σ^2 were estimated by least squares regression. Following Hirano and Imbens (2004), we estimated the generalized propensity scores by inserting the estimates of these parameters into the normal density: $\widehat{R_i} = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\,(-(\frac{A_i-(\hat{\beta}_0+\hat{\beta}_1 W_i)}{2\hat{\sigma}^2}))$.

In the second stage of this approach, we created the weights. Stabilized weights for binary biomarkers A were: $SWi = \frac{p(Ai=a)}{p(Ai=a\,|Wi)}$. For continuous biomarker variables, the estimated stabilized weights were constructed as a ratio of densities (i.e, $SWi = \frac{f_{ai}}{f_{ai}|Wi}$) as proposed in Robins et al. (2000); f_{ai} is the marginal density of the continuous biomarker (A), and $f_{ai}|Wi$ is the conditional density of the biomarker (A) given the set of covariates (W). To estimate the numerator f_{ai} we specified a normal distribution (i.e. $A_i \sim N\,(\alpha_0, \sigma^{*2})$, and then plugged the mean $\hat{\alpha}_0$ and the empirical variance $\hat{\sigma}^{*2}$ of the biomarker (A) values into the normal density. The denominator was estimated based on the generalized propensity score method of Hirano and Imbens (2004) described above.

4.2.3.2. Censoring mechanism

To correct for informative censoring, a process similar to the one followed for binary exposure variables was implemented. For this purpose, we derived a censoring indicator variable that took the value of 1 if the subject was censored and 0 otherwise. For each subject i at each time point j, we estimated through a pooled logistic regression model the probability of being uncensored conditionally on baseline covariates and biomarkers, and the predicted probability from this model was used to generate weights for each participants at all time points. Only non-administrative censoring was corrected for. Administrative censorings were treated differently in the censoring model than earlier drop-out: while earlier drop-outs were assigned a 1 for the censoring indicator at their last visit, subjects administratively censored had a 0 if they never had the event and were in the study at least 6 months preceding the analysis cutoff date. The censoring weights, in the presence of non time-varying covariates, were defined as:

$SWi^*(t)=\prod_{j=0}^{t}\frac{P[C(t)=0|W_i]}{P[C(t)=0|A_i,W_i]}$, where C(t)=1 if a subject was right censored by time t, and 0 otherwise.

4.2.4. Weighted Cox Proportional Hazards Model

In order to generate final adjusted estimates of the overall effect of each biomarker, each subject's observation was weighted by SWi x $SWi^*(t)$. Based on a number of applications available in literature, the conventional approach to fitting such models has been to use the time-varying, subject-specific stabilized weights in a weighted pooled logistic regression model in order to approximate the parameters

of a time-dependent Cox model. Relevant examples can be found in Hernán et al. (2000), Choi et al. (2002); Cole et al. (2003, 2008), Westreich et al. (2010). This approximation of the Cox regression based on pooled logistic regression works well when events are rare, but tends to produce biased estimates of the exposure effect in the case of frequent events (Young at al, 2009). Xiao et al. (2010) argued that an alternative approach might be to fit a directly weighted time-dependent Cox proportional hazards model and used evidence from simulations to demonstrate that this approach always yields unbiased estimates whether or not the outcome under study is rare.

Based on the work of Young at al. (2009) and of Xiao et al. (2010), in this analysis we used the counting process style of input to fit a weighted time-dependent Cox regression using the PHREG procedure in SAS software (version 9.2, SAS Institute, Cary, NC). We defined an indicator of failure (Dj) at time j within each subject-visit [start, stop]. Biomarker exposure served as the lone regressor in the Cox model. We accounted for any within-subject correlation induced by the individual weights by computing robust estimates of the standard errors based on the sandwich estimator (Lin and Wei, 1989). An additional advantage of this approach is the fact that survival estimates were readily available, thus could be used to generate weighted survival curves for each biomarker of interest.

4.3. Application

4.3.1. Study Population

The dataset used for application in this paper came from the Hormonal Contraception and HIV Genital Shedding and Disease Progession Study (thereafter referred to as GS study), a prospective multicenter study of 306 HIV infected women aged 18 to 45 years old from Uganda and Zimbabwe. This study started in 2001 as an add-on to the HC-HIV study (Morrison et al., 2007) and was completed in the field in December 2009. Women who seroconverted during the course of the HC-HIV study were recruited for the GS study, based on procedures outlined in Morrison et al. (2010). The study specific objectives are described in details elsewhere (Morrison et al., 2007, 2010), but one key research question is the effect of hormonal contraception on the biological parameters of the infectivity of women with primary and chronic HIV infection to their sex partners.

The GS Study consisted of a baseline visit and follow-up visits at 2, 4, 8 and 12 weeks following HIV seroconversion, and then every 12 weeks for up to 9 years. Women who developed severe HIV infection or who had successive CD4 cell counts at or below 200 cells per mm^3 were offered highly active antiretroviral therapy (HAART) and were seen twice a month initially, then monthly thereafter. In addition to baseline demographic characteristics, at each time point, information on various laboratory parameters, reproductive variables, contraceptive exposure, and recent sexual behavior was collected. Laboratory data that was collected included HIV plasma viral load, HIV sub-type, CD4, CD8 and total lymphocyte counts, serum chemistries, lipid profile, specimens for the detection of chlamydial, gonococcal, syphilis, *herpes simplex virus 2* (HSV-2), and Human papillomavirus (HPV) infections. The study also collected information on hormonal contraceptive use,

diisease progression parameters, as well as virologic, immunologic, and clinical response to HAART among hormonal and non-hormonal contraceptive method users. For a detailed description of the study population and procedures, the reader is directed to Morrison et al. (2010).

4.3.2. Outcome Definition and Censoring

In this analysis, the time variable was defined as the time from estimated HIV infection date to either the second of two successive CD4 cell counts below 350 cells/mm^3 or the second CD4 cell count of 350 cells/mm^3 or below within a six month period. In both cases, we considered two CD4 counts below 350 cells/mm3 instead of a single count below 350 cells/mm^3 because CD4 cell count measurements tend to be highly variable both from person to person and within an individual patient. Taking more than one measurement helps mitigate the effect of large individual variability in CD4 cell counts values.

The choice of the threshold of 350 cells/ mm^3 for CD4 cell counts was based on current guidelines for the use of antiretroviral agents in adults and adolescents infected with HIV-1 in the absence of an AIDS-defining illness (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2008; Hammer et al., 2008). Also, evidence from the literature suggests that early initiation of antiretroviral therapy before the CD4 cell count falls below 350 cells per mm^3 significantly improves survival, as compared to deferred therapy (Sax and Baden, 2009; Kitahata et al., 2009; When to Start Consortium, 2009).

In this study, censoring occurred either by death, loss to follow-up, or by end of study observation. In this analysis, any subject who was still alive, did not experience the outcome and was still in the study at the end of 2009 was censored at their last recorded visit. This kind of censoring, referred to as administrative censoring, was treated differently in the censoring mechanism model, than drop outs that occurred 1 year or more prior to the data cutoff for this analysis. As mentioned in section 2, drop-outs as a result of death or loss to follow-up (depending on the type of loss to follow-up) could be non-random. Thus, the need to control for both confounding and informative censoring motivates the application of inverse probability of treatment and censoring weighting to this data set.

4.3.3. Biomarker Exposure Variables and Covariates

In the analysis, we controlled for a host of risk factors reported in the literature to be predictive of incident HIV infection or clinically associated with HIV disease progression (Van Der Pol et al., 2008). These covariates include age, country, primary partner risk, STI history, having more than one sex partner, number of coital acts in previous 3 months, frequency of nights away from home by study subject's partner, condom use consistency, study subject's partner's sexual behavior and risk, and breastfeeding.

The biomarkers for which we estimated a measure of variable importance included plasma viral load, HIV sub-type (A, C, D), hemoglobin level, CD4 cell counts at baseline, CD8cell counts, total lymphocyte counts, HSV-2 status, CD4/CD8 ratio, and CD4 percentage.

Finally, baseline in this analysis was defined as the latest available biomarker and covariate measurements at 6 months after the estimated HIV infection date. A number of participants, however, were missing baseline data related to a few biomarkers. For these subjects, baseline was redefined as the first GS visit with available biomarker data.

4.4. Data Analysis

Because the censoring process was time-dependent, we used the counting process formulation to create the dataset required for the analysis. Following the work of Anderson and Gill (1982) who developed the notion of the counting process style of input, we allowed each individual in the sample to have multiple records, one per measurement occasion, containing a time interval (start, stop), a censoring indicator showing the status of the interval, and a vector of explanatory variables, including the biomarker of interest. This analysis was restricted to biomarkers and covariates selected 6 months after estimated date of infection. They remained fixed at all follow-up visits. No time-varying exposure or covariates were involved.

In the raw data, the continuous biomarker variables were measured on vastly different scales. The fact that the regression coefficients in the weighted Cox proportional hazards model depended on the units of measure made it difficult to do meaningful comparison based on metric regression coefficients alone. To circumvent this problem, we incorporated standardized continuous biomarker variables (created by subtracting the mean and dividing by the standard deviation) rather than the raw continuous variables in the different regression models.

Essentially, a Z-score was computed for each continuous variable: CD4 cell counts at baseline, CD8 cell counts, HIV RNA, Hemoglobin level, CD4/CD8 ratio, CD4 percent, and lymphocyte counts. The standardization resulted in a mean of 0 and a variance of 1 for each standardized variable, making the regression coefficients for the affected variables directly comparable to one another. No change was made to the binary biomarker variables. The standardized coefficients should then be interpreted in terms of change in the clinical response variable resulting from a change of one standard deviation in the continuous biomarker variable of interest. The p-value for the test of the hypothesis of no exposure effect stays the same whether or not one uses raw or standardized variables.

4.4.1. Weights Estimation

To estimate the weights, we fitted four different pooled logistic regression models. From the first two models, we generated the predicted probabilities required to compute the stabilized weights associated with the biomarker exposure. The numerator of the weights was obtained from an-intercept only regression model with biomarker (A) as the dependent variable, while estimates for the denominator came from a multivariate regression model with the same covariates listed above plus a smooth function of study duration represented by natural cubic splines with 4 knots at the 5^{th}, 35^{th}, 65^{th}, and 95^{th} percentiles (Harrel, 2001). The use of the cubic splines in lieu of a linear term relaxed the dependency on the strong linearity assumption with regards to duration of follow-up while allowing for time-varying hazards.

Two additional pooled logistic models were used for estimating the censoring weights. All the covariates listed above, including the cubic splines, were used in the estimation of the numerator. For the denominator, the biomarker exposure variable of interest was included as an additional regressor.

For each subject i at each time point j, we subsequently computed an overall weight that was the product of stabilized weights obtained from the biomarker exposure model and the censoring mechanism. These weights were then entered into the Cox model to generate the adjusted estimates of the variable importance measure for each biomarker. This process was followed separately for each biomarker A_p, (p = 1, … … k). There were then k estimates of $\overline{\hat{\theta}_b}$ of variable importance measures and k p-values denoted as P_1, P_2,…, P_k. We adjusted the p-values for multiplicity testing to control the false discovery rate. The Benjamini and Yekutieli (2001) False Discovery Rate (FDR) controlling procedure was used to account for the dependence of the test statistics. Significance for each biomarker was assessed by comparing related adjusted p-value to the 0.05 alpha level. A lower p-value denoted a better measure of importance.

4.4.2. Results

In the GS study, the median age at enrollment was 27 years. Women from Zimbabwe accounted for 58.5% of the population while those from Uganda made up the remaining 41.5%. About 8% of all study subjects had at least two sex partners while 14% had a STI history, 8% were breastfeeding, and 45% displayed STI symptoms. These women averaged 11.2 (Standard Deviation [SD] = 15.5) sex acts

per month, but only 35% of them reported consistent condom use. On average, the partners of these women spent 10 (SD=15.2) nights away from home, and 75% of those partners had been reported to have had sex with another woman in the three months prior to enrollment in the study. Finally, 59% of the subjects' partners met the study definition for primary partner risk, a composite variable that included having a partner with HIV, urethral discharge, weight loss, nights spent away from home, or a history of sex with female sex workers.

Table 7 presents the distribution of the stabilized weights. Overall, the mean for the weights computed for each biomarker was clustered around 1, which is a desired result. All biomarkers displayed small variability in the weights.

For each biomarker, we derived estimates of 5-year cumulative probability of survival with corresponding 2-sided 95% confidence intervals, using the Kaplan Meier estimator (Kaplan and Meier, 1958). In this context, survival was defined as the probability of not experiencing the event of interest in the first 5 years since estimated infection date. For the purpose of survival curves estimation, all biomarkers measured on a continuous scale were dichotomized based on meaningful clinical values suggested in literature (Table 8). In short, having the following biomarker characteristics was associated with a 5-year cumulative probability of survival ≤ 40%: CD4 at baseline ≤ 500 cells/mm^3 (35% survival rate), Lymphocyte count < 1200 cells/mm^3 (35%), and CD4 Percentage ≤ 20 % (30%).

We computed both weighted and un-weighted estimates of the importance of each of the 11 biomarkers under consideration (Tables 9 and 10). The un-weighted estimates were obtained from a standard Cox proportional hazards model with the

following covariates: age, country, primary partner risk, STI history, having more than one sex partner, number of coital acts in previous 3 months, frequency of nights away from home by study subject's partner, condom use consistency, study subject's partner's sexual behavior and risk, and breastfeeding. The results in Table 9 suggest that, when the standard Cox model was used, the following biomarker variables had a significant impact on CD4 cell counts: Baseline CD4 Cell count, CD4/CD8 T-cell Ratio, Plasma Viral Load, and Lymphocyte Count.

In the weighted analysis (Table 10), the same four biomarkers were found to exert a significant impact on the time to the second successive drop of CD4 cell counts below the threshold of 350 cells/mm^3. Based on the magnitude of the p-values, the most important biomarkers were baseline CD4 Cell count, CD4 Percentage, CD4/CD8 Ratio, Lymphocyte Count and HIV Subtype A. Note that, in the pseudo-population created by the weights, HIV subtype A only reached borderline statistical significance.

Evidence from both tables 10 and 11 suggests that a lower hazard (better survival) was associated with increases in Baseline CD4 Cell count, CD4 Percentage, CD4/CD8 Ratio, Lymphocyte Count, hemoglobin levels, and with being of HIV Subtype A. For these biomarkers, the log hazard ratio is negative. Conversely, a higher hazard (lower survival) was linked to increases in Plasma Viral Load (Log10/mL), in CD8 cell counts, and with being of HSV-2 positive or of HIV subtype C or D (positive log hazards ratio).

4.5. Discussion

In this study, we applied a flexible tool for estimating biomarker exposure effects in observational data, taking into account covariates information. The list of biomarkers deemed significant as well as the direction of the associations noted - as suggested by the sign of the log hazards ratio or by a visual inspection of the adjusted survival curves (not shown) - is consistent with current clinical and medical knowledge of HIV infection. For instance, a negative log hazards ratio was expected - and found - for increases in lymphocytes counts. It is known that total lymphocytes counts (TLC) tend to decrease as a result of HIV infection and disease progression. Also, the significant association found with the outcome is consistent with research findings of a relatively high positive correlation between absolute values of TLC and CD4 cell counts or between changes in TLC and CD4 cell counts (Badri and Wood, 2003; Mwamburi et al., 2005). This finding could have practical applications for HIV medical care. As a measure of a patient's immune capacity, CD4 cell count is considered as a standard method for determining eligibility for highly active antiretroviral therapy (HART) and HIV disease progression. However, its measurements require highly skilled personnel and costly maintenance of sophisticated equipment, and these costs could be prohibitive in resource-deprived countries. Cheaper alternatives identified in this study (e.g. TLC), upon further evaluation, could potentially support decision-making with regards to the initiation of antiretroviral therapy or could help monitor patients' immune status during therapy in the absence of expensive CD4 measurements. Overall, the application of the weighted Cox proportional hazards model to the GS study data provides valuable

information for HIV medical care, and should be considered in the panoply of techniques used in biomarker assessment.

In addition to its ability to produce a consistent estimate of the effect of a given exposure, the Weighted Cox proportional hazards model is appealing because it makes it easier to create adjusted survival curves, which can be viewed as a graphic summary of the data averaged over the covariates used in the weights models. While adjusted estimates from the Cox proportional hazards model have been ubiquitously used in the reporting of results from survival analysis, survival curves have been used less frequently in observational studies (Hernán, 2010) due to the lack of a standard method for dealing with confounding. There have been attempts in the literature to generate adjusted survival curves from the conventional Cox model, but these applications were fraught with problems (Nieto and Coresh, 1996). One notable shortcoming identified in Nieto and Coresh's paper was the inability to adjust for continuous covariates. In their 2004 paper, Cole and Hernán proposed and demonstrated the idea of using survival estimates from the weighted Cox model to generate adjusted survival curves. This method was simple, easily implemented using standard statistical software, did not involve stratification on any covariate, and accommodated both continuous and time-varying covariates. Thus, even when results and conclusions from standard covariate adjustment through the conventional Cox proportional hazards model are identical to those from a Weighted Cox proportional hazards model performed on the same data, the latter method has the advantage of readily generating adjusted survival curves. Certainly, the use of survival curves to report results from time to events analyses is encouraged (Hernán,

2010) because survival curves have served as effective tools for displaying informative and meaningful summary of study findings over the span of the entire study period.

This study has limitations. We applied the weighted Cox proportional hazards model under the assumption of no unmeasured confounders for biomarker exposure and censoring. There is no direct way to verify whether there remained any putative confounders that were not part of the vector of covariates (W) used in this study. All measured confounders are controlled for in the weights, and bias could still exist if some important confounding variables were not measured and, therefore, were not included in the weighted models. To guard against violation of this assumption, we included process a number of covariates believed to be related to the exposure of interest and the outcome (based on existing literature and expert knowledge) in the modeling process (Van Der Pol, 2008). Another potential limitation could be the occurrence of practical violations of ETA. Research by Wang et al (2006), Neugebauer and van der Laan (2005), and by Moore et al (2010) has demonstrated how ETA violations could result in significant bias in the inverse probability weighted estimator of causal effect models. In real life applications, it is not uncommon for an exposure to occur with a small probability or even with 0 probability within a given stratum of subjects. Also there may just be practical violations of ETA, defined as the occurrence of random 0 or 1 probability by chance. In this study, we have used a number of biomarkers measured on continuous scales, and it is known that ETA violation tends to be frequently associated with the use of continuous exposure variables. To reduce the impact of practical violations of ETA

on the stability of our estimator and to ensure an adequate bias-variance trade-off, we set all estimated probabilities from both the censoring and exposure models below 0.01 to 0.01, as suggested in Bembom et al. (2008). Another technique implemented in this analysis to mitigate the effects of possible ETA violations was the use of stabilized weights, which allowed for a weaker form of the ETA assumption (Wang et al, 2006).

4.6. Conclusion

This paper has provided an overview of the inverse probability weighting estimator and its application to the problem of biomarker selection in a survival setting. We used an example of observational HIV infection data to illustrate the appropriateness of this method as a tool for generating marginal variable importance measures for each biomarker of interest. This example, however, involved only static exposures and covariates. In actuality, biomarker data may involve time-varying covariates that could be both a risk factor for the outcome and a predictor of subsequent exposure. Furthermore, in those data, past exposure history might predict the risk factor. In such occurrences, the methods used in this paper should be extended to account for time-dependent covariates and exposure. One suitable solution could be the use of Robin's marginal structural models (Robins et al, 2000; Hernan et al, 2000), which have gained widespread use and acceptance in dealing with time-varying confounders. Another possible alternative could be the collaborative targeted maximum likelihood estimation (van der Laan and Gruber, 2009; Stitelman and Van der Laan, 2010), a methodology developed by Van der

Laan specifically with observational data in mind. This extension of the theory of targeted maximum likelihood (Van der Laan and Rubin, 2006) estimation is believed to provide substantial gains in both robustness and efficiency over commonly used methods.

CHAPTER 5: CONCLUSION

5.1. Overview of the study

The objective of this research was to evaluate the use of biomarker data for disease detection, diagnosis, and prognosis. Because of recent development of new biological and molecular techniques, a large number of new biomarkers have become available. Determining the predictive and diagnostic value of these biomarkers, singly or in combination, is essential to their being used effectively. This has spurred the development of new statistical methodologies that can exploit this wealth of information to adequately explain the relationship between biomarkers and outcomes of interest. This research used novel statistical methods to identify biomarkers with good performance characteristics while providing guidelines for biomarker data collection and assessment, such as sample size requirements.

The central question addressed in this dissertation was the following: Given a large set of biomarkers that potentially predict a clinical outcome, how can one make a determination as to which ones are the most important? To answer this question, we applied different estimation methods to generate a marginal variable importance measure (VIM) separately for each biomarker. Then we used the estimated VIM to make inferences about the importance of each biomarker. We performed biomarker evaluation in three different settings: Point treatment, longitudinal repeated measures, and time to event. Methods applied at each setting were as follows:

5.1.1. Point Treatment

Because CD4 cell count, as an indicator of disease progression, is frequently used to determine eligibility for HAART initiation, it is of interest to know as early as possible which subjects from a cohort are most at risk for CD4 cell count depletion. This information could then guide monitoring efforts or HAART initiation policies. The evaluation of biomarker selection methods that could provide such information in the short term is the crux of chapter 2. In this analysis, the outcome was a binary variable representing 2 successive drops in CD4 count to below 350 cells per mm^3 in the first two years following the viral set point (121 days from estimated infection date). Methods discussed in this context included Targeted maximum likelihood estimation (TMLE), flexible propensity score weighting (PSW), and incremental value estimation based on partial area under the Receiver-Operating Characteristic (ROC) curve. TMLE and PSW were both applied in a counterfactual framework and involved maximizing the following objective functions: $\Psi = E_w[E(Y=1|\ A_J=1, W) - E(Y=1|\ A_J=0, W)]$ (for binary A), and $\Psi = E_w[E(Y=1|\ A_J=a, W) - E(Y=1|\ A_J=\bar{a}, W)]$ (for continuous A). Both methods involved two nuisance parameters P(Y|A,W) and P(A|W). In the estimation of the variable importance measure, the TMLE added a covariate created from P(A|W) to the initial regression model, while the PSW method used weights from estimated propensity scores to incorporate the relationship between A and W in the regression P(Y|A,W).

In the third method highlighted in chapter 2, called incremental value estimation for partial area under the ROC curve, we defined variable importance measure as the improvement in classification performance gained by adding each biomarker separately to the set of covariates. More specifically, we generated an ROC curve for a combination of biomarker and covariates using the model P(Y=1|A,W) and another ROC curve for the covariates alone based on P(Y=1|W). The VIM for each biomarker was given by the difference in pAUC between the two models.

5.1.2. Longitudinal Repeated Measures

In this analysis, we extended the TMLE methodology to longitudinal repeated measures to look at trends over a longer term. We addressed the gap caused by the absence of a generally accepted approach for generating a scalar value representing a measure of variable importance over time. We proposed and implemented a methodology integrating both TMLE and the computation of the area under/above the LOESS curve. Computation of the index measure of interest is based on the composite Simpson's rule for numerical integration and is given by:

$\psi = \int_a^b f(x)dx = \frac{h}{3}\left(f(a) + f(b)\right) + \frac{2h}{3}\sum_{k=1}^{m-1} f(x_{2k}) + \frac{4h}{3}\sum_{k=1}^{m} f(x_{2k-1})$, where the time interval [a, b] is subdivided into 2m subintervals $\{[x_{k-1,}\, x_k\,]\}_{k=1}^{2m}$ of equal width $h = \frac{b-a}{2m}$, and f(.) are the predicted values from the LOESS model relating the VIM, ψ, to time. We then compared results from this approach to those obtained from an autoregressive model.

5.1.3. Time to Events

In this analysis, we took right censoring into consideration. The outcome was defined as the time from estimated HIV infection date to either the second of two successive CD4 cell counts below 350 cells/mm^3 or the second CD4 cell count of 350 cells/mm^3 or below within a six month period. Control for measured confounding and potential informative censoring was achieved through the use of stabilized weights in a time-dependent Cox proportional hazards model. From the log hazard given by $\log \lambda(t|A, W) = \log \lambda_0(t) + \beta_1 A + \sum_{j=2}^{p} \beta_j W_j$, the parameter β_1, estimated through partial likelihood, captures the effect of the biomarker and represents the VIM of interest.

5.1.4. General methodology

Under each method, we carried out a separate analysis to estimate a marginal measure of importance of each biomarker, controlling for measured confounding variables. This measure of importance represents the effect of each biomarker on the outcome. For inference, we applied a nonparametric bootstrap. We determined the importance of each target biomarker based on the magnitude of the p-value from the hypothesis test of a non-zero mean bootstrapped estimate for each biomarker. To take into account multiple testing and dependency of the different test statistics, we applied the Benjamini and Yekutieli procedure for controlling the False Discovery Rate (FDR). Statistical significance was reached if the FDR-adjusted p-value was smaller than or equal to alpha=0.05, with smaller p-values indicative of greater importance. A dataset from the Hormonal Contraception and HIV Genital Shedding

and Disease Progression Study (GS Study) that included longitudinal HIV infection data on a sample of 306 HIV-infected adult women from Uganda and Zimbabwe was used to develop and evaluate the methods discussed in this dissertation.

5.2. Summary of the Results and Discussions

In the cross-sectional analysis, the most important biomarkers under the TMLE methodology, among the 11 biomarkers considered, were baseline CD4 cell count and CD4/CD8 T cell ratio. Under the PSW approach, the biomarkers selected as most important were: HSV-2 status, CD4/CD8 T cell ratio, baseline CD4 cell count, and plasma viral load. No biomarker was selected as important by the incremental value method.

We made further statistical evaluation of these methods by performing simulations to assess their finite sample properties. Our results suggest that the PSW tend to perform better than the other methods in small to moderate sample sizes (i.e. $0 < N \leq 200$). The TMLE performed reasonably well in moderate to large sample sizes (i.e. $N > 100$). Finally, the incremental value approach displayed an unsatisfactory ability in detecting significant biomarkers when the sample size is less than 200, but worked well with sample sizes over 200.

In the longitudinal repeated measures analysis, the two methods under consideration yielded the same conclusion: Among the 11 biomarkers, the most important ones, based on the magnitude of the p-values, were baseline CD4 cell counts, HIV subtype, and HSV-2 status. Simulation studies assessing sample size issues indicate that performance of both methods again depends on sample size.

In the time to events analysis, the most important biomarkers were baseline CD4 Cell count, CD4 percentage, CD4/CD8 ratio, lymphocyte count and HIV subtype.

The above results appear to indicate that the list of biomarkers selected as important depends on the type of analysis performed. It would be tempting to expect the list of biomarkers deemed important to be consistent across all three analyses (point treatment, longitudinal repeated measures, and time to events). It should be noted however, that the three major kinds of analyses conducted on the GS data did not address the same research question, as the definition of the outcome in each analysis incorporated different durations of time. In the point treatment analysis, for instance, only the first two years following viral set point was of interest. This analysis inherently used less information from the GS data than the time to event analysis or the longitudinal repeated measures analysis, and the 2-year time period may not have been sufficient to detect the effect of certain biomarkers. Another possible explanation for the lack of consistency in the results could be the behavior of the biomarkers over time.

Insights gained from the LOESS curves in chapter 3 suggest that the longevity of the predictive effect of certain biomarkers may increase, decrease, or stay relatively constant over time. Thus, it is quite possible that some biomarkers could have an effect only in the short term, while the impact of other biomarkers may have been more pronounced in the longer term.

If we were to consider results from the longitudinal repeated measures as the gold standard in this dissertation, on the basis of the amount of information available in the data, we would conclude that the most important biomarkers that

predict CD4 cell counts are: Baseline CD4 cell counts, HIV subtype, and HSV-2 status. Of those three biomarkers, two (baseline CD4 cell count, HIV subtype) were selected as important in the time to events analysis, the second method in terms of amount of information used. In all three analyses, baseline CD4 cell count appeared on the list of the most important biomarkers. This is consistent with results from other studies that found a similarly strong association the baseline CD4 count and the subsequent CD4 response for patients on HAART therapy (Byakwaga et al., 2009; Florence et al., 2003; Le Moing et al., 2007; Robbins et al., 2009).

While none of the results of the three types of analysis done on the GS data seem to contradict current medical knowledge, we should approach them with caution. The definition of all 3 types of outcomes in this research involved time since infection. In the GS Study, investigators knew the last time point where a participant was HIV-uninfected, so they have been able to combine this information with the date of the first visit where that person was confirmed infected to generate their best estimate of infection date. In standard public health settings, this information is generally not available; thus, one may not know at what point in time relative to the infection date the biomarkers have been measured. An additional cautionary note concerns the use of HIV subtype results from this research. The HIV subtypes A, C, and D are prevalent in African but not in Europe or North America. Therefore, the results reported here may not be applicable to Western developed countries where subtype B is largely dominant.

5.3. Contributions of the Study

From a public health perspective, this research is relevant for several reasons. It enabled us to identify from the GS study potentially useful candidate biomarkers based on their importance with regards to the outcome of low CD4 cell count. For instance, analyses of the GS data in cross-sectional, longitudinal repeated measures, and survival contexts each identified baseline CD4 cell counts as one of the most important biomarkers. Based on this finding, a potential action item in public health practice could be the identification of patients at risk for CD4 depletion so that they could be monitored more closely and started on HAART, when necessary. To accomplish this, it might be necessary to obtain an initial CD4 measure when infection is discovered, followed by repeated CD4 but perhaps less frequently for those with a high initial CD4 (say > 500 cells/mm^3). Tools such as the LOESS curve used in this research could be used to determine exactly when additional measurements are needed. This could potentially help save lives, time, and money, especially in resources-deprived countries where the costs to measure CD4 are often prohibitive.

From a more global standpoint, research of this kind has the potential to contribute to the advancement of both clinical and public health practice. Application of analytic methods used in this research to generate a list of useful candidate biomarkers based on their true importance in predicting a given outcome, could potentially help reduce waste and time by directing biologists' focus on the best biomarkers, and by allowing practitioners to direct resources towards the most promising candidate biomarkers. From a statistical standpoint, such a list can

further direct research on establishing the level of surrogacy of the significant set of candidate biomarkers.

5.4. Strengths of the study

This research draws its strength from the significance of the subject matter and the comprehensiveness of the implemented research plan.

5.4.1. Significance and timeliness of the subject matter

Biomarker identification has recently been the focus of tremendous research activity, from basic laboratory research to clinical and epidemiological investigations. The work accomplished in this dissertation contributes to the statistical literature by addressing the issue of biomarker selection in various contexts (cross-sectional, longitudinal, survival) and by proposing a novel procedure based on non parametric regression (LOESS) to compute a longitudinal variable importance measure for biomarker evaluation. This research provides also valuable information for HIV medical care.

As biomedical research is increasingly moving towards a new era of predictive, preventive, and personalized medicine, successful biomarker selection and validation through a combination of enhanced genomic research techniques and novel robust statistical methods could speed early detection, diagnosis, and treatment of disease. For instance, if biomarker identification is improved, this could accelerate introduction of treatment early in the disease process potentially leading to reduction in disease severity, complications and mortality. The end result

will be a reduction in the burden of disease and enhancement in life expectancy. Thus, patients and the public at large stand to benefit from any new reliable statistical method for biomarker selection that contributes to an early disease diagnosis or to a rapid, efficient, and economical drug development process. Even though an HIV infection dataset is used in this research, the methods implemented in this study can be applied to other areas of biomedical research and public health practice, including vaccine studies.

5.4.2. Comprehensiveness of the plan

This research encompasses the assessment of innovative new statistical methodologies for biomarker identification that incorporate covariate information. The use of these methods for not only cross-sectional data, but also in the longitudinal and time-to event settings, combined with their application to a unique and rich data set of HIV infection data, and the simulation studies to develop sample size guidelines, create a comprehensive plan.

5.5. Limitations of the Study

This study has limitations. The methods implemented in this research (TMLE, PSW, Weighted Cox regression) rely on the assumption of no unmeasured confounders, i.e. that within strata of covariates (W), the target biomarker (A) is randomized. In this analysis, we tried to identify all important covariates (W) believed to be related to both potential outcomes and exposure (A) based on the literature or expert knowledge, and include them in the estimation of the effect of each biomarker. There is, however, no direct way to verify whether there remained

any putative confounders that were not part of the vector of covariates (W) used in this study.

Another potential limitation could be the occurrence of practical violations of experimental treatment assignment (ETA) in the three methods that used information from the biomarker exposure assignment P(A|W). In real life applications, it is not uncommon for an exposure to occur with a small probability or even with zero probabilities within a given stratum of subjects. In this study, we have used a number of biomarkers measured on continuous scales, and it is known that ETA violation tends to frequently be associated with the use of continuous exposure variables. Such violations could result in biased exposure effects. In the propensity score-based methods, we aimed at reducing the impact of practical violations of ETA on the stability of our estimators by either truncation of estimated probabilities from the censoring and exposure models or by use of stabilized weights.

Sample size is another limitation of this study, especially in the cross-sectional analysis. Over half of the biomarkers under consideration in this analysis had 200 or fewer non-missing observations because women not seen for more than 6 months following estimated infection date have missing data for biomarker measures. In the marginal analyses that we performed for each biomarker one at a time, all records with missing biomarker information were excluded from the models. As shown in simulation assessing the finite sample properties of the methods outlined in chapter 2 (TMLE, PSW, and incremental value estimation), sample size does affect the ability of all these methods to detect "true" significant biomarkers; all three showed a decreased ability in pinpointing significant biomarkers at smaller sample sizes (e.g.

N=100). Hence, sample size constraints may have hindered us from detecting additional significant biomarkers in the GS dataset. The sample size cut-points we identified through our simulations in chapter 2 and chapter 3 should serve only as starting point towards establishing sample size requirement guidelines for future data collection and assessment in biomarker studies where these selection methods are used.

Finally, in this research, both biomarkers and covariates were chosen at 6 months from estimated infection date. The inclusion of time-varying biomarkers and covariates in both the longitudinal repeated measures analysis and in the time to events analysis might be a more suitable strategy for capturing the various dimensions of the clinical outcome, as compared to single fixed measurements. It is possible that some biomarkers would exhibit a significant effect on CD4 cell count only when they are allowed to vary over time along with covariates of interest. In such situations, measures taken at baseline might not have been good enough predictors of the outcome over the long term.

5.6. Recommendations for future Research

In this research, both biomarkers and covariates were chosen at 6 months from estimated infection date and remained fixed. One future area of research could be the inclusion of time-varying biomarkers and covariates. Biomarker data often include time-varying covariates that could be both risk factors for the outcome and predictors of subsequent exposure. Furthermore, past exposure history might predict those risk factors. One suitable framework for dealing appropriately with

time-dependent covariates and exposure could be Robin's marginal structural models (Robins et al, 2000; Hernan et al, 2000). Another possible alternative could be the collaborative targeted maximum likelihood estimation (van der Laan and Gruber, 2009; Stitelman and Van der Laan, 2010), a methodology developed by Van der Laan specifically with observational data in mind. This extension of the theory of targeted maximum likelihood (Van der Laan and Rubin, 2006) estimation is believed to provide substantial gains in both robustness and efficiency over commonly used methods.

Another logical and intuitive step from this research could be the evaluation of the role of a multi-marker strategy in improving diagnostic accuracy and prediction. Specifically, it would be of clinical significance to determine whether the use of a 2 or 3-marker combination (e.g. baseline cd4, HIV subtype, and HSV-2 status) would be superior to a single marker (e.g. baseline CD4 cell count) in terms of risk prediction. To this end, one can utilize expert knowledge to make appropriate combinations of all biomarkers selected as important in this research. Once the biological plausibility of these marker combinations has been established, one can then embark on a rigorous statistical evaluation aimed at selecting the optimal biomarker combination. Of course, a drawback is that several biomarkers would have to be measured, perhaps even more than once, thus possibly making this approach less feasible in resource-constrained settings.

REFERENCES

Alaiya A, Al-Mohanna M, Linder S. (2005). Clinical cancer proteomics: promises and pitfalls. J Proteome Res;4(4):1213-22.

Anderson and Gill (1982). Cox's Regression Model Counting Process: A Large Sample Study. Annals of Statistics, vol. 10, pp. 1100-1120.

Badri M, Wood R. (2003). Usefulness of total lymphocyte count in monitoring highly active antiretroviral therapy in resource-limited settings. AIDS. 7;17(4):541-5.

Baker SG. (2003). The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. J. Natl Cancer Inst;95(7),511–515.

Bamber D. (1975). The Area Above the ordinal Dominance Graph and the Area Below the Receiver Operating Graph. J Math Psych; 12: 387-415.

Begg CB. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. Statistics in Medicine; 10: 1887-1895.

Bembom O, Fessel JW, Shafer RW, van der Laan MJ. (2008). "Data-adaptive Selection Of The Adjustment Set In Variable Importance Estimation". *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 231. Available at: http://www.bepress.com/ucbbiostat/paper231

Bembom O, Petersen M. L, Rhee S, Fessel J, Sinisi S. E, Shafer R. W, Van der Laan M. J. (2008). Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant HIV infection. Statistics in Medicine 28, 152-172.

Bembom o, Petersen ML, van der Laan MJ. (2006). Identifying important explanatory variables for time-varying outcomes. In W. Dubitzky, M. Granzow, and D.P. Berrar (eds.), Fundamentals of Data Mining in Genomics and Proteomics, Springer, Chapter 11, p.227-250.

Benjamini Y. Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society B, 1995; 57: 289 -300.

Benjamini Y, Hochberg Y. (2000). The adaptive control of the false discovery rate in multiple hypotheses testing. J. Behav. Educ. Statist; 25: 60-83.

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical Pharmacology and Therapeutics; 69: 89-95.

Blatt SP, Lucey CR, Butzin CA, Hendrix CW, Lucey DR. (1993). Total lymphocyte count as a predictor of absolute CD4+ count and CD4+ percentage in HIV-infected persons. JAMA; 269(5):622-6.

Benjamini Y, Yekutieli D. (2001). The Control of the False Discovery Rate Under Dependency. Ann Stat. 29,1165-1188.

Breiman L. (1996). Bagging predictors. Machine Learning 24,123–140.

Breiman L. Random forests. (2001). Machine Learning 45, 5-32.

Breiman L, Friedman J. H, Olshen R. A, Stone C.J. (1984). Classification and Regression Trees. Kluwer Academic Publishers.

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. (2006). Variable selection for propensity score models. Am J Epidemiol;163(12):1149-56.

Brown BW, Russell K. (1997). Methods Correcting for Multiple Testing: Operating Characteristics. Statistics in Medicine;16: 2511–2528.

Brown E. R, Otieno P, Mbori-Ngacha D. A, Farquhar C, et al. (2009). Comparison of CD4 cell count, viral load, and other markers for the prediction of mortality among HIV-1-infected Kenyan pregnant women. J Infect Dis;199(9):1292-300.

Burges C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121 – 167.

Buyse M, Molenberghs G. (1998). The validation of surrogate endpoints in randomized experiments. Biometrics. 1998;54:1014-1029.

Buyse M, Molenberghs G, Burzykowski T. Renard D, Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics.2000;1: 49-68.

Byakwaga H, Murray JM, Petoumenos K *et al.* (2009). Evolution of CD4+ T cell count in HIV-1-infected adults receiving antiretroviral therapy with sustained long-term virological suppression. *AIDS Res. Hum. Retroviruses;* 25(6); 756–776.

Carpenter J, Kenward M, Vansteelandt S. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. J.Roy. Statist. Soc. Ser. A. 2006;169:571–584.

Chen R. Y, Westfall A. O, Hardin J. M, Miller-Hardwick C, et al. (2007). Complete blood cell count as a surrogate CD4 cell marker for HIV monitoring in resource-limited settings. J Acquir Immune Defic Syndr;44(5):525-30.

Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots". Journal of the American Statistical Association 74 (368): 829–836.

Cleveland, W.S.; Devlin, S.J. (1988). "Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting". Journal of the American Statistical Association 83 (403): 596–610.

Cole SR, Frangakis CE. (2009). The consistency statement in causal inference: A definition or an assumption? Epidemiology, 20, 3-5.

Cole SR, Hernán MA. (2004).Adjusted survival curves with inverse probability weights. Computer Methods and Programs in Biomedicine 75: 45—49.

Cole SR, Hernán MA. (2008). Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology, 168(6): 656-664.

Cole SR, Hernán MA, Robins JM, Anastos K, Chmiel J, et al. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. American Journal of Epidemiology, 158(7):687-694.

Cook, T. & Campbell, D. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton-Mifflin.

Cox, D.R. (1972). Regression models and life tables. Journal of the Royal Statistical Scociety, Series B 34, 187-220.

Cross Continents Collaboration for Kids (3Cs4kids) Analysis and Writing Committee. (2008). Markers for predicting mortality in untreated HIV-infected children in resource-limited settings: a meta-analysis. AIDS;22(1):97-105.

Desai M, Stockbridge N, Temple R. (2006). Blood Pressure as an Example of a Biomarker That Functions as a Surrogate. AAPS Journal; 8(1): E146-E152.

Devlin B, Roeder K, Wasserman L.(2003). Statistical Genetics: False discovery or missed discovery? Heredity, 2003; 91: 537-538.

Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen, W (2005). Analysis of Clinical Trials Using SAS: A Practical Guide, Cary, NC: SAS Institute Inc.

Dodd LE, Pepe MS. (2003). Partial AUC Estimation and Regression. Biometrics 59, 614-623.

Dudoit S, Shaffer JP, Boldrick JC. (2003). Multiple Hypothesis Testing in Microarray Experiments. Statistical Science;18: 71–103.

Dudoit S, Yang Y.H, Speed T.P, Callow M.J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica12, 111–139.

Dutkowski J, Gambin A. (2007). On consensus biomarker selection. BMC Bioinformatics 2007,S5.

Duvignac J, Thiébaut R (2006). CD4 Natural History and Informative Censoring in Sub-Saharan Africa. Letter to the Editor. JAIDS Journal of Acquired Immune Deficiency Syndromes.1; 43(3): 380-381.

Easterbrook PJ, Smith M, Mullen J, O'Shea S, Chrystie I, Zuckerman M (2010). Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. Journal of the International AIDS Society; 13:4.

Echt DS, Liebson PR, Mitchell LB, et al. (1991). Mortality and morbidity in patients receiving encainide, flecainide, or placebo. N Engl J Med; 324: 781–8.

Efron B, Tibshirani R. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC.

Ellenberg SS. (1991). Surrogate end points in clinical trials. British Medical Journal 302, 63-4.

Fiscus SA, Hughes MD, Lathey JL, Pi T, Jackson B, et al. Changes in virologic markers as predictors of CD4 cell decline and progression of disease in human immunodeficiency virus type 1-infected adults treated with nucleosides. AIDS Clinical Trials Group Protocol 175 Team. (1998). J Infect Dis.;177(3):625-33.

Fleming TR. (1994). Surrogate Markers in AIDS and Cancer Trials. Statistics in Medicine 13,1423-35.

Fleming TR, DeMets DL. (1996). Surrogate endpoints in clinical trials: are we being misled? Annals of Internal Med;125, 605-613.

Florence E, Lundgren J, Dreezen C *et al. (2003).* Factors associated with a reduced CD4 lymphocyte count response to HAART despite full viral suppression in the EuroSIDA study. *HIV Med.;* 4(3);255–262.

Folsom AR, Chambless LE, Ballantyne CM et al. (2006). An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study. Arch. Intern. Med; 166(13): 1368–1373.

Frangakis CE, Rubin DB. (2002). Principal stratification in causal inference. Biometrics; 8:21–29.

Freund Y, Schapire R.(1997). A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences;55(1):119–139.

Gelman A, Meng XL. (2004). Applied Bayesian modeling and causal inference from incomplete-data perspectives. New York: Wiley.

Genovese C, Wasserman L. (2004). A Stochastic Process Approach to False Discovery Control. The Annals of Statistics; 32 (3):1035-1061.

Genovese CR, Wasserman L.(2002). Operating characteristics and extensions of the false discovery rate procedure. Journal Of The Royal Statistical Society Series B; 64(3): 499–518.

Gilbert PB, Hudgens MG. (2008). Evaluating candidate principal surrogate endpoints. Biometrics;64(4):1146-54.

Goddard MJ, Hinberg I. (1989). Receiver operator characteristic (ROC) curves and non-normal data: An empirical study. Statistics in Medicine;9:325–337.

Goodsaid F, Frueh F. (2007). Biomarker Qualification Pilot Process at the US Food and Drug Administration. AAPS Journal; 9(1): E105-E108.

Green D.M, Sweets, JA (1996). Signal Detection Theory and psychophysics, New York: Wiley.

Guoan C, Tarek G, Chiang-Ching H, et al.(2002). Proteomic analysis of lung adenocarcinoma: identifcation of a highly expressed set of proteins in tumors. Clin. Cancer Res;8:2298-2305.

Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. (1997). A Comparison of Parametric and Nonparametric Approaches to ROC Analysis of Quantitative Diagnostic Tests. Medical Decision Making;17:94–102.

Hammer SM, Eron JJ, Reiss P, et al.(2008). Antiretroviral treatment of adult HIV infection: 2008 recommendations of the International AIDS Society-USA panel. JAMA;300:555-570.

Hanley JA.(1988). The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests. Medical Decision Making;8:197–203.

Hanley JA.(1989). Receiver operating characteristic (ROC) methodology: The state of the art. Critical Reviews in Diagnostic Imaging; 29: 307-335.

Hanley JA. (1998) Receiver operating Characteristics (ROC) Curves. Encyclopedia of Biostatistics; 5: 3738-3745.

Hanley JA, McNeil BJ. (1982a). A Method of Comparing the Areas Under the Operating Characteristic Curves Derived from the Same Cases. Radiology. 1982;148:839-843.

Hanley JA, McNeil BJ. (1982b). The Meaning and Use of the Area Under the Operating Characteristic (ROC) Curve. Radiology 1982;143:29-36.

Hastie T. J, Tibshirani R.J. (1990). Generalized Additive Models, New York: Chapman and Hall.

Heagerty PJ, Lumley T, Pepe MS. (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics;56(2):337-44.

Harrel Jr FE. Regression Modeling Strategies. Springer: New York 2001.

Hirano K, Imbens GW (2004). The propensity score with continuous treatments. In "Applied Bayesian modeling and causal inference from incomplete data perspectives. An essential journey with Donald Rubin's statistical family". Wiley Series in Probability and Statistics, 73-84.
Hernán, Miguel A. (2010). The Hazards of Hazard Ratios. Epidemiology; 21(1): 13-15.

Hernán MA, Brumback B, Robins JM. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology, 11(5):561-570.

Hernán MA, Brumback B, Robins JM.(2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association, 96(454):440-448.

Hernán MA, Hernandez-Diaz S, Werler MM, et al.(2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol;155 (2) : 176-184.

Hernán MA, Robins JM. (2006).Estimating causal effects from epidemiological data. J Epidemiol Community Health;60;578-586.

Holland PW. (1986). Statistics and causal inference. Journal of the American Statistical Association; 81: 945–970.

Huang Y, Pepe M, Feng Z. (2007). Evaluating the predictiveness of a continuous marker. Biometrics;63:1181-8.

Janes H, Longton G. M, Pepe M. (2008). Accommodating Covariates in ROC Analysis. UW Biostatistics Working Paper Series. Working Paper 322.

Janes H, Pepe M. Adjusting for covariate effects on classification accuracy using the covariate-adjusted ROC curve. Technical Report 283, UW Biostatistics Working Paper Series, 2006.

Janes H, Pepe MS.(2008). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. Am J Epidemiol 168, 89–97.

Janes H, Pepe MS. (2008a). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. UW Biostatistics Working Paper Series. Working Paper 283.

Janes H, Pepe MS. (2008b). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. American Journal of Epidemiology;168:89–97.

Janes H, Pepe MS. (2008c). Matching in studies of classification accuracy: implications for bias, efficiency, and assessment of incremental value. Biometrics; 64:1–9.

Kaleebu, P.; et al. (2000). "Molecular epidemiology of HIV type 1 in a rural community in southwest Uganda". AIDS Research and Human Retroviruses **16** (5): 393–401.

Kaleebu, P.; et al. (2002). Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. Journal of Infectious Disease **185** (9): 1244–1250.

Kanki, P.J.; et al. (1999). Human immunodeficiency virus type 1 subtypes differ in disease progression. Journal of Infectious Disease **179** (1): 68–73.

Kang J, Schafer JL. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science, 2007: 22(4): 523–539.

Kaplan EL, Meier P.(1958). Non parametric estimation from incomplete

observations. J Am Stat Assoc 53:457–81.

Kawado M, Hashimoto S, Yamaguchi T, Oka S, Yoshizaki K, Kimura S, Fukutake K, Higasa S, Shirasaka T. (2006). Difference of progression to AIDS according to CD4 cell count, plasma HIV RNA level and the use of antiretroviral therapy among HIV patients infected through blood products in japan. J Epidemiol;16(3):101-6.

Kjetil S.(2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. Journal of Clinical Pathology;62:1-5.

Kish, L. (1965), Survey Sampling. John Wiley & Sons, New York.

Kish, L. (1990). Why, When, and How? Proceedings of the Survey Research Methods Section. American Statistical Association, PP. 121-130.

Kitahata MM et al. (2009). Effect of early versus deferred antiretroviral therapy for HIV on survival. The New England Journal of Medicine 360,1815–26.

Kiwanuka N, Robb M, Laeyendecker O, Kigozi G, Wabwire-Mangen F, Makumbi FE, Nalugoda F, Kagaayi J, Eller M, Eller LA, Serwadda D, Sewankambo NK, Reynolds SJ, Quinn TC, Gray RH, Wawer MJ, Whalen CC (2009). HIV-1 Viral Subtype Differences in the Rate of CD4+ T-Cell Decline Among HIV Seroincident Antiretroviral Naive Persons in Rakai District, Uganda. J Acquir Immune Defic Syndr;00:000–000).

Lagakos SW, Hoth DF. (1992). Surrogate markers in AIDS: where are we? Where are we going? Annals of Internal Medicine;116(7):599-601.

Lassere MN. (2007). The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker –surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. Statistical Methods in Medical Research;17:303-340.

Lassere MN, Johnson KR, Boers M, Tugwell P, Brooks P, Buyse M, Altman D, et al. (2007)Definitions and Validation Criteria for Biomarkers and Surrogate Endpoints: Development and Testing of a Quantitative Hierarchical Levels of Evidence Schema. J Rheumatol; 34:607–15.

Le Moing V, Thiebaut R, Chene G *et al.*(2007). Long-term evolution of CD4 count in patients with a plasma HIV RNA persistently <500 copies/ml during treatment with antiretroviral drugs. *HIV Med.;* 8(3);156–163.

Levy S.E, Statnikov A, Aliferis C. (2005). Biomarker Selection from High-Dimensionality Data. Microarray Technology.

Li J, Fine JP. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics; 9: 566–76.

Liang, K.-Y., and Zeger, S. L. (1986), .Longitudinal Data Analysis Using Generalized Linear Models,. Biometrika, 73, 13.22.

Liotta G, Perno CF, Ceffa S, Gialloreti LE, Coehlo E, Erba F, Guidotti G, Marazzi MC, Narciso P, Palombi L. (2004). Is total lymphocyte count a reliable predictor of the CD4 lymphocyte cell count in resource-limited settings? AIDS;18(7):1082-3.

Lin, D.Y. and Wei, L.J. (1989), "The Robust Inference for the Proportional Hazards Model," Journal of the American Statistical Association, 84, 1074–1078.

Lunceford JK, Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine; 23: 2937–2960.

Machado SG, Gail MH, Ellenberg SS. (1990). On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of treatment for HIV infection. J Acquir Immune Defic Syndr;3(11):1065-73.

Maclachlan E, Mayer KH, Barnabas R, Sanchez J, Koblin B, Duerr A. (2009). The Potential Role of Biomarkers in HIV Preventive Vaccine Trials: A Critical Review. J Acquir Immune Defic Syndr; 51(5): 536-545.

Martin DJ, Sim JG, Sole GJ, Rymer L, Shalekoff S, van Niekerk AB, Becker P, Weilbach CN, Iwanik J, Keddy K (1995). CD4+ lymphocyte count in African patients co-infected with HIV and tuberculosis. J Acquir Immune Defic Syndr Hum Retrovirol.;8(4):386-91.

Mayeux, R.(2004). Biomarkers: potential uses and limitations. NeuroRx;1(2):182-8.

Mellors JW, Muñoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, Detels R, Phair JP, Rinaldo CR Jr. (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. Ann Intern Med;126(12):983-5.

McClish, D. (1989). Analyzing a portion of the ROC curve. Medical Decision Making 9,190–195.

McIntosh M, Pepe M. (2002). Combining several screening tests: Optimality of the risk score. Biometrics 58, 657-64.

Mendoza C, Holguin A, Soriano V (1998). False positives for HIV using commercial viral load quantification assays. AIDS; 12(15); 2076-2077.

Metz CE. Basic Principles of ROC analysis. Semin nucl med. 1978;8:283–298.

Molenberghs G, Burzykowski T, Alonso A, Buyse, M. (2004). A perspective on surrogate endpoints in controlled clinical trials. Statistical Methods in Medical Research; 13:177-206.

Molenbergh G, Burzykowski T, Alonso A, Buyse M. (2009). A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. Statistical Methods in Medical Research;00 ;1-32.

Montaner JS, Le TN, Craib KJ, Schechter MT. (1992). Application of the World Health Organization system for HIV infection in a cohort of homosexual men in developing a prognostically meaningful staging system. AIDS;6(7):719-24.

Moore K, Neugebauer R, Lurmann F, Hall J, Brajer V, Alcorn S, Tager I. (2010). . Ambient Ozone Concentrations and Cardiac Mortality in Southern California 1983-2000: Application of a New Marginal Structural Model Approach. Am J Epidemiol. (Advance Access published May 3, 2010).

Moore KL, Van der Laan MJ.(2009). Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. Statistics in medicine 28,39-64.

Morrison CS, Demers K, Kwok C, Bulime S, Rinaldi A, Munjoma M, Dunbar M, Chipato T, Byamugisha J, Van Der Pol B, Arts E, Salata RA. (2010). Plasma and cervical viral loads among Ugandan and Zimbabwean women during acute and early HIV-1 infection. AIDS. 24(4):573-82.

Morrison CS, Richardson BA, Mmiro F, Chipato T, Celentano DD, Luoto J, Mugerwa R, Padian N, Rugpao S, Brown JM, Cornelisse P, Salata RA. (2007). Hormonal contraception and the risk of HIV acquisition. AIDS 21, 85-95.

Mwamburi DM, Ghosh M, Fauntleroy J, Gorbach SL, Wanke CA. (2005). Predicting CD4 count using total lymphocyte count: a sustainable tool for clinical decisions during HAART use. Am J Trop Med Hyg. 73(1):58-62.

Musial J, Swadzba J, Motyl A, et al. (2003). Clinical significance of antiphospholipid protein antibodies. Receiver operating characteristics plot analysis. J Rheumatol;30: 723–30.

Neugebauer R, Van der Laan MJ. (2005). Why Prefer Double Robust Estimators in Causal Inference? Journal of Statistical Planning and Inference 19, 405-426.

No authors listed. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The

Cardiac Arrhythmia Suppression Trial (CAST) Investigators. N Engl J Med. 1989, Aug 10;321(6):406-12.

Obuchowski NA. (2005). ROC Analysis. AJR Am J Roentgenol 184, 364-72.

Obuchowski NA, Lieber ML, Wians FH Jr. (2004). ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem;50:1118–25.

Paice JA, Ferrans CE, Lashley FR, Shott S, Vizgirda V, Pitrak D (2000): Topical Capsaicin in the Management of HIV-Associated Peripheral Neuropathy. Journal of Pain and Symptom Management; 19:45-52.

Panel on Antiretroviral Guidelines for Adults and Adolescents. (2008). Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Washington, DC: Department of Health and Human Services,1-139.

Pencina MJ, D'Agostino RB. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat. Med; 23 (13),2109–2123.

Petersen ML, Wang Y, Van der Laan MJ, Bangsberg DR (2006). Assessing the Effectiveness of Antiretroviral Adherence Interventions: Using Marginal Structural Models to Replicate the Findings of Randomized Controlled Trials." JAIDS 43.Supplement 1: S96-S103.

Planella T, Cortés M, Martínez-Brú C, Barrio J, Sambeat MA, González-Sastre F. (1998).The predictive value of several markers in the progression to acquired immunodeficiency syndrome. Clin Chem Lab Med;36(3):169-73.

Polley EC, van der Laan MJ (2010). "Super Learner In Prediction". *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 266. Available at http://www.bepress.com/ucbbiostat/paper266.

Pepe, MS. (2000). Receiver Operating Characteristic Methodology. Journal of the American Statistical Association 95, 308-311.

Pepe, MS. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.

Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. (2008). Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol;167:362-8.

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am. J. Epidemiol, 2004; 159(9): 882–890.

Pepe MS, Janes HE. (2008). Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer (Editorial). J Natl Cancer Inst.;100:978-9.

Pepe M, Longton GM, Janes H. Estimation and Comparison of Receiver Operating Characteristic Curves. (2008). UW Biostatistics Working Paper Series.Working Paper 323. Available at: http://www.bepress.com/uwbiostat/paper323.

Pollard KS, Dudoit S, Van der Laan MJ. Multiple Testing Procedures. (2004). R multtest Package and Applications to Genomics. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 164. Available at http://www.bepress.com/ucbbiostat/paper164.

Prentice RL. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. Stat Med; 8: 431–440.

Qu Y, Adam B.L, Yasui Y, et al. (2002). Boosted Decision Tree Analysis of Surface-enhanced Laser Desorption/Ionization Mass Spectral Serum Profiles Discriminates Prostate Cancer from Noncancer Patients. Clin Chem; 48(10):1835–1843.

Qu Y, Case M. (2006). Quantifying the indirect treatment effect via surrogate markers. Statistics in Medicine; 25: 223-231.

Robbins GK, Spritzler JG, Chan ES *et al.*(2009). Incomplete reconstitution of T cell subsets on combination antiretroviral therapy in the AIDS Clinical Trials Group protocol 384. *Clin. Infect. Dis.;*48(3); 350–361.

Robins JM. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy survivor effect. Mathematical Modelling 7, 1393–1512.

Robins JM (1995). An analytic method for randomized trials with informative censoring: Part I. Lifetime Data analysis 1: 241-254.

Robins JM (1999). Robust estimation in sequentially ignorable missing data and causal inference models. In Proceedings of the American Statistical Association: Sectionon Bayesian Statistical Science, pages 6-10.

Robins JM (2000). Marginal structural models versus structural nested models as tools for causal inference. In Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997), pages 95-133. Springer, New York.

Robins JM, Hernán MA, Brumback B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology 11:550-560.

Rosenbaum, PR. (2002). Observational studies (2nd ed.). New York: Springer.

Rosenbaum P, Rubin, DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika,. 1983;70: 41.55.

Rosenblum M, van der Laan MJ (2010). Targeted Maximum Likelihood Estimation of the Parameter of a Marginal Structural Model. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 257. http://www.bepress.com/ucbbiostat/paper257.

Rubin DB: Bayesian-inference for causal effects: the role of randomization. *Ann Stat* 1978, 6:34-58.

Rubin DB. (2004). Causal effects via potential outcomes. Scandinavian Journal of Statistics. 2004; 31, 161–170.

Rubin DB. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association. 2005;100: 322–331.

Rubin DB, Thomas N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc; 95: 573–85.

Rubin DB, Thomas N. (1996). Matching using estimated propensity scores: relating theory to practice. Biometrics;52:249–64.

Rutter MK, Meigs JB, Sullivan LM, D'Agostino RB Sr, Wilson PW. (2004). C-reactive protein, the metabolic syndrome, and prediction of cardiovascular events in the Framingham Offspring Study. Circulation; 110(4): 380–385.

Sato T, Matsuyama Y.(2003). Marginal structural models as a tool for standardization. Epidemiology;14:680–6.

Sax PE, Baden LR. (2009). When to start antiretroviral therapy-ready when you are. The New England Journal of Medicine 360,1897–9.

Schafer JL, Kang J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods 13, 279-313.

Shapiro NI, Karras DJ, Leech SH, Heilpern KL. (1998). Absolute lymphocyte count as a predictor of CD4 cell count. Ann Emerg Med;32(3 Pt 1):323-8.

Shlipak MG, Fried LF, Cushman M et al. (2005). Cardiovascular mortality risk in chronic kidney disease: comparison of traditional and novel risk factors. JAMA; 293(14): 1737–1745.

Sinisi SE, van der Laan MJ. Deletion/Substitution/Addition algorithm in learning with applications in genomics. Statistical Applications in Genetics and Molecular Biology, 3(1):Article18.

Stitelman OM, Van der Laan MJ. (2010). "Collaborative Targeted Maximum Likelihood For Time To Event Data". *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 260. Available at http://www.bepress.com/ucbbiostat/paper260.

Storey JD. A direct approach to false discovery rates. (2002). Journal of the Royal Statistical Society Series B; 64(3): 479-498.

Storey, JD. The Positive False Discovery Rate. (2003). A Bayesian Interpretation and the q-Value. The Annals of Statistics; 31(6): 2013-2035.

Storey JD, Tibshirani R. (2001). Technical Report 2001-28. Department of Statistics, Stanford University, 2001.

Storey JD, Tibshirani R. (2003). "Statistical Significance for Genomewide Studies" in Proceedings of the National Academy of Sciences of the United States of America, volume, 2003;100: 9440–9445.

Temple, RJ. (1995). A regulatory authority's opinion about surrogate endpoints. Clinical Measurement in Drug Evaluation. W.S. Nimmo & G.T. Tucker eds, Wiley, New York.

Thompson ML, Zuccchini W. (1989). On the Statistical Analysis of ROC Curves. Statistics in Medicine; 8: 1277-1290.

Thomson Scientific. Establishing the standards in biomarker research. White Paper, March 2008.

Touloumi G, Pocock SJ, Babiker AG, et al. (1999)Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Stat Med*. 18:1215-1233.

Tuglus C, Van der Laan M. J. (2008).Targeted Methods for Biomarker Discovery, the Search for a Standard. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 233.

Tusher V.G, Tibshirani R, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. PNAS; 98(9):5116–5121.

US Food and Drug Administration. Innovation/Stagnation: Critical Path Opportunities Report. March 2006. Available at http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalP

athOpportunitiesReports/default.htm.

Van der Laan, M J. (2006). Statistical Inference for Variable Importance. The International Journal of Biostatistics 2, Issue 1, Article 11.

Van der Laan MJ, Polley EC, Hubbard AE. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(25):Article 25.

Van der Laan MJ, Rubin D. (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics 2, Issue 1, Article 11.

Van der Laan MJ, Rose S, Gruber S. (2009). Readings in Targeted Maximum Likelihood Estimation. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 254. http://www.bepress.com/ucbbiostat/paper254.

Van der Meer IM, De Maat MP, Kiliaan AJ, van der Kuip DA, Hofman A, Witteman JC.(2003). The value of C-reactive protein in cardiovascular risk prediction: the Rotterdam Study. Arch. Intern. Med; 163(11):1323–1328.

Van Der Pol B, Kwok C, Pierre-Louis B, Salata RA, Chen PL, Morrison CS. (2008). Trichomonas vaginalis infection and human immunodeficiency virus acquisition in African women. J Infect Dis. 15,548-54.

Van der Ryst E, Kotze M, Joubert G, Steyn M, Pieters H, et al. Correlation among total lymphocyte count, absolute CD4+ count, and CD4+ percentage in a group of HIV-1-infected South African patients.(1998). J Acquir Immune Defic Syndr Hum Retrovirol;19(3):238-44.

Vapnik V.N. (1998). Statistical Learning Theory. Wiley-Interscience.

Wagner JA. (2009). Biomarkers: Principles, Policies, and Practice. Clinical Pharmacology & Therapeutics 86, 3–7.

Wang Y, Petersen ML, Bangsberg D, van der Laan MJ (2006). Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. UC Berkeley Division of Biostatistics working paper series, 2006.

Wieand S, Gail MH, James BR, James KL. (1989). A Family of Nonparametric Statistics for Comparing Diagnostic Markers With paired or Unpaired Data. Biometrika, 1989; 76:585-592.

Weir CJ, Walley RJ. (2006). Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Statistics in Medicine;25(2):183-203.

West SG, Biesanz JC, Pitts SC. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T.Reis & C. M.Judd (Eds.), Handbook of research methods in social and personality psychology (pp. 40–84). New York: Cambridge University Press.

Westfall PH, Tobias RD, Rom D, Wolfinger RD,Hochberg, Y. (1999), Multiple Comparisons and Multiple Tests Using the SAS System, Cary, NC: SAS Institute Inc.

Weston AD, Hood L. (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res; 3(2):179-96.

Westreich D, Cole SR, Tien PC, Chmiel JS, Kingsley L, Funk MJ, Anastos K, Jacobson LP. (2010). Time scale and adjusted survival curves for marginal structural cox models. Am J Epidemiol. 2010 Mar 15;171(6):691-700. 5.

When to Start Consortium. (2009). Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. Lancet 373, 1352–63.

Wilson PW, Nam BH, Pencina M, D'Agostino RB Sr, Benjamin EJ, O'Donnell CJ. (2005). C-reactive protein and risk of cardiovascular disease in men and women from the Framingham Heart Study. Arch. Intern. Med; 165(21): 2473–2478.

Winship, C., Morgan, S. L (1999). The estimation of causal effects from observational data. Annual Review of Sociology; 25: 650–707.

Winship, C., Sobel, M. E. Causal inference in sociological studies. In M.Hardy (Ed.), Handbook of data analysis (pp. 481–504). Thousand Oaks, CA: Sage.

Woo MJ, Reiter JP, Karr AF. 92008). Estimation of propensity scores using generalized additive models. Stat Med; 27(19):3805-16.

Wu B, Abbott T, Fishman D, et al. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics; 19(13): 1636–1643.

Xiao Y, Abrahamowicz M, Moodie EM. (2010). Accuracy of Conventional and Marginal Structural Cox Model Estimators: A Simulation Study. The International Journal of Biostatistics, Vol. 6, Iss. 2, Art. 13.

Yekutieli D, Benjamini Y. Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics. (1999). J Statist Planning Inference; 82: 171–196.

Young JG, Hernán MA, Picciotto S, Robins JM. (2009). Relation between three

classes of structural models for the effect of a time-varying exposure on survival. Lifetime Data Analysis, 16(1):71-84.

Yu W, Wu B, Huang T, Li X, Williams K, Zhao H. (2006). “Statistical Methods In Proteomics" in Springer Handbook of Engineering Statistics (editor: H. Pham), Springer-Verlag, London, UK, pp. 623-638.

Zhang, DD, Zhou, XH, Freeman, JM, Freeman, DH. (2002). A Non-parametric Method on the Comparison of Partial Areas Under ROC Curves and its Application on Large Data Sets. Statistics in Medicine; 21:5, 701-715.

Zeger, S. L., and Liang, K.-Y. (1986) .Longitudinal Data Analysis for Discrete and Continuous Outcomes,. Biometrics, 42, 121.130.

Zheng Y, Cai T, Feng Z. (2006). Application of the time dependent ROC curves for prognostic accuracy with multiple biomarkers. Biometrics;62:279–87.

Zou KH, Hall WJ, Shapiro DE. (1996). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in Medicine. 1996;16:2143–2156.

Zou KH, O’Malley AJ, Mauri L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation;115:654–7.

Zweig MH, Campbell G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem;39:561–77.

TABLES AND FIGURES

Table 1: Estimates of Variable Importance Measures for Each Biomarker and Associated p-values

	TMLE			Weighted Logistic Regression			Partial ROC (0.05)		
Biomarker	**VIM**	**Unadjusted p-value**	**p-value [a]**	**VIM**	**Unadjusted p-value**	**p-value [a]**	**Incremental partial AUC [b]**	**Unadjusted p-value**	**p-value [a]**
CD4 at baseline	0.2064	0.0004	0.0070	0.2316	0.0007	0.0079	0.0082	0.3100	1.0000
Plasma Viral Load	0.0519	0.2784	1.0000	0.1537	0.0024	0.0200	0.0029	0.2400	1.0000
CD8 Count	0.0535	0.2705	1.0000	0.1410	0.0920	0.5092	0.0013	0.7300	1.0000
CD4/CD8 T cell Ratio	0.1071	0.0003	0.0070	0.1113	0.0004	0.0059	0.0002	0.9100	1.0000
Lymphocyte count	0.1044	0.1714	1.0000	0.1843	0.2298	0.8720	0.0006	0.6100	1.0000
CD4 percent	0.2779	0.0103	0.1138	0.3663	0.0514	0.3414	0.0091	0.2700	1.0000
HIV subtype A	0.1908	0.5362	1.0000	0.2318	0.3434	1.0000	0.0000	1.0000	1.0000
HIV subtype C	-.06687	0.8731	1.0000	0.1827	0.1809	0.8586	0.0000	0.8400	1.0000
HIV subtype D	0.2200	0.4789	1.0000	0.1295	0.4519	1.0000	0.0001	1.0000	1.0000
Hemoglobin	0.0471	0.4225	1.0000	0.1261	0.2363	0.8720	0.0001	0.9700	1.0000
HSV-2 status	0.0078	0.8667	1.0000	0.1088	<.0001	<.0001	0.0000	1.0000	1.0000

Note: Results based on 5000 bootstrap samples.

[a] Adjusted for multiplicity based on Benjamini & Yekutieli Dependent False Discovery Rate (FDR) procedure.

[b] At a false positive rate of 0.05.

Table 2: Results from 5000 bootstrap samples for five simulated biomarkers based on Targeted maximum Likelihood Estimation for Variable importance Measure

	N=150			N=200			N=250		
Biomarker	**VIM**	**Unadjusted p-value**	**p-value***	**VIM**	**Unadjusted p-value**	**p-value***	**VIM**	**Unadjusted p-value**	**p-value***
A1	0.087	0.7040	1.0000	0.028	0.8923	1.0000	0.021	0.9103	1.0000
A2	0.339	<.0001	<.0001	0.314	<.0001	<.0001	0.321	<0.0001	<.0001
A3	0.218	0.0239	0.1367	0.211	0.0113	0.0643	0.218	0.0029	0.0166
A4	-0.036	0.6873	1.0000	-0.057	0.3855	1.0000	-0.056	0.3379	1.0000
A5	-0.043	0.4745	1.0000	-0.034	0.4958	1.0000	-0.034	0.4713	1.0000

* Adjusted for multiplicity based on Benjamini & Yekutieli Dependent False Discovery Rate (FDR).

Table 3: Results from 5000 bootstrap samples for five simulated biomarkers based on Propensity Score Weighting for Variable importance Measure

	N=150			N=200			N=250		
Biomarker	**VIM**	**Unadjusted p-value**	**p-value***	**VIM**	**Unadjusted p-value**	**p-value***	**VIM**	**Unadjusted p-value**	**p-value***
A1	0.111	0.2173	0.6203	0.111	0.0666	0.2536	0.084	0.1693	0.4832
A2	0.310	<.0001	<.0001	0.284	<.0001	<.0001	0.301	<.0001	<.0001
A3	0.189	0.0088	0.0505	0.193	0.0001	0.0006	0.204	<.0001	<.0001
A4	-0.044	0.6438	1.0000	-0.070	0.1533	0.4374	-0.056	0.0298	0.1134
A5	-0.059	0.2027	0.6203	-0.042	0.2543	0.5806	-0.036	0.3678	0.8398

* Adjusted for multiplicity based on Benjamini & Yekutieli Dependent False Discovery Rate (FDR).

Table 4: Results from 5000 bootstrap samples for five simulated biomarkers based on Incremental Value estimation for partial area under the curve

	N=150			N=200			N=250		
Biomarker	**ΔpAUC (0.05)**	**Unadjusted p-value**	**p-value***	**ΔpAUC (0.05)**	**Unadjusted p-value**	**p-value***	**ΔpAUC (0.05)**	**Unadjusted p-value**	**p-value***
A1	.001	0.0660	0.3768	.001	0.0600	0.3425	.001	.1600	0.6089
A2	.006	0.1400	0.5328	.004	0.1300	0.4947	.010	.0016	0.0091
A3	.005	0.0200	0.2283	.005	0.0007	0.0081	.007	<.0001	<.0001
A4	-.001	0.3800	1.0000	-.001	0.4800	1.0000	.000	.9800	1.0000
A5	.001	0.5200	1.0000	.000	0.7400	1.0000	.001	.5100	1.0000

* Adjusted for multiplicity based on Benjamini & Yekutieli Dependent False Discovery Rate (FDR).

Table 5: Longitudinal Targeted Variable Importance Estimates for HIV Infection Biomarkers

	Method 1: Regression with Autocorrelated Errors			Method 2:Area Under LOESS Curve Estimation		
Biomarker	**EstimatedTime Slope Coefficient**	**Unadjusted p-value**	**p-value [a]**	**Estimated Area under the LOESS Curve**	**Unadjusted p-value**	**p-value [a]**
CD4 at baseline	0.0083	<.0001	<.0001	0.7817	<.0001	<.0001
Plasma Viral Load	0.0001	0.9021	1.0000	0.0099	0.9490	1.0000
CD8 Count	0.0042	0.1083	0.5994	0.3287	0.1381	0.7647
CD4/CD8 T-cell Ratio	0.0039	0.0811	0.5388	0.4362	0.0376	0.2497
Lymphocyte count	0.0002	0.9126	1.0000	0.1213	0.5520	1.0000
CD4 percent	-.0012	0.9144	1.0000	0.0731	0.9339	1.0000
HIV subtype A	-.0147	<.0001	<.0001	1.580	<.0001	<.0001
HIV subtype C	-.0133	0.1765	0.8374	1.139	0.3525	1.0000
HIV subtype D	0.0109	<.0001	<.0001	0.9169	<.0001	<.0001
Hemoglobin	0.0013	0.5066	1.0000	0.0053	0.9706	1.0000
HSV-2 positive	-.0092	<.0001	<.0001	0.8209	<.0001	<.0001

Note: Results based on 5000 bootstrap samples.
[a] Adjusted for multiplicity based on Benjamini & Yekutieli Dependent False Discovery Rate (FDR) procedure

Table 6: VIM as a Function of Time in Simulated Data

	Method 1: Regression with Autocorrelated Errors	Method 2:Area Under LOESS Curve Estimation
Number of time points	**p-value** [a]	**p-value** [a]
5	0.9448	0.3879
10	0.2619	0.0042
20	0.0003	<0.0001
25	0.0002	<0.0001

Note: Results based on 5000 bootstrap samples.

Table 7: Distribution of Stabilized Weights by Biomarker

Biomarker	Mean (SD)	Minimum - Maximum
Baseline CD4 Cell count	1.000 (0.003)	0.966 - 1.029
CD8 Cells Count	1.001 (0.060)	0.503 - 1.941
CD4/CD8 T Cell Ratio	1.001 (0.026)	0.821 - 1.261
CD4 Percentage	1.000 (0.008)	0.832 - 1.038
Lymphocyte Count	0.997 (0.069)	0.493 - 1.929
Plasma Viral Load	1.001 (0.032)	0.737 - 1.340
Hemoglobin Level	1.003 (0.067)	0.521 - 2.236
HSV-2 Status	1.006 (0.071)	0.696 - 1.535
HIV Subtype A	1.040 (0.164)	0.329 - 3.372
HIV Subtype C	1.015 (0.126)	0.409 - 3.179
HIV Subtype D	1.028 (0.118)	0.425 - 2.781

Abbreviation: SD=Standard Deviation

Table 8: Overall survival Through 5 Years Post Estimated Infection Date

Biomarker		N	n(%)	5-year Cumulative probability of survival (%) and 2-sided 95% CI [a]
CD4 at baseline ≤ 500 cells/mm^3	No	176	30(17.05)	84.86(79.11-90.62)
	Yes	126	82(65.08)	35.12(26.55-43.68)
CD4 Percentage ≤ 20 %	No	244	70(28.69)	72.05(66.08-78.03)
	Yes	58	42(72.41)	30.44(18.46-42.43)
CD8 Count ≤ 1000 cells/mm^3	No	210	75(35.71)	63.85(56.98-70.72)
	Yes	92	37(40.22)	64.09(54.07-74.11)
Lymphocyte count < 1200 cells/mm^3	No	250	79(31.60)	69.98(64.03-75.93)
	Yes	52	33(63.46)	34.83(21.25-48.40)
CD4/CD8 Ratio ≤ 1	No	66	10(15.15)	87.71(79.08-96.34)
	Yes	236	102(43.22)	57.28(50.69-63.87)
HIV RNA > 55000 copies/ml	No	143	45(31.47)	69.37(61.59-77.15)
	Yes	79	37(46.84)	58.50(47.46-69.54)
HIV Subtype A	No	113	56(49.56)	53.18(43.73-62.63)
	Yes	76	20(26.32)	74.62(64.35-84.88)
HIV Subtype C	No	107	35(32.71)	67.26(57.71-76.80)
	Yes	82	41(50.00)	53.78(42.85-64.72)
HIV Subtype D	No	158	61(38.61)	63.62(55.89-71.35)
	Yes	31	15(48.39)	48.77(28.82-68.72)
HSV-2 positive	No	42	16(38.10)	60.12(44.78-75.47)
	Yes	96	49(51.04)	54.20(44.04-64.35)
Hemoglobin Level < 11g/dl	No	233	81(34.76)	66.21(59.88-72.54)
	Yes	69	31(44.93)	56.13(43.75-68.50)

Abbreviations: N=number of subjects in category, n=number of subjects with event, CI=Confidence Intervals.
[a] Estimates based on the Kaplan-Meier method.
Note: survival time for those with either two successive drops of <=350 cells/mm^3 in CD4 cell counts or two drops of <=350 cells/mm^3 in CD4 cell counts within six months is defined as date of second drop minus estimated date of infection.

Table 9: Estimates of Variable Importance Measures for Each Biomarker and Corresponding 95% Confidence Intervals from Standard Cox Proportional Hazards Model

Biomarker	Log Hazard Ratio	Hazard Ratio (2-sided 95% CI)	Unadjusted p-value	p-value [a]
Baseline CD4 Cell count	-0.763	0.466(0.379,0.573)	<.0001	<.0001
CD4/CD8 T cell Ratio	-0.593	0.552(0.447,0.683)	<.0001	<.0001
Lymphocyte Count	-0.351	0.704(0.586,0.846)	0.0002	0.0021
Plasma Viral Load	0.349	1.417(1.171,1.715)	0.0003	0.0028
Hemoglobin Level	-0.156	0.855(0.736,0.994)	0.0408	0.2713
HIV Subtype C	0.936	2.551(0.590,11.032)	0.2100	1.0000
CD8 Cell Counts	0.072	1.075(0.909,1.272)	0.3998	1.0000
HIV Subtype A	-0.196	0.822(0.468,1.443)	0.4941	1.0000
CD4 Percent	-0.045	0.956(0.810,1.129)	0.5963	1.0000
HIV Subtype D	0.102	1.107(0.620,1.979)	0.7303	1.0000
HSV-2 Status	-0.013	0.987(0.637,1.529)	0.9537	1.0000

Abbreviation: CI=Confidence Intervals.
[a] Adjusted for multiplicity based on Benjamini & Yekutieli Dependent False Discovery Rate (FDR) procedure.

Table 10: Weighted Estimates of Variable Importance Measures for Each Biomarker and Corresponding 95% Confidence Intervals

Biomarker	Log Hazard Ratio	Hazard Ratio (2-sided 95% CI)	Unadjusted p-value	p-value [a]
Baseline CD4 Cell count	-0.674	0.510(0.418,0.621)	<.0001	<.0001
CD4/CD8 T- Cell Ratio	-0.508	0.602(0.480,0.754)	<.0001	0.0002
Lymphocyte Count	-0.297	0.743(0.640,0.862)	<.0001	0.0010
Plasma Viral Load	0.306	1.358(1.141,1.615)	0.0006	0.0046
Hemoglobin Level	-0.163	0.850(0.744,0.970)	0.0162	0.1073
CD4 Percent	-0.065	0.937(0.836,1.051)	0.2688	1.0000
HIV Subtype D	0.184	1.202(0.720,2.006)	0.4815	1.0000
CD8 Cell Counts	0.051	1.052(0.895,1.237)	0.5359	1.0000
HIV Subtype A	-0.111	0.895(0.608,1.317)	0.5739	1.0000
HIV Subtype C	0.042	1.043(0.708,1.536)	0.8318	1.0000
HSV-2 Status	-0.013	0.987(0.638,1.527)	0.9542	1.0000

Abbreviation: CI=Confidence Intervals.

[a] Adjusted for multiplicity based on Benjamini & Yekutieli Dependent False Discovery Rate (FDR) procedure.

Figure 1: Plots of TVIM Data with Robust Smoother For Biomarkers with a P-value <.05

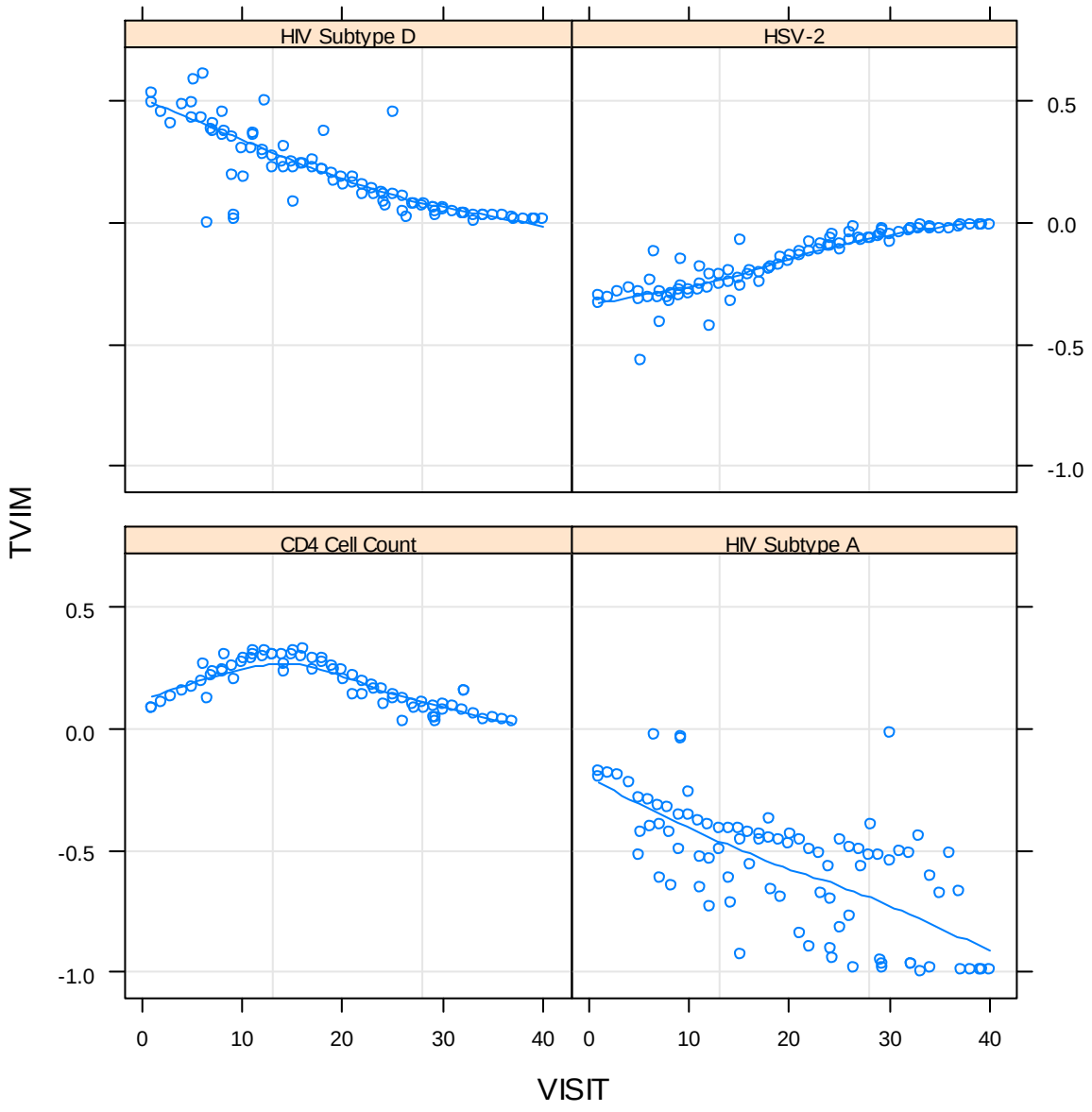

Figure 2: Plots of TVIM Data with Robust Smoother For Biomarkers with a P-value > .05

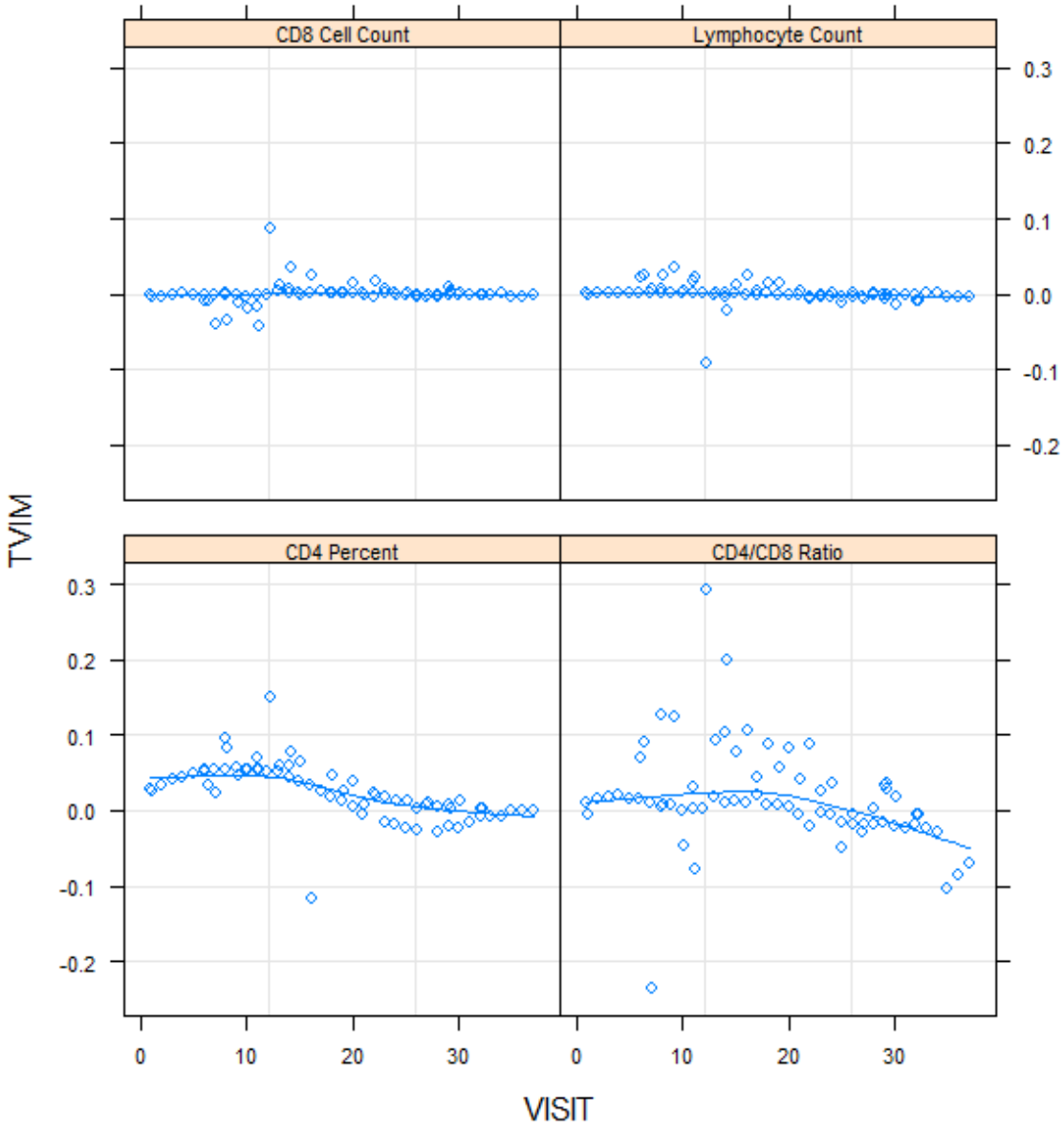

Figure 2: Plots of TVIM Data with Robust Smoother For Biomarkers with a P-value > .05