Evaluation of Potential Surrogate Endpoints

Erin E. Gabriel

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Peter B. Gilbert, Chair

Holly Janes

Ying Huang

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Evaluation of Potential Surrogate Endpoints

Erin E. Gabriel

Chair of the Supervisory Committee:
Professor Peter B. Gilbert
Department of Biostatistics

Valid surrogate endpoints can make clinical trials more efficient, allowing for more trials to be conducted and more rapid development of effective treatments. Identifying useful surrogates is a statistically challenging but extremely valuable endeavor. This work develops statistical methods for the evaluation and comparison of biomarkers as correlates of protection (CoP). Methods herein were developed with a focus on a time-to-event clinical endpoint and possible time-varying effects of treatment, an important and thus far neglected topic in CoP evaluation. We propose a novel Weibull model and three methods of estimation for use in CoP evaluation. Simulations and real data examples demonstrate the characteristics of these methods.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

INTRODUCTION AND BACKGROUND

1.1 Introduction

Valid surrogate endpoints can make clinical trials more efficient, allowing for more trials to be conducted and more rapid development of effective treatments. The measurement of immune responses to treatment, particularly to vaccination, that are correlated with clinical outcome are of great interest due to their potential to be used as surrogates. Not all immune responses that are correlated with risk will be useful surrogates. Identifying correlates that will be useful surrogates is a statistically challenging but extremely valuable endeavor.

The use of surrogates has had a long and controversial history. Many assumed surrogate endpoints were actually misused correlates of risk that led to ineffective and sometimes harmful treatments being approved for use (Fleming and Demets, 1996). The need to identify true surrogates of protection has been recognized as a top priority in many fields. This is particularly true for Human Immunodeficiency Virus (HIV) vaccine research.

For many, HIV has become a chronic illness that can be controlled by the use of highly active antiretroviral therapy (HAART). Due to the efforts of groups like the Gates Foundation, even in developing nations, where as recently as five years ago HAART was unavailable or too costly, many people are receiving the necessary medication to control their HIV symptoms. However, even in more developed nations, HIV medication does not always make it to those who need it and not all patients can tolerate the side effects of HAART. Every year HIV infection costs billions of dollars in health care expenditures. A safe, relatively inexpensive and effective preventative vaccine would reduce costs significantly and save lives. Due to the unique challenges of HIV vaccine development, identifying an immune response that reliably predicts protection against HIV infection would be a major breakthrough.

There are many existing paradigms for evaluation of potential surrogates. We focus on the principal stratification paradigm of Frangakis and Rubin (2002), under which they define a principal

surrogate. Gilbert et al. (2008) and Qin et al. (2007) refine the definition of principal surrogate in the vaccine efficacy setting to a specific surrogate of protection (SoP). Gilbert et al. (2008) define a SoP to be correlated with clinical outcome for those treated and serve as a reliable predictor of vaccine efficacy. Qin et al. (2007) also defined a correlate of risk (CoR) to be a biomarker that is correlated with the clinical outcome in the vaccine arm.

In this dissertation, we describe methods for testing and comparing the quality of potential CoRs and SoPs. Herein, new methods developed as part of this research are tested both in simulation and real data examples. Real data examples are from the RV144 and Step HIV vaccine trial (Rerks-Ngarm et al., 2009; McElrath et al., 2008) and the Zoster efficacy and safety trials (ZEST) (Schmader et al., 2010). Although all data used in this dissertation are from vaccine trials, that is not the only context in which these methods could be useful.

1.2 Motivation for Extension of Exiting Methods

Although among all the methods for surrogate evaluation there are some that are designed to be used with a time-to-event outcome, this is not the case for SoP evaluation. As time-to-event is the primary outcome currently used in many clinical trials, this is a weakness of the literature. No method of specific surrogate of protection evaluation, to our knowledge, allows for the characterization of changes in the surrogate and clinical outcome relationship over time. These time-varying effects can be very important when evaluating a candidate surrogate as durability of surrogacy and treatment effect can impact the usefulness of the surrogate as well as the treatment. This is especially true in the vaccine setting. Figure 1.1 displays the observed waning of VE in the RV144 HIV vaccine trial. Without methods to evaluate and compare candidate SoP allowing for time variation, useful SoP may be missed due to rapid VE waning and SoP with time-varying relationships with the clinical outcome could be used inappropriately to estimate lasting treatment effects.

1.3 Chapter Outline

In the remainder of Chapter 1, we review some of the unique biological challenges to HIV vaccine development. We outline the Phase IIb and Phase III HIV vaccine trials to date. We review several of the statistical features of HIV vaccine trial data. Specifically, we outline the features of immune

response to vaccination data, as these are the data where potential CoR, SoP are found. We then give a similar background for Varicella- Zoster and the licensed vaccine Zostavax®.

In Chapter 2, we outline the concept of a CoR evaluation and introduce a time-to-event method that allows for evaluation of CoR in the presence of time-varying associations. We apply this method to potential correlates identified in the primary RV144 correlates analysis (Haynes et al., 2012). We extend the pseudoscore method of Chatterjee et al. (2003) to estimate a novel Weibull model while accounting for the two-phase case-control sampling that is common in immune correlate studies.

In Chapter 3, we review a number of paradigms proposed for surrogate endpoint evaluation. We begin with Prentice's paradigm and after showing some limitations, we outline principal stratification (Prentice, 1989; Frangakis and Rubin, 2002). Principal stratification (PS) is the paradigm under which we introduce our proposed methods of surrogate evaluation. We then summarize the direct and indirect effects paradigm (Taylor et al., 2005a; Robins and Greenland, 1992a), the meta-analysis paradigm (Daniels and Hughes, 1997) and the recently introduced paradigm of Pearl (2011).

In Chapter 4, we outline the existing methods for identification and estimation of SoP estimands of interest under principal stratification. We examine the augmented trial designs of Follmann (2006) and the proposed two-phase sequential trial design of Gilbert et al. (2011b). We then summarize the estimated likelihood methods of Follmann (2006) and Gilbert and Hudgens (2008); the semi-parametric method of Huang and Gilbert (2011), the non-parametric method of Gilbert and Hudgens (2008), the Bayesian estimation method of Li et al. (2010) and the pseudoscore method of Wolfson (2009). We also review the discrete time-to-event evaluation method of Qin et al. (2008) which, although not causal in its estimand, is of interest in evaluating surrogates.

In Chapter 5, we introduce our proposed extension to the SoP evaluation framework to allow for time-varying treatment effects. We extend the concept of surrogate-dependent vaccine efficacy (VE) of Gilbert and Hudgens (2008). We adapt the time-dependent and surrogate-specific positive predictive value (PPV_x) of Zheng et al. (2008) for use in SoP evaluation. We broaden the concept of standardized total gain (STG) (Huang and Gilbert, 2011) and partial total gain (pTG) (Sachs, 2011) to allow for time dependence. We conform several other risk model based summary statistics from Gu and Pepe (2011) to the SoP evaluation framework.

In Chapter 6, we introduce three methods of estimating the time-dependent SoP estimands from Chapter 5. We present a novel Weibull structural risk model, extending the fully parametric esti-

mated likelihood (EML) methods of Follmann (2006) and Gilbert and Hudgens (2008) to accommodate this models estimation. We propose a semi-parametric EML method of estimation, an extension of Huang and Gilbert (2011), and finally, we propose the use of pseudoscore estimation adapted from Chatterjee et al. (2003). The pseudoscore estimation is unique as it allows for a closed form variance estimates, something not previously available with any existing SoP evaluation methods. We evaluate all three methods of estimation via simulation.

In Chapter 7, we apply the proposed SoP evaluation methods to a real data example from the ZEST and Step vaccine trials. We apply and compare the methods and the summary statistics proposed on these data. In Chapter 8, we briefly discuss the scientific interest of our proposed methods. We give conclusions based on our finding from the data analyses and outline our ideas for future work.

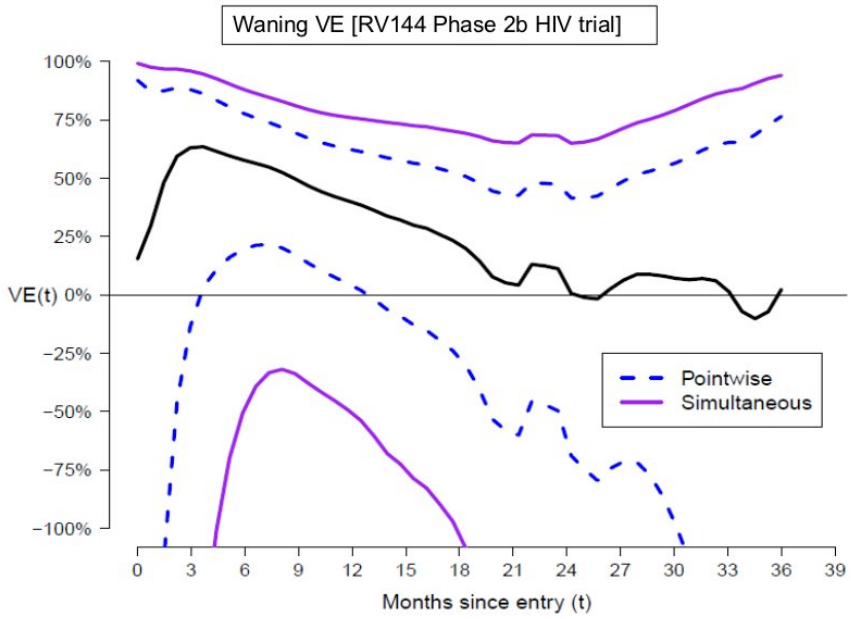

Figure 1.1: Gilbert et al. (2011b) vaccine efficacy wanes over time in the RV144 trial

1.4 Human Immunodeficiency Virus (HIV) Background

The concept of viral vaccines is not new. By priming the immune system to recognize infecting virions, it is hoped that a protective response prevents or aborts disease. An antibody or humoral protective response reacts to free floating antigen, while cell-mediated or T-cell protective response is reactive to markers found on the surface of infected cells. Cell-mediated and antibody protective responses work synergistically to prevent or clear infection (Murphy et al., 2008). Vaccines target a response from either or both of these protective pathways to prevent clinically significant infection and lasting disease (Plotkin, 2009). Although there have been HIV vaccine formulas that attempted to target both responses, there is still no effective, licensed HIV vaccine.

1.4.1 Challenges to HIV Vaccine Development

There are many unique and significant challenges to developing a safe and effective HIV vaccine. The target cells of HIV are immune cells. If a vaccine induces an active but ineffective immune response to HIV exposure, vaccination could increase the risk of infection or decrease the likelihood of viral control after infection. Guaranteeing vaccine harmlessness for healthy individuals is a primary goal of all vaccine development, but is of particular importance for HIV vaccines.

HIV mutates quickly, on average there are three nucleotide mutations per complete RNA copy. Rapid mutation leads to a genetically diverse HIV population within and between individuals. In order for a vaccine to be effective in preventing infection it must be priming the immune system against the virus of exposure. With such a diverse population of viruses, it is very difficult to develop a vaccine that will be effective against all possible exposures. Also, due to the extremely rapid mutation of HIV, live attenuated virus vaccines are currently considered inadvisable.

Vaccines often do not prevent infection, but rather prevent disease (Clements-Mann, 1998). Mutation after infection, possibly caused by immune pressure from a primed system, can make the infecting virus the common ancestor to many quasispecies that may become the dominant specie, or majority strain, in all or part of the body. Although advantageous mutations are relatively rare when compared to the number of total mutations, they allow HIV to evade recognition by the immune system. Quasispecies can remain dormant in the body in resting T-cells and may be reintroduced when the heightened targeted response has diminished. Within these resting T-cell reservoirs, the

full history of an individual's HIV quasispecies is contained. Therefore, even the primed immune response of a vaccinated subject is unlikely to control HIV in the long term.

1.4.2 HIV Vaccine Trials to Date

The RV144 trial, conducted for 2003-2009, was the first HIV vaccine efficacy trial to show significant reduction in infection (Rerks-Ngarm et al., 2009). The results of RV144 have been highly debated due to the slim margin of significance (Gilbert et al., 2011a). RV144 was a community-based, randomized, multicenter, double-blind, placebo-controlled efficacy trial. Over 16,000 eighteen to thirty year old male and female subjects from two different provinces in Thailand were randomized $1:1$ to receive four priming injections of a canarypox vector plus two booster injections of a recombinant glyco-protein vaccine. Two primary end points were monitored, HIV-1 infection and early HIV-1 viremia, every 6 months for 3.5 years per subject. A Cox regression model was used for the primary modified intention-to-treat (MITT) analysis. The MITT sample included all subjects who were not found to have HIV infection at baseline and the estimated vaccine efficacy (VE) was 31.2 percent (95% confidence interval (CI), 1.1 to 51.2; P-0.04).

The first two Phase III HIV vaccine trials, VaxGen 003 and 004, tested two different bivalent subtype recombinant glyco-protein 120 (rgp120) subunit vaccines. The 004 trial tested a bivalent subtype B/B vaccine and was conducted in the population of men who have sex with men (MSM) and high risk females (HRF) in North America (N.A.) and The Netherlands (NL). The vaccine did not prevent HIV-1 acquisition; infection rates were 6.7 percent in 3598 vaccine recipients and 7.0 percent in 1805 placebo recipients. Vaccine efficacy was estimated to be 6 percent (95% CI -17% to 24%; p- 0.59) (Flynn et al., 2005). Subgroup analyses suggested a trend toward increased acquisition risk among non-whites and women. Findings were not significant due to the small number of infections in these subgroup. VaxGen 003 tested a subtype B/E rgp120 vaccine in the population of injection drug users in Thailand. Among the 2527 subjects randomized, the estimated VE was 0% (p-0.99) (Gilbert et al., 2005; Pitisuttithum et al., 2005).

In 2004, the Step and Phambili trials began to enroll participants. Both trial vaccines were formulated to elicit a T-cell response by using a modified Ad5 virus that contained three core HIV proteins. Step was conducted at 34 sites in North America, the Caribbean, South America, and

Australia. Three-thousand participants were randomized in Step to receive three injections of the MRKAd5 HIV-1 gag/pol/nef vaccine (n=1494) or placebo (n=1506). The MRKAd5 vaccine was found to have no evidence of efficacy. In the primary analysis that assessed the subgroup of subjects with baseline Ad5 antibody titer 200 or less, 24 (3%) of 741 vaccine recipients became HIV-1 infected versus 21 (3%) of 762 placebo recipients, 95% CI for RR included 1 (Buchbinder et al., 2008a). There was some evidence that vaccination increased risk of HIV infection among uncircumcised males with Ad5 titers greater than 200. The Phambili trial was unblinded once the results of Step were made public. As there is still no licensed HIV vaccine, new trials and new vaccines formulas are needed. One such trial, HVTN 505 is the only currently active Phase II or higher trial of a HIV vaccine.

HVTN 505 is currently enrolling subjects to evaluate the safety and effect of a multiclade HIV-1 DNA plasmid vaccine followed by a multiclade HIV-1 recombinant adenoviral vector vaccine on infection and post-HIV acquisition viremia in HIV-uninfected, Ad5 neutralizing antibody negative, circumcised men (HVTN, 2009). HVTN 505 will enroll a study population of 2200 US MSM. The trial population was selected based on the Step trial results, excluding any subgroup that showed possible increase in risk of HIV infection due to the Ad5 vaccine. Subjects will be followed and tested every 3 months until 66 cases post full vaccination have been established. Each case will have 196 days of post-infection follow-up prior to primary statistical analysis. Although it is unknown whether the HVTN 505 will show significant vaccine efficacy, correlates of protection found due to the trial will move the field forward, possibly toward finding a SoP (Mulligan, 2009). Table 1.1 outlines the Phase II and 3 HIV vaccine trials that have concluded or are currently planned.

1.4.3 HIV Vaccine Trial Features

All clinical trials have unique features that need to be addressed in any statistical analysis of their data. Vaccine trials have several unique challenges to SoP evaluation, some that pose problems to be solved by statistical methods and some that support assumptions that make statistical evaluation easier. Although all the features below are present in HIV vaccine trials, they are not unique to HIV and are often present in other vaccine trials.

Table 1.1: Human immunodeficiency virus (HIV) vaccine efficacy trials

Trial	Phase	Time Period	Vaccine Type	Study Population	Primary Endpoint(s)	Trial Design	Number Events	Outcome (95% CI)
VaxGen 004	3	1998-2003	antibody rgp120 protein	N.A. and NL MSM HRF	Infection	5403, 2:1 V:P	368, 241:127	$\widehat{VE}$ =6% (-17% to 24%); p.59
VaxGen 003	3	1999-2003	rgp120 protein	Thailand IDU	Infection	2527, 1:1 V:P	211, 106:105	$\widehat{VE}$ =0%
Step HVTN 502	2b	2004-2008	Ad5 vector	Americas MSM, women	Infection; Viral load	3000, 1:1 V:P	83, 49:33	$\widehat{VE}$=(-48)% (-140% to 5%); p.07
Phambili HVTN 503	2b	2004-2008	Ad5 vector (Step)	South African Hetero	Infection; Viral load	3000, 1:1 V:P	11	Unblinded
HVTN 505	2b	2010-2015	DNA and Ad5 vector	US MSM	Infection; Viral load	1350, 1:1 V:P	45 Targeted	In progress
RV144	2b	2004-2009	Prime/Boast: ALVAC gp120	Thailand General Population	Infection; Viral load	16.395, 1:1 V:P	125, 51:74	$\widehat{VE}$=31.2% (1.7% to 51.8%); p.04

Post-treatment Biomarker Measurement

A useful surrogate predicts the effect of treatment on outcome making all potential surrogates post-randomization measurements. Estimates assessing surrogate quality ignoring this fact can suffer from bias, as unmeasured confounding can affect the observed values of the surrogate. This concern was one of the reasons for the development of the principal stratification (PS) causal framework for surrogate evaluation (Frangakis and Rubin, 2002).

Immune responses to treatment take time to develop, often a six-week to six-month post-treatment period is given prior to immune assay measurement. Pre-infection assay measurements cannot be obtained from infected subjects; treatment can affect the probability of being able to obtain a pre-infection biomarker by affecting outcome prior to assay measurement. Most methods of potential surrogate evaluation, including ours, assume that the treatment has no individual effect on outcome prior to potential surrogate measurement. This untestable assumption that is sometimes dubious has been relaxed in some work (Wolfson, 2009), this relaxation makes identifiability more difficult and we believe it is unnecessary in our motivating examples.

Constant Biomarker In Placeboees

In vaccine trials where the uninfected placebo population is unlikely to have previous exposure to the infecting agent, placebo recipients will often have negative immune response assay measurements. This will generally be true of HIV vaccine trials, as unexposed uninfected placebo recipients will not have measurable immune responses to HIV agents. The assumption of this feature is referred to in the PS surrogate evaluation literature as constant biomarker (CB). Under CB, S has the same level for all subjects in a given arm of the trial. The assumption of CB in the placebo arm is made in much of the PS literature and is well supported by past HIV vaccine trials.

Low Infection Rate

Even in high risk populations, HIV infection rates in previous trials have not exceeded 10% for 2 to 3 years of follow-up. With such low rates of infection, it is often not feasible to wait for large numbers of infection events in an HIV vaccine trial. As in all trials, efficient statistical methods are desirable to have sufficient power to detect surrogate quality at the lowest cost. This motivates the use of the

parametric method presented in Chapter 4, as well as two-phase sampling methods for increased power and reduced measurement cost. Maximizing efficiency through sequential monitoring is also a motivating factor in the newly proposed trial design of Gilbert et al. (2011b).

1.5 Varicella Zoster Virus (VZV) Background

Chickenpox are the symptoms of varicella Zoster virus (VZV) resulting from primary infection. One episode of chickenpox results in lifelong immunity to the disease, and second episodes are rare, even among immunocompromised patients (Oxman, 2010). After primary infection the virus remains dormant in the spinal dorsal root ganglia or cranial sensory nerves (RW et al., 2008). VZV can become reactivated. Although it is not completely understood why the virus reactivates, it is known that Herpes-zoster, shingles, occurs when VZV-specific cell-mediated immunity (CMI) declines and the body is unable to control the reactivated virus.

All individuals infected with VZV are at risk for shingles, but those with weakened immune systems due to age or disease are at greater risk. This is why shingles is commonly associated with the elderly. The VZV seroprevalence rate is 95 to 100% in adults aged >= 30 years, in most of the world (Araujo et al., 2007). The reported lifetime risk of developing Herpes-Zoster is 25 to 35%, but the risk is higher in the elderly and the immunocompromised (van Hoek et al., 2009). Adults fifty years of age or older account for approximately half the cases of Herpes-zoster and with increasingly elderly populations in many developed nations this is expected to increase (Johnson and Rice, 2007).

Shingles symptoms include a unilateral dermatomal rash that is often intensely painful or burning (RW et al., 2008). Neuralgic pain may persist after the rash has cleared, a condition referred to as postherpetic neuralgia (RW et al., 2008). Other, more life threatening, complications can occur in some rare cases including: secondary bacterial infections, Guillain-Barr syndrome, motor paresis, cerebral angulitis, visceral dissemination and ophthalmic damage (RW et al., 2008).

There are some treatments for Herpes-Zoster outbreak, but early treatment is important, as the effectiveness of the treatments is uncertain when they are started >72 hours after rash outbreak (RW et al., 2008; Schmader, 2001). Early treatment, although efficacious in treating the primary rash, have not been shown to reduce pain or prolonged postherpetic neuralgia. Postherpetic neuralgia is

difficult to treat (Schmader, 2001). The difficulties and cost in treating Herpes-Zoster and the severe pain associated with it are a strong argument for the development of a vaccine to prevent symptoms.

1.5.1 Development of a Vaccine Against Herpes-Zoster

In 1965, Dr. Hope-Simpson hypothesized that immunity to VZV, induced by Varicella, prevents the development of Herpes-Zoster on the basis of his observation from his general practice (Hope-Simpson, 1965). He further hypothesized that this immunity wanes over time, but that exogenous exposure to Varicella boost this immunity in healthy adults. VZV immunity may fall below some critical level, permitting latent VZV reactivation to proliferate into disease. Dr. Hope-Simpson noted that second episodes of Herpes-Zoster were relatively rare and concluded that virus replication during Shingles boosted immunity, reducing risk of a second episode. Using this hypothesis and an advanced understanding of immunology, a Zoster vaccine was developed that would increase the cell-mediated immune response (Oxman et al., 2005).

VZV is an alphaherpesvirus (Roizman and Pellett, 2007). A strain of VZV called OKa was isolated from a healthy Japanese child with varicella and attenuated by serial passage in cell culture. Due to the attenuation it was believed that injection of the strain would result in no increased risk of chickenpox and increased CMI immunity to VZV. In Phase I and II trials Oka vaccine was found to increase VZV-specific cell-mediated immunity (VZV-CMI) (Oxman, 2010). This same strain of the virus was later used at a higher concentration as a vaccine against Shingles outbreak, (Table 1 in Oxman (2010)). The minimum potency of Zoster vaccine is at least 14 times the minimum potency of varicella vaccine (Oxman et al., 2005).

The Shingles prevention study was the first major study of the varicella vaccine for the prevention of Zoster outbreak. It was a placebo-controlled, double-blind trial in which 38,546 adults aged $\geq$ 60 years of age were randomized to receive either the Zostavax vaccine or placebo with a primary endpoint of burden of illness due to Herpes-Zoster. The results suggested a significant decrease in both burden of illness, with VE estimated to be greater than 60%, although efficacy declined with age (Oxman et al., 2005).

A second Phase III trial was conducted in subjects $\geq$ 50 years of age, with estimated VE of 70% in this more immunocompetent group (Schmader et al., 2010). This is the trial that our Zoster data

example comes from. The Zostavax vaccine was approved by the FDA for use in patients 60 years of age or older in 2006 and expanded to include patients 50-59 years of age in 2011. Although approved for use, there is room for improvement as the desired VE for a licensed vaccine is higher than any of the estimated VE in any of the trials. For this reason finding and validating a surrogate for use in this improvement process is important.

1.5.2 Zoster Vaccine Trial Features

The trial and data features of the Zoster vaccine differ from those of HIV, not the least of which is that there is an effective vaccine licensed by the FDA, Zostavax. Also, as everyone at risk of shingles has the latent virus in their system, CB will not hold in the Zoster setting. The vaccine is also a live attenuated virus, something that has not yet been investigated for HIV.

However, all potential surrogates of VE are measurements of immune response to vaccine, and will therefore be post-vaccination. As such, these measurements cannot be treated as baseline variables without biasing the results. Although the life-time risk of shingles is relatively high in comparison to HIV, during the period of an efficacy trial this will still lead to a low incidence rate. Both of these features will need to be considered when evaluating candidate surrogates.

Chapter 2

CORRELATES OF RISK (COR)

Gilbert et al. (2008) and Qin et al. (2007) categorized surrogate value into three levels. First, they define a correlate of risk (CoR) to be a biomarker that is correlated with the clinical outcome. CoRs are generally evaluated in the vaccine arm alone. Secondly, they define specific surrogate of protection (SoP). The concept of SoP is a refinement of the Frangakis and Rubin (2002) definition of a principal surrogate for the vaccine trial setting. A specific surrogate of protection is a CoR that is also correlated with VE. Finally, a general surrogate of protection (GSoP) is defined as a SoP that can serve to bridge into a new trial setting. Only the SoP is defined within the principal stratification paradigm, which is outlined in Chapter 3.

Table 2.1: Definition of terms

CoR	Vaccine-induced immune response associated with the rate of HIV-1 infection among vaccine recipients. Can be identified in standard vaccine efficacy trials.
Surrogate	Correlate that reliably predicts the level of protection from HIV-1 infection. Its evaluation requires subject predictors of the correlate and/or vaccination of placebo recipients. Provides more rigorous guidance on trial-endpoint decisions for vaccine development.
Level 1	SoP: predictive value for protective efficacy in the same or similar vaccine efficacy trials.
Level 2	GSoP: predictive value for protective efficacy in different vaccine efficacy trial settings (bridging)

Definitions from Qin et al. (2007)

As CoR are only evaluated among the treated arm, or the vaccine recipients, in a trial, evaluation can be as simple as a linear regression, logistic regression or Cox model all modified to account for the two-phase sampling and depending on the clinical outcome of interest. Correlates are of interest both because they can give insight into biological mechanism and because they are potential

candidates for SoP. The term correlates of protection (CoP) encompasses both CoR and SoP. The first Step in a CoP analysis is to identify a CoR in the treatment arm that can then be considered as a candidate SoP for VE in the full trial.

2.1 Two-phase Sampling

Often the potential CoR are expensive assays that are not feasible to measure on all vaccinated subjects in a trial and therefore a sub-sample of the vaccinated group are selected for testing. Less expensive baseline covariates are often measured on all the vaccine recipients. These covariates can be used to stratify the vaccine group into risk sets for the selection of the sample to receive the more expensive or difficult assay testing. This is often referred to as two-phase stratified sampling design, most matched case-control trials use two-phase sampling.

Two-phase stratified sampling designs were proposed by Neyman (1938) when a variable of interest was difficult to measure. Under this sampling methodology a large sample of less expensive or easier to measure variables are collected; based on categories determined by these variables, a second phase sample is selected for which the more expensive or difficult to measure variables are also obtained. White (1982) developed methods that allowed for the use of both disease status and exposure for stratification of the second phase of sampling. White (1982) found that when both disease and exposure are rare, large efficiency gains could be made by including exposure in the stratification of case-control status. White (1982) outlines how complete sampling in the smaller stratum, such as exposed cases and sub-sampling in the larger stratum, such as unexposed controls, increase efficiency. Borgan et al. (2000) considered an exposure stratified versions of the case-cohort study, where true exposure is only measured at the second phase and stratification is actually based on a predictor of exposure measured on all subjects at phase one. Borgan et al. (2000) proposes a few methods of accounting for such stratification in estimation.

There are at least five schools of thought on two-phase sampling estimation (Chatterjee et al., 2003). One strategy is the use of inverse probability weights in the second-phase sample to account for the first-phase data based on the sampling probability; Horvitz and Thompson (1952) suggested an inverse probability weighted (IPW) version of estimation to account for the bias sampling. Versions of this IPW have been used in papers addressing two-phase sampling (Borgan et al., 2000;

Barlow, 1994; Barlow et al., 1999). This approach is robust to model misspecification but can be highly inefficient due to underutilization of cohort information (Robins et al., 1994).

Breslow and Wellner (2007) developed a two-phase weighting scheme for semi-parametric models and Breslow et al. (2009) suggest an improved form of the Horvitz-Thompson estimator for the Cox model. This method however, is limited by its proportional hazards assumption, which Breslow et al. (2009) highlights as future work. In yet unpublished work, Dr. Youyi Fong of the Fred Hutchinson Cancer Research Center has developed an extension to Breslow et al. (2009) to general estimating equations. This work is based on linear transformation models and therefore may be able to characterize time-varying effects, a possible extension to our proposed methods, we intend to investigate in future work. All of theses methods use weights and therefore cannot be used to account for restricted two-phase sampling, under which there is zero probability of some group of subjects being part of the second phase sample.

The second strategy for estimation with two-phase data is to consider the complete data likelihood for the second phase data using conditional logistic regression, conditioning on being in the second phase sample. Breslow and Cain (1988) developed methods of this nature for two-phase case-control sampling and Carroll et al. (1995) applied this method for survey and measurement error problems. These methods involve conditioning on being in phase-two sample.

A third strategy explores the fact that when the distribution of the second-phase covariate given the first-Phase is known, the likelihood contribution for those missing the second-phase covariate is the expected value of their complete data likelihood with respect to that distribution. One can estimate these likelihood contributions by estimating the distribution of the second-phase covariate given the first-phase. This method was first developed by Pepe and Fleming (1991) to account for a covariate measured with error in the majority of a sample but measured accurately in a validation sample. Pepe and Fleming (1991) suggest a non-parametric estimate of the distribution of the missing or mis-measured covariate from the validation sample. However, this method has also been applied using a parametric or semi-parametric estimation of the missing variable distribution. Although this method does not rely on IPW, the asymptotics developed in Pepe and Fleming (1991) require that there be a non-zero probability of any subject being in the validation sample.

A fourth approach, uses the second-phase data but the complete cohort data likelihood. Breslow and Holubkov (1997) suggest a method of this nature for case-control data and binary clinical out-

come using constrained likelihood, where it is constrained by the marginal probabilities of being a case. Scheike and Martinussen (2004) propose this same concept for the Cox model and propose the EM algorithm to solve the full data likelihood using only the incomplete or second-phase data.

The fifth approach, called pseudoscore, was developed by Chatterjee et al. (2003). They adapted the estimated likelihood method by proposing a parametric estimation of the distribution of the missing second-phase covariate that accounts for the bias sampling. They develop the pseudoscore solution to the two-phase sampling problem for a discrete baseline and a discrete or continuous outcome, although the proofs of the asymptotic properties of the pseudoscore estimator are stated for a discrete clinical outcome. Chatterjee and Chen (2007) extends Chatterjee et al. (2003) to allow for continuous baseline covariates via smoothing. The pseudoscore method is unique in that it does not require that there be a positive probability of being in the phase two sample for all subjects, merely that there is a positive expected probability of being in the phase two sample conditional on the baseline covariate(s) and the outcome.

Although there are two-phase sampling methods that allow for estimation of time-varying effects, few have been developed for the investigation of potential CoP, of which CoR identification is the 1st step. Li et al. (2007) allows for a time-dependent CoR with estimation in a two-phase sample for discrete time. Borgan et al. (2000) via extension of Prentice (1986) is a two-phase sampling method that allows for time-dependent covariates via a Cox model. Both of these methods however can not be used for the full CoP analysis, the next step of which is SoP evaluation.

In this Chapter we focus on the pseudoscore method of Chatterjee et al. (2003) with application to a Weibull parametric model that allows for time-varying associations. This method is developed below after introduction to the Weibull model of interest. The pseudoscore method is appealing for CoP evaluation due to its relaxed assumption of positive expected probability of being in the validation sample. This is useful in the evaluation of SoP, as there is always zero probability that infected placebo recipients will be in the validation sample. The pseudoscore method of estimation can be used in the CoR and SoP analysis and the asymptotic properties will apply in both.

2.2 CoR Evaluation Allowing for Time-varying Effects

2.2.1 Weibull Model

The Weibull model allows for time-varying effects to be characterized. Let X be a potential CoR measured at a fixed time-point τ. Let T be the time from potential CoR measurement τ to event and C be censoring time from τ with Y indicating that $T < C$. Let $Q = min(T, C)$. If $T < \tau$ or $C < \tau$, X is considered undefined and we remove that subject from the analysis. Thus, the cohort for analysis is subjects at risk at time τ. Let W represent all other baseline variables measured on all cohort subjects in the trial. One can also consider a partition of the baseline variables into those that were involved in second-phase sample stratification and those that are correlated with X. For simplicity of notation, we use W to denote any set of the baseline variables. Let δ be the indicator that X is measured.

Using a Weibull parametric model with both the scale and shape parameters characterized by inclusion of the potential CoR allows one to investigate a varying effect on the correlation between the fixed time immune response and the clinical outcome. We define this Weibull model by its conditional hazard function:

$$\begin{aligned}\lambda(t|x,w;\gamma,\beta) \equiv \lambda(t|x,w) &= \frac{\exp(\beta_0+\beta_1 x)}{\exp(\gamma_0+\gamma_1 x+\gamma_2 w)} \\ &\times \left(\frac{t}{\exp(\gamma_0+\gamma_1 x+\gamma_2 w)}\right)^{(\exp(\beta_0+\beta_1 x)-1)}\end{aligned}$$

and conditional survival function;

$$S(t|x,w;\gamma,\beta) \equiv S(t|x,w) = \exp\left\{-\left(\frac{t}{\exp(\gamma_0+\gamma_1 x+\gamma_2 w)}\right)^{\exp(\beta_0+\beta_1 x)}\right\}.$$

From these we can define the model for T, the conditional density allowing for right censoring,

$$g(t|x,w,y;\gamma,\beta) = g(t|x,w,y) = \lambda(t|x,w)^y \times S(t|x,w)\cdot \tag{2.1}$$

Although this is our model for the clinical outcome, the coefficients estimated via this model can be used to estimate several different characterizations of risk including the hazard, the cumulative hazard and the cumulative risk or cumulative distribution function (CDF). We define risk via the hazard by:

$$risk(t|x) = \lambda(t|x,w),$$

but any one of these could be used to define risk and can be used to evaluate the correlation of X with risk. We drop w from the risk notation here, but $risk(t|x)$ can always be considered to include baseline covariate if desired. An important consequence of using the Weibull model is that $risk(t|x)$ will always be monotonic in t at any fixed level of X. We believe this to be an advantage of the model for vaccine trials where monotonicity in time is expected. For CoR evaluation this monotonicity should be supported by the data, as most assay measurements of immune response to treatment are taken at what is believed to be the highest level of response and correlation with outcome. As we observe in the RV144 example, not all changes $risk(t|x)$ over time need be the in the same direction over the levels of x, as risk is not monotone in the potential CoR. Tests of the strength of the CoR are of interest using this model, but one must consider the effect time-dependence has on that correlation.

Before we consider evaluation of a CoR, we prove that the model we are suggesting has a unique solution.

Theorem 1. Global identifiability of the Weibull model 2.1
There are no two distinct sets of parameter values $\{\gamma, \beta\}$ such that the distribution of T is the same as defined by the probability density function PDF $g(t|x, w, y; \beta, \gamma)$.

Proof. Theorem 1:
Let $W_s(\beta, \gamma)$ denote the Weibull Model 2.1 and $I_w(.)$ denote the Fisher information for the model with respect to $\{\beta, \gamma\}$. For an arbitrary set of points $\theta_0 = \{\beta = \mathbf{B}, \gamma = \mathbf{G}\}$; $\mathrm{I}_w(\theta_0)$ is non-singular: calculations done via Mathematica. Then by Theorem 5 and Corollary 6 of Dasgupta et al. (2007) $W_s(\mathbf{B}, \mathbf{G})$ is locally identifiable at θ_0. Therefore, $W_s(\beta, \gamma)$ is globally identifiable as θ_0 is arbitrary. □

2.2.2 Testing for a CoR under the Weibull Model

As CoR is a stepping stone for the evaluation of correlates of protection (CoP), we are interested in any evidence of correlation with risk. Testing if risk is correlated with the potential CoR with or without time dependence is therefore of primary interest. However, the appropriate form of the model for testing correlation can depend on the time-variation of risk. There are three null hypotheses of time independence which can be considered:

1. $\mathrm{H0}^{1}_{CoR}$: $risk(t|x) = risk(x)$ or $\beta_0 = \beta_1 = 0$: No time-variation in risk

2. $\mathrm{H0}^{1a}_{CoR}$ or $\beta_0 = 0$: No time-variation in risk that is independent of the potential CoR X

3. $\mathrm{H0}^{1b}_{CoR}$ or $\beta_1 = 0$: No time-variation in risk that is associated with the potential CoR X

The $H0^{1}_{CoR}$ null hypothesis can be evaluated via a Wald test of both shape coefficients, β_0 and β_1, being zero. Both $H0^{1a}_{CoR}$ and $H0^{1b}_{CoR}$ can be tested directly from the model coefficients via Wald test of the nulls $\beta_0 = 0$ and $\beta_1 = 0$, respectively.

If the data do not support rejection of $H0^{1}_{CoR}$, an exponential model is suggested. In the time-independent case, the null hypothesis of interest for CoR evaluation is then:

$$H0^{2}_{CoR} : risk(x) = risk$$

this can be tested via a Wald test of the null, $\gamma_1^* = 0$, where star indicates that this is a coefficient from the exponential model.

When the data support rejection of $H0^{1}_{CoR}$ the null of interest is:

$$H0^{2}_{CoR} : risk(t|x) = risk(t)$$

This can be evaluated using a Wald test of the null $risk(t|x_1) - risk(t|x_2) = 0$ v. the alternative $risk(t|x_1) - risk(t|x_2) \neq 0$ for $x_1 \neq x_2$ and a fixed time t. In this test risk can be any one of the characterization of risk given above, i.e., CDF, Hazard etc. One might also consider the Wald test of the null $1 - risk(t|x_1)/risk(t|x_2) = 0$ for a fixed t and $x_1 \neq x_2$. Lastly one can consider the joint Wald test that all terms involving X are simultaneously zero, $\{\gamma_1 = \beta_1 = 0\}$. One may also consider the use of pseudo-likelihood ratio tests (Chen and Fan, 2005), for $\mathrm{H0}^{2}_{CoR}$ and $\mathrm{H0}^{1}_{CoR}$. This is ongoing research for this method.

2.2.3 Estimation Under The Pseudoscore Approach

In general, the second phase covariate, X, will be measured on a subsample of the subjects that are stratified into case or control at the close of the study. Often baseline covariate(s), W, will be

measured on the full sample. Although Chatterjee et al. (2003) focuses on case-control sampling and the estimation of odds ratios, it is our proposal that the same method can be used to estimate hazard ratios under two-phase sampling. So that we can follow the same estimation technique as is outlined in Chatterjee et al. (2003), let us assume that W is (are) discrete. Although Chatterjee et al. (2003) also assume the outcome to be discrete, we will show the method holds in the case where X and outcome are continuous and time-to-event, respectively. The assumption of discrete baseline covariate(s), W, was extended in Chatterjee and Chen (2007) and could be directly applied here for a continuous W; however, we do not pursue this investigation. Just as in Chatterjee et al. (2003) we must assume that

- Ps1: $\int_t \phi(t, W)dt > 0$ for all ϕ in the neighborhood of the true ϕ_0,
  where $\phi(t, W) = P(\delta = 1|T = t, W = w)$, positive expected probability of selection into second phase with respect to outcome.

- Ps2: $g(t|x, w, y; \beta, \gamma) > 0$ for almost all observed data in the neighborhood of the true β_0 and γ_0. Strictly positive value given the assumption of the parametric model for outcome T.

- Ps3: $P(\delta = 1|T, X, W) = P(\delta = 1|T, W) = \phi(T, W)$, X is missing at random, (MAR).

Assumption Ps3 will hold in most clinical trials as sampling should only depend on X via the outcome, the ability to measure X, $Q > \tau$ and the observed baseline covariate(s), W. Assumption Ps1 will hold in almost all CoR analysis as there is no reason that the sampling probability will not be greater than zero for all vaccine recipients. Assumption Ps2 requires the model for outcome to be correct and that there exists positive risk at all times. Thus, in the time-to-event outcome the time being investigated cannot exceed the longest observation time at which there are subjects still at risk.

Let $F(x|W)$ be the conditional distribution of X given W. Then the likelihood of the observed data is given by:

$$L(\beta, \gamma; F) = \prod_{i \epsilon v} g(T_i|X_i, W_i, Y_i; \beta, \gamma) \prod_{j \epsilon \bar{v}} \int g(T_j|x, W_j, Y_j; \beta, \gamma) dF(x|W_j),$$

where $v = \{j : \delta = 1\}$ is the second-phase sample of trial subjects and $\overline{v} = \{j : \delta = 0\}$ is the sample of all non-second-phase trial subjects. As suggested in Pepe and Fleming (1991), outlined below, if we could directly estimate $F(x|W_j)$ from the observed data we could use this to estimate the likelihood contributions for those missing X. However, due to the biased sampling of X we can only observe $F(x|W, \delta = 1)$, which is denoted by Chatterjee et al. (2003) as $F^*(x|W)$. Let,

$$q^{\phi}(X, W; \beta, \gamma) \equiv P(\delta|X, W) = \int \phi(t, W) g(t|X, W, Y; \beta, \gamma) dt.$$

Assumptions Ps1 and Ps2 ensure that $q^{\phi}(X, W; \beta, \gamma) > 0$ almost surely. Using this, we can define $F(x|W = w)$ from the observable data by:

$$F(x|W) = \frac{P(X \leq x|W, \delta = 1)P(\delta = 1|W)}{P(\delta|X = x, W)} \equiv \frac{F^*(x|W)P(\delta = 1|W)}{P(\delta|X = x, W)}.$$

Taking the score of the observed likelihood we have:

$$\begin{aligned} S(\beta, \gamma; F^*, \phi) &= \frac{\partial log L(\beta, \gamma; F)}{\partial(\beta, \gamma)} = \sum_{i \epsilon v} S_{\beta,\gamma}(T_i|X_i, W_i, Y_i) \\ &+ \sum_{i \epsilon \overline{v}} \frac{\int S_{\beta,\gamma}(T_i|x, W_i, Y_i)) h^{\phi}(T_i, x, W_i, Y_i; \beta, \gamma) dF^*(x|W)}{\int h^{\phi}(T_i, x, W_i, Y_i; \beta, \gamma) dF^*(x|W)}; \end{aligned}$$

where

$$h^{\phi}(t|x, w, y; \beta, \gamma) = \frac{g(t|x, w, y; \beta, \gamma)}{q^{\phi}(x, w; \beta, \gamma)}.$$

We denote the piece of the pseudoscore function for those in the validation set by $S_{\beta,\gamma}(T_i|X_i, W_i, Y_i)$ and the piece of the pseudoscore function for those missing X by:

$$S_{\beta,\gamma,F^*}(T_i|W_i, Y_i) \equiv \frac{S_{\beta,\gamma}(T_i|x, W_i, Y_i)) h^{\phi}(T_i|x, W_i, Y_i; \beta, \gamma) dF^*(x|W)}{\int h^{\phi}(T_i|x, W_i, Y_i; \beta, \gamma) dF^*(x|W)}.$$

Using the empirical estimate of $F^*(x|W)$ given by,

$$F_N(x|w) = \frac{\sum_i I_{[X \leq x, W = w, \delta = 1]}}{\sum_i I_{[W = w, \delta = 1]}},$$

one can write the pseudoscore estimating equations $S_{Ps}(\beta, \gamma; F_N, \phi)$ given by:

$$\begin{aligned} S_{Ps}(\beta, \gamma; F_N, \phi) &= \sum_{i \epsilon v} S_{\beta,\gamma}(T_i|X_i, W_i, Y_i) \\ &+ \sum_{j \epsilon \overline{v}} \sum_{i \epsilon v} \frac{S_{\beta.\gamma}(T_j|X_i, W_j, Y_j) h^{\phi}(T_j|X_i, W_j, Y_j; \beta, \gamma) I_{[W_j = W_i]}}{\sum_{l \epsilon v} h^{\phi}(T_j|X_l, W_j, Y_j; \beta, \gamma) I_{[W_j = W_l]}} = 0. \end{aligned}$$

These estimating equations can be solved via Newton-Raphson algorithm, arriving at the pseudoscore estimates $\widehat{\beta}^{Ps}, \widehat{\gamma}^{Ps}$. Chatterjee et al. (2003) points out that the estimating equations can be solved using a reweighting algorithm, which Chatterjee (1999) shows to have better convergence properties than the Newton-Raphson algorithm alone. For this reason we outline the reweighting algorithm.

1. Start with an initial estimate $\{\widehat{\beta}^0, \widehat{\gamma}^0\}$ which we will use as the first current estimate $\{\widehat{\beta}^c, \widehat{\gamma}^c\}$.

2. Use the phase two sample as is and create augmented data sets for each missing x_j, $j \epsilon \overline{v}$, $\{(T_j, X_i, W_j, Y_j), i \epsilon v_{w_j}\}$ where v_{w_j} is the subsample of the validation sample with $W = W_j$.

3. Calculate an associated weight, w_{ij}, for each imputed observation $(T_j, X_i, W_j, Y_j) : j \epsilon \overline{v}, i \epsilon v_{w_j}$, where

$$w_{ij}(\hat{\beta}^c, \hat{\gamma}^c) = \frac{h^{\hat{\phi}}(T_j | X_i, W_j, Y_j; \hat{\beta}^c, \hat{\gamma}^c)}{\sum_{l \epsilon v_{w_j}} h^{\hat{\phi}}(T_j | X_l, W_j, Y_j; \hat{\beta}^c, \hat{\gamma}^c)}.$$

4. Fit the parametric model to the combined validation and augmented datasets using the associated weights for each $(T_j, X_i, W_j, Y_j) : j \epsilon \overline{v}, i \epsilon v_{w_j}$ and update the current estimates $\{\beta^c, \gamma^c\}$.

5. Repeat associated weight calculation (3) and weighted model fitting (4) until convergence.

No standard software implements a generalized linear model (glm) for our Weibull model; so, we postulate two possible ways to fit our model in Step 4. The easier conceptual fitting is via direct weighted solution to the score equations, so the reweighting algorithm amounts to iteratively reweighted score equations. This is the estimation method we choose for our example and simulations. The second possible solution is to use iteratively reweighted least squares to solve a linear-transformation model that is equivalent to the Weibull, as those given in Zeng and Lin (2007). Investigation of the use of linear-transformation models in this setting are of future research interest.

Using either model fitting method for estimation requires calculation of $h^{\phi}(t|x, w, y; \beta, \gamma)$ at each iteration, which would seem to be more computationally difficult in the time-to-event setting. We contend that immunological sub-studies based in vaccine trials with time-to-event endpoints will use case-control sampling. Sampling will not depend directly on the time-to-event, but rather on whether or not an event was observed after τ and prior to the close of the trial. Subjects who have an observed event will be classified as cases; subjects without an observed event will be classified as controls.

We have defined Y to be the indicator that an event was observed before censoring, administrative or otherwise. We can partition $h^{\phi}(t|x, w, y; \beta, \gamma)$ by levels of T defined by Y and eliminate the need for numerical integration in the calculation of $h^{\phi}(t|x, w, y; \beta, \gamma)$, as is pointed out in Chatterjee et al. (2003) for a continuous outcome. Given $\delta \perp T|Y$ we have,

$$\begin{aligned} q^{\phi}(X, W; \beta, \gamma) &\equiv \int \phi(t, W) g(t|X, W, Y; \beta, \gamma) dt \\ &= \phi(Y = 0, W)(1 - G(c|X, W, Y; \beta, \gamma)) + \phi(Y = 1, W) G(c|X, W, Y; \beta, \gamma) \\ &= \phi(0, W)(1 - G(c|X, W, Y; \beta, \gamma)) + \phi(1, W) G(c|X, W, Y; \beta, \gamma) \end{aligned}$$

where $G(c|X, W, Y; \beta, \gamma)$ is the parametric distribution function of T given X and W at the close of the trial c.

As described by Chatterjee et al. (2003), when the clinical outcome and the second-phase covariates X are discrete, the saturated model for ϕ based on (Y, W) will give the observed sampling fractions. However, all the asymptotic properties outlined in Chatterjee et al. (2003) still apply when the model for ϕ is correctly specified.

Observing the fact that the function $\theta \rightarrow \log\{g(t|x, w, y; \beta, \gamma)\}$ is continuously differentiable with respect to (t, x, w), there exists some set $\mathcal{M}$ such that $\phi(t, w) > 0$ for all $(t, w) \epsilon \mathcal{M}$ by Assumption Ps1 and $q_0^{\phi_0}(X, W; \beta, \gamma) < \infty$ for all (x, w), we can assume the existence of the pseudoscore function. The unbiasedness of the $\widehat{\beta}^{Ps}$ and $\widehat{\gamma}^{Ps}$ estimates as the solutions to the S^{Ps} equations holds without further conditions on the form of outcome or estimating equations given independent and identically distributed (iid) data with case-control sampling at the second-phase, as stated in Chatterjee (1999). Due to the complexity of the asymptotic variance and the conditions necessary for it to hold. We first fit using bootstrap estimates of variance in the simulations.

Chatterjee et al. (2003) outlines the asymptotic properties of the β^{Ps} and γ^{Ps} estimates and the conditions under which these properties hold. First, we outline the asymptotic properties following Chatterjee (1999), we then consider the conditions under which these properties hold. We need to define some notation before proceeding. Let $\theta = \{\beta, \gamma\}$ and

$$\Psi_\theta(\theta_0, F_0^*) = \frac{\partial E_0\{S_{Ps}(\beta_0, \gamma_0; F_0^*, \phi_0)/N\}}{\partial \theta},$$

where F_0^* and ϕ_0 represent the true values and E_0 the expectation with respect to the true likelihood. Let $S_{\beta_0,\gamma_0;F_0}(t|w, y) = E_0[S_{\beta_0,\gamma_0}(t|X, w, y)|t, w, y]$, and

$$a(X, W) = \frac{h^{\phi_0}(t|x, w, y; \beta_0, \gamma_0)}{\int (h^{\phi_0}(t|x, w, y; \beta_0, \gamma_0) dF_0^*(x|w)} \{D(t|x, w, y)\},$$

where $D(t|x, w, y) = S_{\beta_0,\gamma_0}(t|x, w, y) - S_{\beta_0,\gamma_0;F_0}(t|w, y)$. Combining Theorem 5.1 and 4.1 and Proposition 5.2 from Chatterjee (1999) as well as Theorem 3.3.1 from van der Vaart and Wellner (1996) we arrive at our Theorem 2 for uniqueness, asymptotic normality and consistency.

Theorem 2. Under regularity conditions 4.1-4.3 of Theorem 4.1 of Chatterjee (1999) listed below the following hold:

- a. The pseudoscore estimating equations $S_{Ps}(\beta, \gamma; F_N, \hat{\phi}) = 0$ have a unique, consistent sequence of solutions, $\{\hat{\theta}_N^{Ps}\}_{N\geq 1}$, and

- b.
$$\sqrt{N}(\hat{\theta}_N^{Ps} - \theta_0) = -\Psi_\theta^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} g^0(T_i|X_i, W_i, Y_i, \delta_i) + o_p(1);$$
where $g^0(T, X, W, Y, \delta) = \delta\{S_{0,\beta_0,\gamma_0}(t|x, w, y) + a(x, w)\} + (1 - \delta)S_{0,\beta_0,\gamma_0;F_0}(t|w, y)$ where the subscript 0 indicates that both the model and the parameters in the model are the truth, and

- c. If $Var_0(g^0(T|X, W, Y, \delta)) < \infty$, then $\sqrt{N}(\hat{\theta}_N^{Ps} - \theta_0) \rightarrow_d N(0, \Omega)$, where Ω is defined by the sandwich formula,

$$\Omega = [\Psi_\theta(\theta_0, F_0^*)]^{-1} Var^0(g_0(T, X, W, Y, \delta))[\Psi_\theta^t(\theta_0, F_0^*)]^{-1}$$

Chatterjee (1999) outlines three regularity conditions 4.1-4.3 from Theorem 4.1, that are equivalent to the conditions of Theorem 3.3.1 in van der Vaart and Wellner (1996) which we will follow. We address the conditions in the order of difficulty to prove.

Proof. **Condition 4.3 (bounded probability):**
As Chatterjee (1999) points out, bounded probability trivially follows since the infinite dimensional parameters are estimated at a $\sqrt{N}$-rate.

Condition 4.1 (stochastic equicontinuity):
Following example 3.3.10 in van der Vaart and Wellner (1996), suppose we observe a sample $(T_1, W_1)...(T_n, W_n)$ for a distribution given by the density

$$\int p_\theta(t|x)d\eta(x)d\eta(w),$$

where η is an unknown distribution. We can express the score function with respect to our particular data as:

$$\dot{\ell}(t|x, w, y; \beta_0, \gamma_0, \eta) = \frac{\int (S_{\beta,\gamma}(t|x, W, Y))g(t|x, W, Y; \beta_0, \gamma_0)dF(x|W)}{\int g(t|x, W, Y; \beta_0, \gamma_0)dF(x|W)}.$$

This is linked to the notation in van der Vaart and Wellner (1996) by, $\dot{\ell}_\theta(t|x) = S_{\beta,\gamma}(t|x, w, y)$, $p_\theta(t|x) = g(t|x, w, y; \beta_0, \gamma_0)$ and $\eta(x) = F(x|w)$. We will use the same notation as van der Vaart and Wellner (1996) for the rest of the sketched proof. Using the suggested one-dimensional submodels $m \to \hat{\eta}$, passing through $\hat{\eta}$ in the log likelihood and differentiating with respect to m, we can obtain the likelihood equations. We assume existence of a bounded, measurable function ξ, then for every sufficiently small number $|m|$, van der Vaart and Wellner (1996) define a probability measure $\hat{\eta}$ by

$$d\hat{\eta} = (1 + m(\xi - \int \xi d\hat{\eta}))d\hat{\eta}.$$

This leads to the score operators for the mixture model:

$$A_{\theta,\eta}\xi(t, w) = \mathsf{B}_{\theta,\eta}\xi(t) + \xi(w) = \frac{\int \xi(x)p_\theta(t|x)d\eta(x)}{p_\theta(t|\eta)} + \xi(w).$$

Let H be the set of all functions $\xi : Z \to [0, 1]$ with the Lipshitz norm of all ξ less than or equal to 1.

Then $\Psi_n(\theta, \eta) = (\Psi_{n_1}(\theta, \eta), \Psi_{n2}(\theta, \eta))$, where these elements are given by,

$$\Psi_{n_1}(\theta, \eta) = \mathbb{P}_n \dot{\ell}_{\theta,\eta},$$

and

$$\Psi_{n2}(\theta, \eta)\xi = \mathbb{P}_n A_{\theta,\eta}\xi - P_n A_{\theta,\eta}\xi.$$

Then, Condition 4.1 is satisfied, as is pointed out in van der Vaart and Wellner (1996), if for some $\delta > 0$,

$$\{A_{\theta,\eta}\xi : \xi \epsilon H, ||\theta - \theta_0|| + ||\eta - \eta_0|| < \delta\}$$

is P_0-Donsker and $A_{\theta,\eta}\xi \to A_0\xi$ and $\dot{\ell}_{\theta,\eta} \to \dot{\ell}_0$ point-wise uniformly in ξ as $\theta \to \theta_0$ and $\eta \to \eta_0$.

Given Theorems 2.10.6 and 2.10.24 (van der Vaart and Wellner, 1996), and as η is an empirical CDF and all other parts of the equation are scores and pdfs from smooth, identified and bounded distributions, $A_{\theta,\eta}\xi$ is P_0-Donsker. Then by the fact that our score equation is identified for the empirical estimate of η we have point-wise uniform convergence.

Condition 4.2 (differentiability): Condition 4.2 is equivalent to the differentiability and continuity of the inverse of the derivative of the map Ψ (van der Vaart and Wellner, 1996). Using the same set-up as for 4.1 and the same example from van der Vaart and Wellner (1996), we introduce a Hilbert-space adjoint $\mathsf{B}^*_{\theta,\eta}$ of the operator $\mathsf{B}_{\theta,\eta} : L_2(\eta) \to L_2(p_\theta(.|\eta))$ given by:

$$\mathsf{B}^*_{\theta,\eta} g(x) = \int g(t) p_\theta(t|x) d\mu(x).$$

As is given in van der Vaart and Wellner (1996), the range of the operator $A_{\theta,\eta}$, is conditional in the subset G of $L_2(p_\theta(.|\eta) \times \eta)$ consisting of functions of the form $(t, w) \to g_1(t) + g_2(w) + c$. As we are working with scores, we have zero-mean functions and as such, this is a unique representation (van der Vaart and Wellner, 1996). Then the adjoint of the operator $A_{\theta,\eta} : L_2(\eta) \to G$ is given by $A^*_{\theta,\eta}(g_1 + g_2 + c) = \mathsf{B}^*_{\theta,\eta} g_1 + g_2 + 2c$. Then, we have the identity $A^*_{\theta,\eta}, A_{\theta,\eta} = I + \mathsf{B}^*_{\theta,\eta}, \mathsf{B}_{\theta,\eta}$ on the set of zero-mean functions in $L_2(\eta)$.

van der Vaart and Wellner (1996) derive the derivative of Ψ at (θ_0, η_0) as being given by the map:

$$(\theta - \theta_0, \eta - \eta_0) \to \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \eta - \eta_0 \end{pmatrix}$$

where

$$\begin{aligned}
\dot{\Psi}_{11}(\theta-\theta_0) &= -I_0(\theta-\theta_0) \\
\dot{\Psi}_{12}(\eta-\eta_0) &= -\int \mathsf{B}_0^*\dot{\ell}_0 d(\eta-\eta_0), \\
\dot{\Psi}_{21}(\theta-\theta_0)\xi &= -P_0A_0\xi\dot{\ell}_0(\theta-\theta_0), \\
\dot{\Psi}_{22}(\eta-\eta_0)\xi &= -\int (I+\mathsf{B}_0^*,\mathsf{B}_0)\xi d(\eta-\eta_0)(x)\cdot
\end{aligned}$$

As we are requiring that the operator be continuously invertible on the linear span of the domain of Ψ, this is equivalent to verifying the continuous invertibility of the two operators: $\dot{\Psi}_{11}$, which holds in our case as Fisher information for θ is positive as given for the proof of Theorem 1, and $\dot{V} = \dot{\Psi}_{22} - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}$.

The matrix $\dot{V}$ has the form:

$$\dot{V}(\eta-\eta_0)\xi = -\int\left[(I+\mathsf{B}_0^*,\mathsf{B}_0)\xi - \frac{P_0A_0\xi\dot{\ell}_0}{I_0}\mathsf{B}_0^*\dot{\ell}_0\right]d(\eta-\eta_0)\cdot$$

As is pointed out in van der Vaart and Wellner (1996), this is continuously invertible if there exists a positive number ν such that:

$$\left\{(I+\mathsf{B}_0^*,\mathsf{B}_0)\xi - \frac{P_0A_0\xi\dot{\ell}_0}{I_0}\mathsf{B}_0^*\dot{\ell}_0 : \xi\epsilon H\right\} \supset \nu H\cdot$$

Let us define K as:

$$K\xi = \mathsf{B}_0^*,\mathsf{B}_0 - \frac{P_0A_0\xi\dot{\ell}_0}{I_0}\mathsf{B}_0^*\dot{\ell}_0\cdot$$

Using the theory of Fredholm operators, if K is compact and one-to-one then the operator $I+K$ from a Banach space to itself is continuously invertible. As we are assuming a smooth map $x \to p_\theta(t|x)$, van der Vaart and Wellner (1996) state that $\mathsf{B}_0^*, \mathsf{B}_0$ is compact. The second part of K is compact as it has a one-dimensional range. It can be shown that $I + K$, in our case, is one-to-one, as presented in van der Vaart and Wellner (1996) Example 3.3.10, and 4.2 holds. □

Theorem 2 holds for our model and we have a form of the variance for the pseudoscore estimates, given by,

$$\Omega = [\Psi_\theta(\theta_0, F_0^*)]^{-1} var_0 g^0(T, X, W, Y, \delta)[\Psi_\theta^t(\theta_0, F_0^*)]^{-1}\cdot$$

Chatterjee (1999) suggests an estimator for the variance, which we will use and outline here. First, $g^0(T, X, W, Y, \delta)$ is given as above by:

$$g^0(T, X, W, Y, \delta) = \delta\{S_{0,\beta_0,\gamma_0}(t|x, w, y) + a(x, w)\} + (1 - \delta)S_{0,\beta_0,\gamma_0;F_0}(t|w, y).$$

Chatterjee (1999) gives an alternative form for $a(x, w)$ for estimation, given by:

$$a(s_1, w_k) = E^0_{t|s_1,w,\delta=1}\left\{\frac{1 - \phi_0(t, w_k)}{\phi_0(t, w_k)}\left[S_{0,\beta_0,\gamma_0}(t|s_1, w_k, y) - S_{0,\beta_0,\gamma_0;F_{k,0}}(t|w_k, y)\right]\right\} I_{[w=w_k]}.$$

Here, K is the number of levels of W with w_k being the value at the kth level, $v_k = \{j : \delta = 1, w_j = w_k\}$, $\overline{v_k} = \{j : \delta = 0, w_j = w_k\}$ and $F_{k,0}$ is the kth conditional distribution of $X|W = w_k$. The alternative form of $a(x, w)$ can be estimated by:

$$\hat{a}(x, w_k) = \sum_{i\epsilon\overline{v_k}} \frac{g(T_j|x, w_k, Y_j; \hat{\beta}, \hat{\gamma})}{\sum_{i\epsilon v_k} g(T_j|X_i, w_k, Y_j; \hat{\beta}, \hat{\gamma})}\left[S_{\hat{\beta},\hat{\gamma}}(T_j|x, w_k, Y_j) - S_{\hat{\beta},\hat{\gamma};\hat{F}}(T_j|w_k, Y_j)\right] I_{[w=w_k]}.$$

Then we can estimate $Var_0 g^0(T, X, W, Y, \delta)$ by:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\delta_i\left\{S_{\hat{\beta},\hat{\gamma}}(T_i|X_i, W_i, Y_i) + \hat{a}(T_i, W_i)\right\} + (1 - \delta_i)S_{\hat{\beta},\hat{\gamma};\hat{F}}(T_i|W_i, Y_i)\right]^{\otimes 2}.$$

It follows that the suggested estimator for Ω is given by:

$$\Omega = [\Psi_\theta(\hat{\theta}, \hat{F}^*)]^{-1}\widehat{Var}(g(T, X, W, Y, \delta))[\Psi^t_\theta(\hat{\theta}, \hat{F}^*)]^{-1}.$$

In order to test for surrogate quality via the null $risk(t|x_1) - risk(t|x_2) = 0$ for $x_1 \neq x_2$, we need to determine the form of the variance of points on the risk curve and of the difference between two points. Let $risk_\theta$, $risk'_\theta$ be risk and the derivative of risk with respect to $\theta = \{\beta, \gamma\}$, evaluated at θ for particular fixed time, T, and value of the potential CoR, X. Then by the Delta method and Theorem 2, we can state that:

$$\sqrt{N}(risk_{\hat{\theta}^{Ps}_N} - risk_{\theta_0}) \rightarrow_d N(0, \Omega[risk'_{\theta_0}]^2)$$

provided $risk'_\theta$ is non-zero at θ_0. It is possible there is some characterization of the risk for which this will be zero, in those cases the second derivative Delta method will need to be used in derivation of the standard error. The Wald test of the null $risk(t|x_1) - risk(t|x_2) = 0$ has the estimated variance:

$$\Omega[risk'_{\theta_0}(t|x_1)]^2 + \Omega[risk'_{\theta_0}(t|x_2)]^2 - 2\Omega[risk'_{\theta_0}(t|x_1)risk'_{\theta_0}(t|x_2)],$$

when $risk'_{\theta_0}(t|x_1) \neq 0$.

We investigate the finite sample properties of the pseudoscore method in simulations and in an data example from RV144 HIV vaccine trial. We use the Monte Carlo standard errors found over the simulations for the power calculations and the bootstrap SE for the example at this time. The estimated variances given by Theorem 2 although estimable are complicated and their implementation is left to the journal papers that will contain this work.

2.2.4 Simulation setting and Results

For simulations we use a trial scenario introduced in Gilbert et al. (2011b) and consider a two-armed one-to-one randomized trial. We considered seven different CoR scenarios for a continuous potential CoR, X, a discrete first-phase covariate, W, with levels determined by by quantiles of the continuous, W, used to simulate the data and a time-to-event clinical outcome, T. These are later considered for SoP evaluation as well under the pseudoscore method. These scenarios include a non-correlated CoR, X, a marginally correlated CoR and a highly correlated CoR with the clinical time-event-outcome all of which have no time-dependence. We also consider two scenarios where there is waning in the correlation over time. Figures 6.1 to 6.5 in Section 6.6 depict the scenarios used in the simulations as they relate to VE.

We also consider a time-dependent scenario where there is time-variation in the hazard, independent of the CoR, for a highly correlated CoR and a marginally correlated CoR. We also investigate a time-dependent scenario where there is time-variation in the hazard, both independent and dependent of the CoR, for a highly and marginally correlated CoR. We consider case:control sampling at a 1:5 and 1:10 ratio among the 2000 vaccine recipients for each of these scenarios.

Table 2.2 suggest that the method is unbiased in finite samples when there is a highly correlated Phase I covariate. Also, in comparison to the Monte Carlo SE for the points on the VE curve, all the biases are low, within 0.5 SE of the true value. Table 2.3 illustrates that there is reasonable power to detect an association between the potential CoR and the clinical outcome based on Wald tests. There is also suggestion of good power to detect time-dependence that is both associated with CoR and unassociated with the CoR.

Table 2.2: Percent Bias: W and X correlation (0.8) pseudoscore CoR evaluation simulations

Estimate	1:5 case:control $S(1)$							1:10 case:control $S(1)$						
	Time Ind			Risk wane		Both wane		Time Ind			Risk wane		Both wane	
	No Cor	Some Cor	High Cor	Some	High	Some	High	No Cor	Some Cor	High Cor	Some	High	Some	High
$\widehat{risk}(1\|0)^{Haz}$	0.40	0.90	1.10	2.50	0.80	1.20	0.60	0.30	0.80	1.30	1.70	0.10	0.50	0.10
$\widehat{risk}(1\|3)^{Haz}$	0.80	1.70	1.60	-0.90	-0.70	-0.60	1.30	0.40	1.50	2.60	-1.20	-0.60	-0.90	1.40
$\widehat{risk}(2\|2)^{Haz}$	0.40	0.40	0.70	-0.20	0.20	-0.10	0.20	0.40	0.40	1.00	0.30	0.10	0.10	0.20
$\widehat{risk}(1\|0)^{CDF}$	0.40	0.90	1.00	2.70	0.80	1.80	0.80	0.30	0.80	1.20	2.80	0.40	1.40	0.40
$\widehat{risk}(2\|2)^{CDF}$	0.40	0.40	0.70	-1.90	-1.50	-2.10	-1.70	0.40	0.40	1.00	-1.70	-1.40	-2.00	-1.60
$\widehat{risk}(2.5\|3)^{CDF}$	0.70	1.60	1.60	0.90	0.60	1.20	2.50	0.30	1.50	2.50	0.80	0.50	1.10	2.40

Table 2.3: Percent Rejection Power: W and X correlation (0.8) pseudoscore CoR evaluation simulations Monte Carlo SE

Null	1:5 case:control $S(1)$							1:10 case:control $S(1)$						
	Time Ind			Risk wane		Both wane		Time Ind			Risk wane		Both wane	
	No Cor	Some Cor	High Cor	Some	High	Some	High	No Cor	Some Cor	High Cor	Some	High	Some	High
$H0^{1}_{CoR}$	0.05	0.05	0.06	0.75	0.88	0.97	0.96	0.06	0.04	0.06	0.76	0.87	0.97	0.96
$H0^{1a}_{CoR}$	0.04	0.04	0.06	0.07	0.07	0.52	0.34	0.06	0.04	0.04	0.09	0.07	0.53	0.36
$H0^{1b}_{CoR}$	0.05	0.05	0.06	0.54	0.69	0.30	0.50	0.05	0.06	0.07	0.56	0.70	0.30	0.54
${H0^{2}_{CoR}}^{a}$	0.05	0.84	0.99	0.60	0.97	0.94	0.98	0.05	0.86	0.99	0.62	0.99	0.94	0.99

a:Wald test based of the null $risk(1.5|x_1) = risk(1.5|x_2)$ with risk based on the CDF at time 1.5 years at τ and $x_1 = 0$ and $x_2 = 4$.

$H0^{2}_{CoR}$ based on the correct model within each scenario, power does not reflect the percentage of models that would have been under the incorrect model due to type I or II error of $H0^{1}_{CoR}$.

2.3 RV144 CoR Analysis

A case-control analysis was conducted within the RV144 vaccine trial to identify immune correlates of infection risk (Haynes et al., 2012). Six primary immune response variables were identified via pilot studies: plasma IgA envelope (Env) binding antibodies, IgA Env antibody avidity, antibody-dependent cellular cytotoxicity, neutralizing antibodies, binding antibodies to Env first and second variable regions (V1V2), and Env-specific CD4+ T cells (Haynes et al., 2012). Each response was measured using a 26 week visit blood draw, subjects were chosen for measurement in a two-phase case:control manner. The case-control group consists of samples from 41 infected vaccine recipients pre-infection and a stratified random sample of 205 uninfected vaccine recipients using blood samples from two weeks after final immunization.

Two of the six primary immune response variables were found to be significantly correlated with infection risk. V1V2 antibody levels were inversely correlated ($\widehat{RR}$=0.57 per standard deviation increase; P-value (0.02)) and plasma IgA HIV-1 Env binding antibodies were directly correlated ($\widehat{RR}$=1.54 per standard deviation increase; P-value (0.03)) with HIV infection (Haynes et al., 2012). These findings generated the hypotheses that V1V2 antibodies may have contributed to protection against HIV-1 infection, while high Env IgA antibodies may have mitigated the effects of protective antibodies.

The primary analysis used univariate and multivariate logistic regression accounting for the two-phase sampling via Breslow and Holubkov (1997) and Cox proportional hazards models accounting for the sampling design via Borgan et al. (2000). Neither of the methods allowed for time-variation in the correlation between the clinical outcome and the CoR.

2.3.1 RV144 Analysis allowing for Time-varying Associations

We use the same case-control data set as was used in the original correlates analysis, but in addition we use all MITT controls that were not censored prior to the mean observed week 26 visit time. There were a small number of cases infected prior to week 26 that we exclude from our analysis, these cases were also removed from the original correlates analysis. We also removed from the sample any subject who did not have a known infection status or who was missing time from randomization to event or censoring. Controls with known time from randomization to event and

infection status but missing time to week 26 visit were removed from the sample if their failure time was less than the mean observed week 26 visit time.

Our phase-I sample consisted of 7843 subjects, 41 cases and 205 controls were also in the phase-II sample. The phase-II sample was selected based on outcome status and stratified by the phase-I covariates sex, per-protocol status and number of vaccinations; leading to 5 categories. Weights were assigned to the second-phase data based on these strata; we used the strata indicators as our categorical phase-I predictor of X. We considered both of the primary immune response variables found to be significantly correlated with infection risk, V1V2 and IgA HIV-1 Env, as potential CoR in our analysis.

We fit the saturated Weibull model and conducted a Wald test for the shape parameter being different from one using bootstrap SE based on 500 bootstrap samples, as the estimated SE have not yet being implemented. We find no evidence of time-dependent risk in the case of V1V2; (Bootstrap based joint-Wald test of the null $\beta_0 = \beta_1 = 0$, (P-value:0.888)). We do find marginal evidence of time-dependence of risk for IgA; (Bootstrap based joint-Wald P-value (0.047)). Tests of null hypotheses $H0^{1a}_{CoR}$ and $H0^{1b}_{CoR}$ suggest that the time dependence is driven by the time-dependence in the association with IgA, (P-value: 0.007) rather than by a main effect of time. For this reason, we consider both the Weibull and exponential models for IgA and only the exponential model for V1V2.

Just as in the original analysis we find that V1V2 is inversely associated with risk, this can be seen in Figure 2.1. We find marginal evidence in these data to support the rejection of null hypothesis $H0_2$ via the risk ratio (RR) based on the CDF at 2.83 years, the mean follow-up time in the trial, P-value (0.07). As is suggested above there is more than one way to test for an association between a potential CoR and outcome, including the RR the HR and direct testing of the potential CoR parameters being equal to zero. In this case, the Wald test based on the null $\gamma_1^* = 0$ v. $\neq 0$ or on the HR provide no additional evidence to suggest an association, P-values greater than (0.14).

We find evidence to support the rejection of the null hypothesis $H0_2$ in the case of the IgA based on the test statistics from the time-dependent model looking at the joint Wald for both IgA parameters, $\beta_1 = \gamma_1 = 0$, being zero, P-value (0.01). Based on the exponential model, we find marginal evidence, P-value (0.08), to reject null hypothesis $H0_2$, using risk ratio based on the CDF at 2.83 years. We find no evidence to support rejection of $H0_2$ based on $\gamma_1^* = 0$ v. $\neq 0$.

Looking at Figure 2.1, the left most panel, we can see that the u-shaped association with risk is reduced over time and in fact it seems like there is a growing inverse association. When we instead look at the time-independent version of the ratio of hazards over IgA, middle panel, we find a very clear direct association. The right most panel depicts the estimated ratio of hazards over V1V2 and the inverse relationship with risk is clear.

There are some weaknesses to this analysis, and prior to journal publication of this example further analyses will need to be preformed to probably support these findings. We preformed the bootstrap sampling here based solely on the second phase data, which may underestimates the true SE. As we intend to implement the estimated standard errors prior to journal publication of this work we did not feel greater investigation into correct bootstrap procedure was useful here. We also did not include any of the possibly confounders or precisions variables included in the original analysis. This maybe have effected our ability to provide strong evidence of an association between IgA and outcome and V1V2 and outcome, which is inconsistent with the primary CoR analysis of these data (Haynes et al., 2012). Investigation of the differences between these two analyses will be undertaken in greater detail once the estimated variance equations are coded.

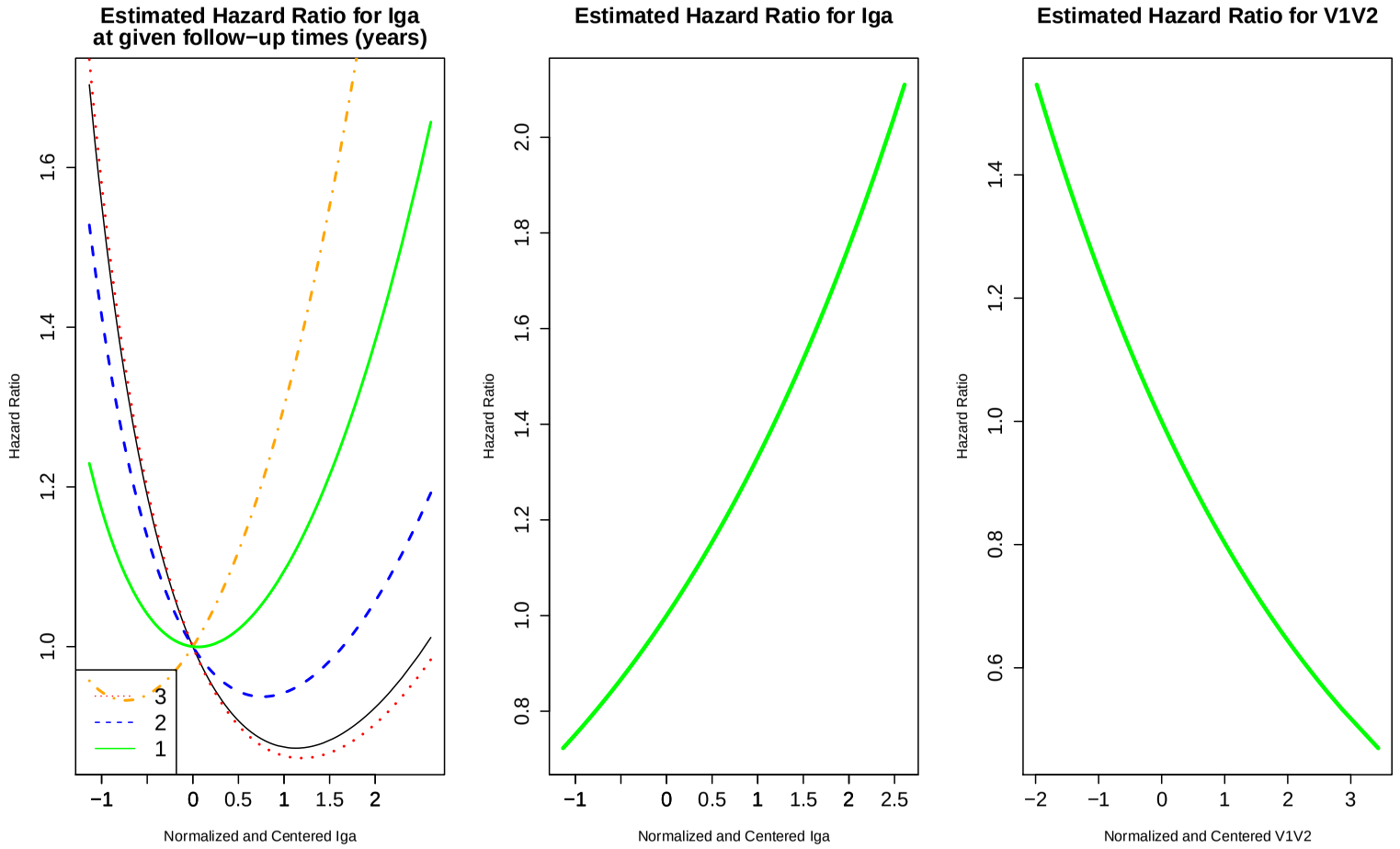

Figure 2.1: The left most panel depicts the estimated time-dependent hazard ratio at the given level of IgA relative to the estimated time-dependent hazard ratio at the mean IgA, (zero). The middle panel depicts the estimated time-independent hazard ratio at the given level of IgA relative to the estimated hazard ratio at the mean IgA, (zero). The right most panel depicts the estimated hazard ratio at the given level of V1V2 relative to the estimated hazard ratio at the mean V1V2, (zero).

Chapter 3

PARADIGMS FOR DEFINING AND EVALUATING SURROGATES

The paradigms discussed below define a surrogate, or treatment effects on the surrogate, as a predictor of outcome in some form. Joffe and Greene (2009) outline two types of prediction approaches and within those approaches four paradigms for defining and evaluating a surrogate. Under the first approach, causal-association (CA), the effect of treatment on the surrogate is associated with the effect of treatment on the clinical outcome. The second approach, causal-effect (CE), uses knowledge about the effects of the treatment on the surrogate, and the surrogate on the outcome, to predict the effect of treatment on the outcome. Below, we outline these fours paradigms for defining and evaluating surrogates, plus an additional paradigm that was recently proposed by Pearl (2011).

The first and third surrogate evaluation paradigm we discuss below, Prentice (Prentice, 1989), and direct and indirect effects (Taylor et al., 2005a; Robins and Greenland, 1992b) fall under the CE approach. The second paradigm of principal stratification is the framework under which we define our novel parametric and semi-parametric methods of SoP evaluation, and is a CA approach. The fourth paradigm, also a CA approach, uses meta-analysis to directly assess the surrogate quality in different study settings. The fifth, and most recently developed, paradigm and definition of a surrogate from Pearl and Bareinboim (2011) is both a CE and a CA paradigm. Pearl's method looks for correlation within a single trial setting and then determines the assumptions under which the surrogate would be transportable to new trial settings.

3.1 Prentice

Prentice (1989) defines a surrogate endpoint "to be a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." In a randomized clinical trial for the treatment Z, let $S(t)$ be the history prior to t of a stochastic process or fixed quantitative measure at given time points $u_j < t$. This can be used to formulate a potential surrogate for the time-to-event

clinical outcome T. Let $F(t)$ be the distribution of failure and censoring histories prior to time t for the clinical endpoint, T, and let $\lambda(t)$ be the hazard function.

Prentice's definition can be stated mathematically for a surrogate $S(t)$ as:

$$P(S(t)|z; F(t)) = P(S(t)|F(t)) \quad \leftrightarrow \quad \lambda_T(t|S(t), z) = \lambda_T(t|S(t)) \quad \forall t\cdot \tag{3.1}$$

Under Prentice's definition, the set of baseline covariates is not considered, but they could be added without changing his meaning.

3.1.1 Prentice Criteria

In Prentice (1989), two criteria and a restriction are given for identifying a surrogate or vector of surrogates from a single trial. Testing of these criteria under the restriction is necessary and sufficient for establishing a biomarker as a Prentice surrogate by equation 3.1. The criteria are:

1. $\lambda_T(t|S(t), z) = \lambda_T(t|S(t))$; T is independent of Z given S(t), and

2. $\lambda_T(t|S(t)) \neq \lambda_T(t)$; S(t) has some predictive value for T.

The restriction removes the possibility that treatment Z is independent of both the surrogate and the outcome on some space. Restricting to spaces where $E[\lambda_T(t|S(t))|z, F(t))] \neq E[\lambda_T(t|S(t))|F(t)]$, we are confining testing to a space where treatment effects on $S(t)$ have an effect on the risk. It eliminates the space where $\int_s \lambda_T(t|S(t) = s)dP(S(t) = s|Z, F(t)) = \lambda_T(t)$ even though $\lambda_T(t|Z) \neq \lambda_T(t)$. In this space we verify that criteria 1 and 2 above establish equation 3.1, for completeness we restate the proof given in Prentice (1989).

If we assume that $P(S(t)|z; F(t)) = P(S(t)|F(t))$, then

$$\begin{aligned}
\lambda_T(t|Z) &= \int_s \lambda_T(t|Z, S(t) = s)dP(S(t) = s|Z; F(t)), \quad \text{and} \\
&= \int_s \lambda_T(t|Z, S(t) = s)dP(S(t) = s|F(t)), \quad \text{by assumption, and} \\
&= \int_s \lambda_T(t|S(t) = s)dP(S(t) = s|F(t)), \quad \text{if criterion 1 holds} \\
&= \lambda_T(t)\cdot
\end{aligned}$$

If we assume that $P(S(t)|Z; F(t)) \neq P(S(t)|F(t))$, then

$$\begin{aligned}
\lambda_T(t|Z, F(t)) &= \int_s \lambda_T(t|Z, S(t) = s, F(t)) dP(S(t) = s|Z; F(t))) \text{ and} \\
&= \int_s \lambda_T(t|S(t) = s, F(t)) dP(S(t) = s|Z, F(t)) \text{ if criterion 1 holds, and} \\
&\neq \lambda_T(t|F(t)) \text{ if criterion 2 holds and the restriction is enforced}
\end{aligned}$$

by contrapositive, $\lambda_T(t|z) = \lambda_T(t) \rightarrow P(S(t)|Z; F(t)) = P(S(t)|F(t))$.

Combining, we find that given criteria 1 and 2 and enforcing the restriction, we have $\lambda_T(t|z) = \lambda_T(t)$ IFF $P(S(t)|Z; F(t)) = P(S(t)|F(t))$ or equation 3.1. While it is possible to verify 3.1 directly by running a series of clinical trials for treatment Z in which both $S(t)$ and T are measured; this construction is not discussed here. As was pointed out by Joffe and Greene (2009), Pearl (2000) and Frangakis and Rubin (2002) and as will be discussed here, Criteria 1 and 2 can yield misleading surrogate identification due to uncontrolled confounding.

3.1.2 Limitations of Prentice Surrogates

There are several works that have examples of both sound surrogates that do not meet the Prentice criteria, and biomarkers with no surrogate value that do meet the criteria (Joffe and Greene, 2009; Pearl, 2000; Frangakis and Rubin, 2002). These examples are due to unmeasured confounders between $S(t)$ and T and unmeasured conditional confounders between Z and T given $S(t)$. The example given in Figure 1 of Frangakis and Rubin (2002) illustrates how confounding can cause a sound surrogate to fail the Prentice criteria.

Consider a randomized trial with a binary clinical outcome Y and potential surrogate S. $Y = 1$ and $S = 1$ denote improved outcome and heightened surrogate response, respectively. Let there be three groups of patients in the trial: sicker patients who have poorer outcomes regardless of treatment, healthier patients who have better outcomes regardless of treatment and normal patients who respond to treatment with improved outcomes. We ignore for the moment those who might be negatively affected by treatment.

If the potential surrogate is highly predictive and follows the same pattern as the outcome: increasing in normal people who are treated and remaining the same and low in sicker people and the same and high in healthier people, the biomarker still may not be a Prentice surrogate. When com-

paring those with zero potential surrogate value over treatment, one would be comparing a group of normal placebo recipients mixed with a group of sicker placebo recipients to a group of treated sicker patients for which there was no effect of treatment on outcome or on the surrogate. The normal placebo recipients will have better outcomes than the sicker treated patients, making it seem that treatment has a negative effect on outcome while surrogate value remains the same.

This seemingly sound surrogate is not a Prentice surrogate by definition, as treatment appears to have an effect on outcome independent of S. One could see how this scenario could be reversed. If the outcome always improved with treatment but the surrogate was only affected by treatment in the normal group, the potential surrogate could be a Prentice surrogate even though it does not reliably predict outcome. When comparing patients at lower surrogate values over the treatment groups, the comparison is again between a mixture of sicker placebo recipients, normal placebo recipients and treated sicker patients. If treatment affects the outcome in the sicker patients so that the average outcome level is the same as the mixture of normal and sicker placebo recipients, the biomarker S would meet the Prentice criteria, yet would not be a useful surrogate. Principal Stratification was developed partially to address this weakness of the Prentice surrogate.

A scenario where the Prentice surrogate does not apply, addressed by Principal Stratification, is the the constant biomarker setting that is often present in vaccine trials. In order for the Prentice criteria to hold, one must observe non-constant responses in the potential surrogate measure in both arms of the trial. In fact, the Prentice method works well only in scenarios were there is similar and large variation in the potential surrogate in both arms of the trial (Wolfson and Gilbert, 2010). Principal Stratification allows for CB by conditioning on the potential measurements of the candidate surrogate under vaccination, rather than the observed value under placebo. However, Principal Stratification can also be used in settings where CB does not hold by conditioning on both the potential SoP under vaccination and under placebo (Follmann, 2006).

3.2 Principal Stratification

Principal stratification was developed in Frangakis and Rubin (2002) to evaluate post-treatment biomarkers as principal surrogates. Let Z denote the type of treatment we are testing in a randomized trial. For simplicity we assume there are two arms in the trial, ($Z \epsilon \{0, 1\}$), Z = treatment, $Z = 0$

placebo. A participant in the trial has a potential outcome, Y(z), under both the treatment arms. Subject i has at least two potential outcomes, $Y_i(1)$ and $Y_i(0)$, regardless of the observed treatment assignment of subject i. This is also true of all post-treatment variables, such as the potential surrogate S. S(z) is the potential biomarker measured under treatment arm z. The full set of potential outcomes and potential biomaker measurements for each individual i, $\{S_i(1), S_i(0), Y_i(1), Y_i(0)\}$, are assumed to be independent and identically distributed. If we were to obtain all potential outcomes and measurements for all individuals, the observed treatment assignment will contain no additional information about the treatment effects. Let τ be the follow-up time at which S(z) is measured. Then, $Y^{\tau}(z)$ is the indicator of the potential disease free status of a subject at τ. If $Y^{\tau}(z)= 0$ then S(z)=Y(z)=$*$, and are undefined.

As the observed value of the potential surrogate, S, is a post-treatment variable, the data should not be partitioned for comparison based on the value of S alone. Instead, Frangakis and Rubin (2002) introduce the concept of principal stratification with respect to a post-treatment variable without inducing bias. As stated in Frangakis and Rubin (2002), a basic principal stratification P_0 with respect to S is a partition of the data units into sets containing individuals who have the same values of the vector $(S(1), S(0))$. These are groups of people who have the same value as each other for the potential surrogate under treatment and under placebo.

A principal stratification, P, with respect to S, is the partition of the data units into sets that are unions of the sets created in P_0. An example can be drawn from subsection 3.1.2, partitioning the data in the first example into those for which $S(1) = S(0)$ and for those for which $S(1) \neq S(0)$. Comparisons in potential outcome made within particular stratum defined in this way with respect to S, S^P, are referred to as principal effects.

As stated in Frangakis and Rubin (2002) an individual's placement in a principal stratum is unaffected by observed treatment and principal effects are always causal. If we knew the S_i^p for each individual i we would fully capture all of the unit level difference measured by S, without inducing selection bias. Based on this framework, a principal effect for a post-treatment variable can be defined as a comparison of the ordered sets of $\{Y_i(1)|S_i^p\}$ and $\{Y_i(0)|S_i^p\}$.

For evaluation of a surrogate the ideal causal comparison would be between one's risk under treatment and one's risk under placebo at a given level of surrogate value under treatment and placebo, an individual-level causal effect. As all participants' potential outcomes and surrogate

measurements cannot be observed, this ideal comparison cannot be made. Thus, we reduce the comparison to the group-average level, taking the average causal comparison over the potential outcomes in each group. Frangakis and Rubin (2002) define risk based on these stratum by:

$$risk_1(s_1, s_0) \equiv \Pr(Y(1) = 1 | S(1) = s_1, S(0) = s_0) \tag{3.2}$$

$$risk_0(s_1, s_0) \equiv \Pr(Y(0) = 1 | S(1) = s_1, S(0) = s_0). \tag{3.3}$$

Following Gilbert and Hudgens (2008), one set of assumptions used to reduce the number of potential surrogates and define the average risks for comparison based on the observable data are given here:

- A1: Stable unit treatment value assumption (SUTVA) plus consistency
- A2: Ignorable treatment assignment
- A3: Equal individual clinical risk up to time τ, $\mathrm{Y}^{\tau}(1) = 0$ if and only if $\mathrm{Y}^{\tau}(0) = 0$

Assumption A1, implies that subjects' potential outcomes are independent of the treatment assignment of others in the trial, and that observation of outcomes does not change them. Assumption A1 will hold in most randomized trial settings. However, A1 may be violated for highly infectious diseases or if trial subjects are in close proximity to each other. Assumption A2, will hold in all randomized trials, contrary to previous statements in the literature blindness in not needed. Assumption A3 is useful for identifying the causal estimand defined below based on observed data from subjects at risk at time τ. Assumption A3 is an untestable assumption that can be violated in some trials. Combined, Assumptions A1, A2 and A3 imply that the risks defined above can be define in the observable data by:

$$\begin{aligned} risk_1(s_1, s_0) &\equiv \Pr(Y = 1 | Z = 1, S = s_1, S(0) = s_0) \text{ and,} \\ risk_0(s_1, s_0) &\equiv \Pr(Y = 1 | Z = 0, S(1) = s_1, S = s_0). \end{aligned}$$

Frangakis and Rubin (2002) define a principal surrogate to be a biomarker such that, $risk_1(s_1, s_0) = risk_0(s_1, s_0)$ for all $s_1 = s_0$. This condition is called average causal necessity. In the risks we are conditioning on the principal strata $\{S(1), S(0)\}$ and therefore, by definition, treatment is independent of potential outcomes. Gilbert and Hudgens (2008) point out that Frangakis and Rubin (2002)

must have implicitly conditioned on principal strata of the form $\{S(1), S(0), Y^{\tau}(1), 1, Y^{\tau}(0)\}$, and state risk as:

$$risk_1(s_1, s_0) \equiv \Pr(Y(1) = 1 | Y^{\tau}(1) = Y^{\tau}(0) = 1, S(1) = s_1, S(0) = s_0) \quad \text{and} \tag{3.4}$$

$$risk_0(s_1, s_0) \equiv \Pr(Y(0) = 1 | Y^{\tau}(1) = Y^{\tau}(0) = 1, S(1) = s_1, S(0) = s_0). \tag{3.5}$$

Frangakis and Rubin (2002) also introduce the concepts of associative and dissociative effects, suggesting that a measure of surrogate quality be based on these associations. Associative effects are differences between the sets, $\{Y_i(1) : S_i(1) \neq S_i(0)\}$ and $\{Y_i(0) : S_i(1) \neq S_i(0)\}$. Looking at this as a subset analysis, this is a comparison over treatment for subjects who would not have the same value of S under treatment as under placebo. Comparing risk in this subset allows us to determine if treatment effects on the potential surrogate imply treatment effects on the outcome. Dissociative effects are differences between the sets, $\{Y_i(1) : S_i(1) = S_i(0)\}$ and $\{Y_i(0) : S_i(1) = S_i(0)\}$. Comparing risk in this subset allows us to see if the lack of treatment effect on the potential surrogate is associated with a lack of treatment effect on the outcome. For a good surrogate, many subjects have associative effects, while few subjects have dissociative effects.

3.2.1 Principal Surrogate of Protection (SoP)

The definition of principal surrogate was refined for the vaccine efficacy setting in Qin et al. (2007), Table 2, to specific surrogate of protection (SoP). Qin et al. (2007) stated that a specific SoP has predictive value for vaccine efficacy. Using this definition there have been several proposed meanings of "predictive value" in this context. Gilbert and Hudgens (2008) suggested predictive value for efficacy can be seen as average causal necessity, just as Frangakis and Rubin (2002) required, along with average causal sufficiency, defined as $risk_1(s_1, s_0) \neq risk_0(s_1, s_0)$ for all $|s_1 - s_0| > c$ for some constant $c \geq 0$. Wolfson and Gilbert (2010) and Gilbert et al. (2011b) relaxed this definition by proposing that potential SoPs for which a comparison of structural risks over z varies greatly is a partial SoP and can be useful. However, ideal SoPs will also satisfy average causal necessity. This relaxed definition is the focus of our work.

Comparisons over the structural risks differing in z are the risk estimands of interest for specific SoP evaluation. The existing risk estimands are discussed in Chapter 4. As these estimands require observation of data not observed in standard clinical trials, they are generally unidentified.

Therefore, identification of the estimand is an important piece of all SoP evaluation methods. There have been many proposed methods of risk estimand estimation under the data restrictions of vaccine efficacy trials (Follmann, 2006; Qin et al., 2007; Gilbert and Hudgens, 2008; Gilbert et al., 2008; Huang and Gilbert, 2011; Wolfson and Gilbert, 2010; Gilbert et al., 2011b). These methods are outlined in Chapter 4.

3.3 Direct and Indirect Effects

The concept of direct and indirect effects has been investigated based on observed data and under the paradigm of causal effects and counterfactuals. Herein we focus on the direct and indirect effects (DIE) framework for surrogate evaluation that uses the causal paradigm. Under a similar line of reasoning as Prentice, but using the counterfactual framework, a perfect surrogate in the direct and indirect effects paradigm fully mediates the effect of treatment on the outcome (Joffe and Greene, 2009). This is the same as no direct effect of treatment given the surrogate, or the indirect effect of treatment through the surrogate accounts for all of the treatment effect on outcome.

Although there are several different assumption sets and trial scenarios under which direct and indirect effects can be evaluated (Robins and Greenland, 1992a), the most common is the scenario where treatment is randomized and the potential surrogate can be manipulated and randomized post treatment. The notation of the direct and indirect effects framework is similar to that of principal stratification; however, the levels of S are being controlled and are not considered an outcome. A subject's potential Y, the clinical outcome of the trial, under each arm of the trial Z, is denoted Y^z for treatment level z. As S can be manipulated and randomized, Y can also be defined under the levels of S, Y^{zs}. For each subject, Y^{zs} can differ by each possible level of Z crossed with S.

Subjects are classified for comparison by their potential outcomes under each possible level. In a two arm trial with binary outcome Y and the manipulated binary potential surrogate S there are 64 subject types (Robins and Greenland, 1992a). The group of subjects who have outcome $Y = 1$ regardless of treatment and regardless of S is an example of a subject type. One view of direct and indirect effects are the differences $E(Y^{1s}) - E(Y^{1s'})$ for an indirect effect and $E(Y^{1s}) - E(Y^{0s})$ for a direct effect. In general direct and indirect effects are not uniquely defined. Assumptions restricting the number of subject types are needed to identify and separate direct effects of treatment Z on

outcome Y and the indirect effects of Z through S on Y (Robins and Greenland, 1992a).

Taylor et al. (2005a) use the Freedman et al. (1992) proportion of the total effect explained by the indirect effects through S (PE) and introduce their modification of tests for assessing surrogate quality. Taylor et al. (2005a) state that complete mitigation is overly restrictive and merely having some PE indicates a possibly useful surrogate. Taylor et al. (2005a) attempt to link the PE statistic to the principal stratification framework counterpart proportion associated (PA) by finding some region of the Z cross S plane where PA and PE are highly correlated. Failing to do so, Taylor et al. (2005a) suggest that there are different scenarios where both paradigms are useful.

Taylor et al. (2005a) just like Robins and Greenland (1992a) consider natural direct effects (NDE) and indirect effects. Under NDE one considers potential clinical outcomes under assignment to both treatment and a particular level of the post-randomization measurement given assigned treatment. Given consistency this is the same as the observed outcome under assigned treatment and the naturally occurring level of the post-randomization measurement under the assigned treatment. PS direct effects (PSDE), like those considered by VanderWeele (2008), investigate particular levels of post-randomization variable regardless of naturally occurring levels. Pearl (2011) suggests natural direct effects are more scientifically interesting.

We, as Gilbert et al. (2011c), agree with Pearl (2011) that "...it is hard to accept the PSDE restriction that nature's pathways should depend on whether we have the technology to manipulate one variable or another". However, as pointed out in Gilbert et al. (2011c), there is a difference between those measurements that we cannot currently conceive of a way to manipulate, and those that are impossible or unethical to manipulate.

The comparison of risk when S(1)=S(0) is a PSDE, and is of clear scientific interest in SoP evaluation. To investigate this via the direct effects framework, it would require either all placebo recipients to have their post-randomization measurement set to the measurement under treatment, or for all the vaccine recipients to have their post-randomization measurement set to the measurement under placebo. In the HIV vaccination setting and post-treatment measurements of specific immune response, both of these scenarios seem highly unlikely.

For example, it seems impossible to cause a HIV naive subject to have a HIV specific immune response, without pre-exposing them or their blood to some component of HIV. This would directly violate the consistency assumption used for identification of the direct effect estimands. Although it

seems possible to cause a non-response in a vaccinee, by killing their immune system via radiation for example, it seems highly unethical. This would also eliminate other possibly pathways of protection, not just the HIV specific response we would generally be measuring as a potential SoP, and therefore might also violate the consistency assumption. For these reasons, principal stratification is a useful tool in SoP evaluation in the vaccine setting where direct effects considering the SoP PSDE of interest would be implausible to assess.

3.4 Meta-analysis

Unlike the above paradigms that evaluate and define surrogacy within a single trial setting, the meta-analysis paradigm defines and evaluates a surrogate over multiple trials. In this framework, a surrogate is biomarker the treatment effects on which allow for precise estimation of treatment effects on clinical outcome. Ideally, the surrogate allows for evaluation of a treatment in a new trial setting where clinical outcome is not measured and only the surrogate is used. This is called a bridging surrogate and is the main use of meta-analysis. Although meta-analysis could adequately evaluate an SoP it seems like a very poor use of resources, as meta-analysis is costly and time consuming.

Meta-analysis requires some set of previous similar trials, numbering n, that can estimate both the treatment effect on the potential surrogate and the treatment effect on the clinical outcome. From these n treatment effect estimates, the joint distribution of the effects over the n trials can be formed. Using this estimated joint distribution, future trials that only measure the potential surrogate can estimate the effect treatment would have on the clinical endpoint.

Not all trials in a meta-analysis need to have the same treatment, geographic location or time frames; a range of trial types may be of greater interest. The units of analysis describing the differences between the trials is of great importance for answering the scientific question of interest when considering meta-analysis. An example of units are different regions. A meta-analysis of n trials testing the same vaccine in different regions would be able to evaluate a surrogate over those regions and then extrapolate to regions outside the trial set. This extrapolation is the major untestable assumption that meta-analysis relies upon, and is one of the main criticisms of this framework. Surrogate evaluation via meta-analysis is attractive as it allows evaluation of GSoP, which, if adequate,

can allow for S to be used as the clinical endpoint in future Phase IIb and III trials that are unit extensions of the meta-analysis trial set; this is called bridging. The use of a true GSoP in a Phase IIb or III HIV vaccine trial would greatly increase vaccine testing efficiency and cost effectiveness.

3.4.1 Methods of Meta-analysis Evaluation of Surrogate Value

Daniels and Hughes (1997) introduce an approach to meta-analysis-based surrogate evaluation. Under the Daniels and Hughes (1997) approach, the ith trial estimates the true clinical treatment effect for the ith treatment, θ_i, and is denoted $\hat{\theta}_i$. The true treatment effect on the potential surrogate, γ_i is also estimated within the ith trial; denoted $\hat{\gamma}_i$. Each trial must be large enough so that the estimates can be assumed to be distributed normally about the true effects. The distribution of the estimates over the trials is then given by:

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\gamma}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_i \\ \gamma_i \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i\sigma_i\delta_i \\ \rho_i\sigma_i\delta_i & \delta_i^2 \end{pmatrix}\right),$$

where σ_i^2 and δ_i^2 are the within-trial estimated variances and ρ_i is the correlation between the estimates conditional on true treatment effect. This distribution reflects the within-trial variation of treatment effects. The between-trial variation of treatment effects can then be modeled by treating the set of true treatment effects of the potential surrogate, $\{\gamma_i\}$ as fixed effects. For a large enough number of trials, n, the true treatment effects on clinical outcome, $\{\theta_i\}$ can be assumed to be distributed normally given the true treatment effects on the potential surrogate, $\theta_i|\gamma_i \sim N(\alpha + \beta\gamma_i, \tau^2)$. The value of τ^2 reflects the between trial variation. Combining these distributions we can remove the true θ_i from the distribution.

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\gamma}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha + \beta\gamma_i \\ \gamma_i \end{pmatrix}, \begin{pmatrix} \sigma_i^2 + \tau^2 & \rho_i\sigma_i\delta_i \\ \rho_i\sigma_i\delta_i & \delta_i^2 \end{pmatrix}\right).$$

Applying priors to the unknown parameters, Daniels and Hughes (1997) estimate this distribution, which allows for prediction of treatment effect on the clinical outcome in future trials that only measure treatment effects on the surrogate.

Buyse et al. (2000) and Gail et al. (2000) extend the Daniels and Hughes (1997) method to more general modeling of the distributions of the treatment effect on the clinical outcome and the treatment effect on the potential surrogate. The methods of Gail et al. (2000) allow for the determination of the accuracy of estimated treatment effects on clinical outcomes in new trial setting, using bootstrap confidence intervals.

3.5 Pearl Paradigm

Recent papers, Pearl (2011) and Pearl and Bareinboim (2011), call into question the usefulness of the principal stratification framework and introduce a new definition of a surrogate using directed acyclic graphs (DAGs) that depict the relationship between treatment, the surrogate and the clinical outcome. Pearl maps the causal relationships using DAGs and determines what assumptions apply to each edge. The basis of Pearl's paradigm is that any variable correlated with the outcome in a single trial should be investigated as a bridging surrogate and the assumptions under which the bridging is valid should be identified and verified. A future research goal of the author is to determine the conditions under which a SoP or a GSoP is a Pearl surrogate and how these two paradigms can be used in compliment.

Chapter 4

EXISTING METHODS OF SPECIFIC SURROGATE OF PROTECTION (SOP) EVALUATION

4.1 Risk Estimands

Comparisons of structural risk over z, risk Equations 3.4 and 3.5, are the estimands of interest for specific surrogate evaluation. Estimands can be based on $risk_z(s_1, s_0)$ or on the predictiveness curve of Huang et al. (2007), which was introduced for use in specific SoP evaluation by Gilbert and Hudgens (2008). The predictiveness curve for a potential SoP, S, give that $Z = z$, is given by:

$$R_z(v_1) \equiv Pr(Y(z) = 1|S(1) = F^{-1}_{S(1)}(v)).$$

This is a marginal form of risk, that conditions on quantiles of $S(1)$ and ignores the $S(0)$ values. Under CB this is equivalent to the joint predictiveness curve.

Gilbert and Hudgens (2008) introduce the causal effect predictiveness (CEP) estimand, defining the overall causal effect of treatment on the clinical outcome as $CE \equiv h(Pr(Y(1) = 1), Pr(Y(0) = 1))$, for a contrast function $h(x, y)$, such that $h(x, y) = 0$ if and only if $x = y$. The CEP surface is a conditional causal effect estimand. Gilbert and Hudgens (2008) defines the CEP for surrogate-dependent risk as $CEP^{risk}(s_1, s_0) = h(risk_1(s_1, s_0), risk_0(s_1, s_0))$ and for the predictiveness curve $CEP^R(v_1) = h(R_1(v_1), R_0(v_1))$. The CEP surface allows us to quantify the associative and dissociative effects of a potential surrogate. When $CEP^{risk}(c, c) = 0$ for all c, this indicates causal necessity or no dissociative effects. When $|CEP^{risk}(s_1, s_0)| > 0$ for all $|s_1 - s_0| > c$, this indicates average causal sufficiency or the presence of associative effects in a region of (s_1, s_0).

The most widely used CEP^{risk} in the SoP evaluation literature is surrogate-dependent vaccine efficacy, $VE(s_1, s_0)$, defined as:

$$VE(s_1, s_0) = 1 - \frac{risk_1(s_1, s_0)}{risk_0(s_1, s_0)}. \tag{4.1}$$

A figure depicting VE(s_1, s_0) v. (s_1, s_0) allows for a visual evaluation of the surrogate over the whole surface of VE. Gilbert and Hudgens (2008) suggest that for vaccine trials a good surrogate

will tend to have $|VE(s_1, s_0)|$ increasing in $|s_1 - s_0|$. However, this is not always true as the largest $|VE(s_1, s_0)|$ may be associated with a lower $|s_1 - s_0|$. A better indicator of surrogate quality is the $VE(s_1, s_0)$ figure. In one of the motivating scenarios of HIV vaccine efficacy, where CB is likely, the $VE(s_1, s_0)$ surface can be collapsed to the $VE(s_1)$ curve. When Assumptions A1 through A3 and CB hold, the marginal VE curve conditioning on s_1 alone is equivalent to the joint curve $VE(s_1, s_0)$, conditioning on both s_1 and s_0.

Figure 4.1 depicts different levels of surrogate quality. The solid horizontal line depicts a useless surrogate, while the two dashed lines depict a surrogate of medium quality and a highly sensitive surrogate; the surrogate starting at zero and arcing to 90% VE being the better of the two. The $VE(s_1)$ curve for the highly sensitive surrogate depicts the ideal zero average VE when surrogate value is zero and average 90% VE when surrogate value is high. This is an excellent surrogate due to the large amount of variability in VE over the range of s_1, the monotonic increase of VE in s_1 and that $VE(0) = 0$, such that average causal necessity holds.

Gilbert and Hudgens (2008) introduce a CEP based summary of surrogate quality estimand called PAE^w. Huang and Gilbert (2011) introduce standardized total gain (STG) based on the risk difference; $risk_0(s_1, s_0) - risk_1(s_1, s_0)$. There have been several summary statistics suggested in the principal surrogate evaluation literature (Taylor et al., 2005a; Li et al., 2010; Huang and Gilbert, 2011; Gilbert and Hudgens, 2008).

4.1.1 Estimands for Summarizing Surrogate Value

Summary statistics are useful in evaluation and comparison of candidate specific surrogates of protection. The proportion associative (PA) was introduced by Taylor et al. (2005a) and defined as:

$$PA \equiv Pr(S(1) > S(0), Y(1) = 0, Y(0) = 1)/Pr(Y(1) = 0, Y(0) = 1).$$

This is an interesting summary as it is the proportion of subjects with positive clinical effects of treatment that also have positive treatment effects on the surrogate. As pointed out in Gilbert and Hudgens (2008), this measure is dependent on the underlying joint strata distribution as it does not condition on $(S(1), S(0))$. Thus, if it is likely that surrogate response under treatment will be greater than under placebo, PA will be large.

Gilbert and Hudgens (2008) remedy this weakness in PA by introducing the proportion associative effect PAE^w which conditions on $\{S(1), S(0)\}$. The two components of PAE^w are the weighted expected associative effect (EAE^w) and the expected dissociative effect (EDE). EAE is weighted to reflect the utility that finding associative effects is of greater importance than penalizing candidate surrogates with dissociative effects. EAE^w is defined by:

$$\text{EAE}^w \equiv \frac{E[w(S(1), S(0))CEP^{risk}(S(1), S(0))|S(1) > S(0)]}{E[w(S(1), S(0))|S(1) > S(0)]},$$

where $w(S(1), S(0))$ is a weight function, and EDE is defined by,

$$EDE \equiv E[CEP^{risk}(S(1), S(0))|S(1) = S(0)].$$

From these PAE^w is constructed, $PAE^w \equiv |EAE^w|/\{|EDE| + |EAE^w|\}$. PAE^w has a range of $[0, 1]$ and values in $(0.5, 1]$ indicate the presence of surrogate value. Gilbert and Hudgens (2008) offer examples of weight functions that could be used; different weight functions lead to different clinical meanings of PAE^w. For example, when CEP^{risk} is the risk difference, $Pr(S(1) > S(0)) = 0.5$ and no active harm by treatment is assumed then the unweighted PAE$^w = 1$ is the PA of Taylor et al. (2005a).

Huang and Gilbert (2011) consider a different type of summary measured based on the CEP of risk difference. Huang and Gilbert (2011) denote the risk difference curve as,

$$\Delta\{S(1), S(0)\} = risk_1(s_1, s_0) - risk_0(s_1, s_0)$$

and the difference in potential outcomes $\{Y(0) - Y(1)\}$ as D. They propose to characterize surrogate value as the amount of variability in D explained by $\Delta\{S(1), S(0)\}$. Huang and Gilbert (2011) consider the quantile plot of $\Delta\{S(1), S(0)\}$, denoting the vth quantile of $\Delta\{S(1), S(0)\}$ by $R^\delta(v)$. It was shown in Gilbert and Hudgens (2008) that the area under the quantile curve is equal to the difference in prevalence, where prevalence can be denoted as $\rho_z = Pr\{Y(z) = 1\}$, and the difference in prevalence over z by $\rho_0 - \rho_1$. Huang and Gilbert (2011) suggest that formal comparison of these two curves can be undertaken via the total gain (TG) of Bura and Gastwirth (2001) and the standardized version (STG) of Huang et al. (2007). TG is the area between the quantile curve $R^\delta(v)$ versus v and the line $\rho_0 - \rho_1$ under CB. This is given by:

$$\widehat{TG}(W) = \int |\widehat{risk}_0(s_1) - \widehat{risk}_1(s_1) - \{\hat{\rho}_0(W) - \hat{\rho}_1(W)\}|d\hat{F}^{S(1)|W}(s_1).$$

The larger the TG the better the model of risk and the more useful the surrogate. Standardization of this statistic is appealing because under the additional assumption of no active harm, $Y(1) \leq Y(0)$, it has a clinically meaningful interpretation. The STG is given by:

$$\widehat{STG}(W) = \widehat{TG}(W) / \left[2\{\hat{\rho}_0(W) - \hat{\rho}_1(W)\}\{1 - \hat{\rho}_0(W) + \hat{\rho}_1(W)\}\right] \cdot$$

Given the no harm assumption, STG can then be defined as: STG= $\max_c$\{Sensitivity(c)+Specificity(c)\}$-$ 1, where $Sensitivity(c) = P(D = 1|\Delta\{S(1), S(0)\} > c)$ and $Specificity(c) = P(D = 0|\Delta\{S(1), S(0)\} < c)$. As Sensitivity and Specificity are well known classification measures, this definition of STG makes it clinically meaningful and easily interpretable.

Li et al. (2010) consider other summary statistics for the setting of binary clinical endpoints and binary potential SoP. They define causal effects similarly to Gilbert and Hudgens (2008) as the fraction of patients who respond to treatment. Using the probabilities from Table 4.4, CE can be defined as: $CE = p_{12} + p_{22} + p_{32}$. The associative effect, p_{22}, is the proportion of subjects with effects on S and Y, and the dissociative effect, $p_{12} + p_{32}$, is the proportion of subjects with effects on Y but not S. Therefore PA is given by p_{22}/CE. Li et al. (2010) introduce the surrogate associative proportion (SAP), given by $SAP = p_{22}/(p_{21} + p_{22} + p_{23})$ and the surrogate dissociative proportion (SDP), given by $SDP = (p_{12} + p_{32})/(p_{1.} + p_{3.})$, where $p_{1.} = \sum_{k=1}^{3} p_{1k}$. SAP is equivalent to positive predictive value (PPV) and SDP is equivalent negative predictive value (NPV) the setting of Li et al. (2010). SAP can also be thought of as the ratio of the associative effect and the proportion of subjects with treatment effect on S. Similarly, SDP is the ratio of the dissociative effect to the proportion of subjects with no treatment effect on outcome.

Li et al. (2010) suggest associative proportion (CAP) as a summary statistic, defined by $CAP = p_{22}/(p_{12} + p_{21} + p_{22} + p_{23} + p_{32})$. They state that when the denominator of CAP is p_{22}, S is a perfect surrogate, and CAP will be 1.

Most vaccines are tested via a large, stratified, randomized trial design. RV144 is an example of this type of trial. All of the SoP evaluation estimands suggested above condition on $S(1)$. Therefore, due to the missing S(1) values in the placebo arm, estimands are unidentified given the observed data from a standard clinical trial. Identification of the estimand is an important piece of all SoP evaluation methods. Follmann (2006) introduced trial augmentation designs to add in SoP estimand identification; these methods are outlined below.

4.2 Trial Designs for Identification of SoP Estimands

4.2.1 Augmented Vaccine Trial Designs

Follmann (2006) outlines three augmented trial designs: closeout placebo vaccination (CPV), baseline irrelevant variable (BIV), which is a special case of a baseline immunogenicity prediction (BIP) of Gilbert and Hudgens (2008), and their combination (BIP + CPV). In BIP, a baseline predictor, W, of the surrogate, S, is measured in some number of the trial participants. This baseline predictor is ideally highly correlated with S. In the special case of BIV, W is also independent of the clinical outcome given S. In CPV, uninfected placebo recipients at the close of the trial are given the vaccine and S^C, the same biomarker used for $S(1)$, is measured at the same timepoint τ after closeout vaccination.

The BIP+CPV setting is merely the combination of the two designs in which all or a subset of the uninfected placebo recipients receive closeout vaccination and all or some subset of the trial participants have a BIP measured. Follmann (2006) only discusses the case where all uninfected placebo recipients and all vaccine recipients have $S(1)/S^C$ measured for CPV or BIV+CPV and where all participants have a BIV. Follmann's augmented trial designs have been extended in multiple works to two-phase sampling of CPV and BIP (Gilbert and Hudgens, 2008; Huang and Gilbert, 2011; Qin et al., 2008). The use of the augmented trial designs requires assumptions about the nature of S^C relative to S(1). In order for CPV to be useful for identification we must make some assumptions about the CPV values. For example the following two assumptions suffice:

- A5: Time constancy of the immune response distribution, for subjects with $Y(0) = Y^c(0) = 0$ $S^C = S(1) + e_i$ were e_i are iid random errors with mean zero.

- A6: No infections in the uninfected placebo group during the close-out period, $Pr\{Y(0)^c = 0|Y(0) = 0\} = 0$.

$Y(0)^C = 0$ indicates no infection occurred during the close-out period. The close-out period is the time after the close of the trial during which placebo receive close-out vaccination and have S^C measured. We use the A5 and A6 notation here to be consistent with the literature; Assumption A4 will be introduced in a later section (Gilbert and Hudgens, 2008; Huang and Gilbert, 2011; Wolfson

and Gilbert, 2010). There are cases, such as infant trials, where there will most likely be an absolute time effect due to rapid changes in the subjects over the length of the trial, and in those cases A5 will not hold. The validity of Assumption A5 should always be considered on a case by case basis. When A5 fails, CPV does not provide the information needed to replace the missing S(1) value.

For the motivating example of Zostavax® it is debatable if there would be an absolute time effect as there are not rapid changes in the circulating strain and natural exposure has already occurred prior to randomization. However, aging two more years may change the immune response in this elderly population. If CPV had been taken in the Zostavax example we would have tested for evidence of this by looking at the immune responses in the vaccine recipients over age groups based on average length of study follow-up. In HIV vaccine trials A5 will hold in general as subjects immune responses to HIV peptides after vaccination will not tend to change over the standard 3 to 7 years of the trial.

Assumption A6 is needed to identify the relationship of interest for use of CPV. The distribution we are interested in, S(1) conditional on $\{Y(0) = 0, T(1) > \tau, T(0) > \tau, W\}$, can only be identified via CPV if we can identify $Pr\{S^C = s_1 | Y(0) = 0, T(1) > \tau, T(0) > \tau, W\}$. This is not the case if $Pr\{S^C = s_1 | Y(0) = 0, Y^c(0) = 0, T(1) > \tau, T(0) > \tau, W\} \neq Pr\{S^C = s_1 | Y(0) = 0, T(1) > \tau, T(0) > \tau, W\}$. Unlike Assumption A5, A6 can easily be tested in a trial by observing the closeout outcomes of CPV subjects. There is biological support of A6 in vaccine trials, as most placebo recipients who remain uninfected at the end of the trial will be relatively low risk subjects who are now vaccinated with a possibly efficacious vaccine. However, as subjects will generally be from high risk populations the stringent restriction of no infections seems unlikely to hold. Minor deviations from A6 are most likely allowable and provided that close-out infections do not exceed the expected rate observed in the trial, sensitivity analysis can be preformed to determine how much deviation from A6 change the results of the SoP analysis.

4.2.2 Sequential Trial Design

The newly proposed sequential Phase IIb multi-drug trial design of Gilbert et al. (2011b) addresses many of the difficulties in HIV vaccine trials. This trial design leverages the high risk of a large group of South Africans in order to test multiple vaccine candidates simultaneously; evaluating

VE and bringing new candidate vaccines into the trial as needed. The sequential trial design was created to address four main objectives, two of which are the proposed methods. Main objective two, to evaluate durability of VE can be accomplished using our proposed method that characterizes waning in VE, outlined below. Main objective three, to expeditiously evaluate the immune correlates of protection, will yield opportunities for use of our proposed methods in trials that have adequate events to allow for CoP detection.

The design uses prespecified sequential monitoring for the events of vaccine harm, non-efficacy, and high efficacy. Monitoring plans are selected to weed out poor vaccines as rapidly as possible while guarding against prematurely abandoning vaccines that require more time to confer vaccine efficacy. The trial will operate until an effective vaccine is found or no new candidate vaccines remain. Potential surrogates can be evaluated in the large multi-arm trial allowing for more power to detect surrogacy, while individual vaccine candidates can be tested against the pooled control group alone. The trial is augmented with two-phase sampling of both CPV and BIP. Simulation studies of the design showed power to detect surrogacy increases dramatically with an increase in the number of trial arms.

4.3 Estimation of SoP Estimands

4.3.1 Estimated Maximum Likelihood (EML)

In Pepe and Fleming (1991) the concept of using estimated maximum likelihood (EML) in the presence of missing or mismeasured covariates is introduced. Although the paper suggests an empirical form of estimation, it has been applied parametrically in much of the SoP evaluation literature.

If X denotes the true value of a missing covariate and W denotes a predictor of X, then let the probability function for an outcome Y as a function of X and W be parameterized by β and denoted by $P_\beta(Y|X, W)$, $P_\beta(Y|W) = \int P_\beta(Y|x, W)dP(x|W)$. If there were some group of subjects, a validation group, that had both X and W measured, and we knew $P(X|W)$, the likelihood for β given iid data would be,

$$L(\beta) = \prod_{i \in Val} P_\beta(Y_i|X_i, W_i) \prod_{i \in \overline{Val}} P_\beta(Y_i|W_i),$$

where Val and $\overline{\text{Val}}$ indicate inclusion in and exclusion from the validation group. As we do not know

P($X|W$) exactly, the observed likelihood is given by,

$$\mathrm{L}(\beta,\theta) = \prod_{i\in Val} \mathrm{P}_\beta(Y_i|X_i,W_i) \prod_{i\in\overline{Val}} \int \mathrm{P}_\beta(Y_i|x,W_i) d\mathrm{P}_\theta(X|W)\cdot$$

One could maximize over β and θ simultaneously, but this is a computationally intensive procedure. Instead, it was shown in Pepe and Fleming (1991) that if one estimates $\mathrm{P}_\theta(X|W)$ consistently and independently of the β, it will lead to consistent estimates of β from the resulting estimated likelihood. The estimated likelihood is given by:

$$\widehat{\mathrm{L}}(\beta) \equiv \mathrm{L}(\beta,\hat{\theta}) \prod_{i\in Val} \mathrm{P}_\beta(Y_i|X_i,W_i) \prod_{i\in\overline{Val}} \int \mathrm{P}_\beta(Y_i|x,W_i) d\mathrm{P}_{\hat{\theta}}(x|W)\cdot$$

One can then maximize this estimated likelihood over β. After establishing the β estimates derived from the estimated likelihood will be consistent for the true β, Pepe and Fleming (1991) establish the asymptotic normality of the estimated score functions. This property of the EML estimates will not apply in the SoP evaluation environment, however, as a minimal required condition of asymptotic normality is a non-zero probability of being in the validation set for all subjects. The zero probability of infected placebo recipients being in the validation set violates this condition.

Binary Outcome Parametric EML

As was introduced in the principal stratification paradigm Section 3.2, let Y(z) be the potential binary clinical endpoint under the treatment arm z and S(z) the potential biomarker. Let τ be the follow-up time at which S(z) is measured. Then, Y^τ(z) is the indicator of the potential disease-free status of a subject at τ. If Y^τ(z)= 0 then S(z)=Y(z)=$*$, and are undefined. The full set of potential outcomes and potential biomaker measurements for each individual i,

$$\{S_i(1), S_i(0), Y_i(1), Y_i(0), W_i, Y_i^\tau(1), Y_i^\tau(0)\}$$

are assumed to be iid. Let $F_{S(1)|W}$ be the distribution of S(1) given W and μ_S, μ_W, σ_S^2 σ_W^2 be the first and centered second moments of S(1) and W, respectively. Let ρ_{SW}^2 denote the correlation of S(1) and W. Let $i\epsilon V$ denote subjects in the vaccine arm of the trial, and $i\epsilon P(U)$, $i\epsilon P(I)$ denote uninfected and infected placebo recipients, respectively.

Follmann (2006) develops methods for estimating the usefulness of a potential SoP using EMLE with the assumption of a parametric form of the clinical endpoint Y. Assuming CB holds and in the

binary clinical endpoint setting risk can be defined as:

$$risk_z(s_1) \equiv \Pr(Y(z) = 1 | Y^\tau(1) = Y^\tau(0) = 1, S(1) = s_1)\cdot$$

Follmann (2006) assumes a parametric link function $g(\cdot)$ to connect risk to the variables S(1) and Z; $g(\cdot)$is assumed to be a Probit. This link is their assumption A4, referred to in Gilbert and Hudgens (2008) as A4-P when the assumption has a parametric form, denoted $risk_z(s_1;\beta)$. It allows for efficient estimation of the risk estimand,

$$risk_z(s_1;\beta) = \Phi(\beta_{z0} + \beta_{z1} * s_1),$$

where Φ represent the standard normal CDF. To link this risk to a likelihood, Follmann (2006) assumes Y has a binary distribution. Follmann (2006) also assumes that S(1) and W are distributed bivariate normal. Under any one of the three augmented trial designs described in subsection 4.2.1, risk can be identified. As stated above, Follmann (2006) only considers the case where all trial participants have the BIV W measured for BIV and BIV+CPV trials and where all vaccines and all uninfected placebo recipients have S(1) or S^C measured and $Y_i^\tau = 1$ for all i.

With the BIV+CPV full sampling, where all subjects have W measured and all uninfected placebo recipients receive CPV, there are 4 groups of trial participants defined by vaccination and infection: infected vaccine recipients, uninfected vaccine recipients, uninfected placebo recipients and infected placebo recipients. Under assumptions A1 through A6 and CB, the estimated likelihood can be identified from the observed data. The estimated likelihood is given by:

$$L(\boldsymbol{\beta}, \hat{\nu}) = \left[\prod_{i\epsilon V} risk_1(s_{i1};\beta)^{y_i}\{1 - risk_1(s_{i1};\beta)\}^{1-y_i}\right] \cdot \left[\prod_{i\epsilon P(u)} \{1 - risk_0(s_{i1};\beta)\}\right] \left\{\prod_{i\epsilon P(I)} risk_0^*(w_i;\beta)\right\},$$

where P(I), P(u) and V stand for the group of infected placebo recipients, uninfected placebo recipients and vaccine recipients respectively. Here $\nu = F_{S(1)|W}$ and $risk_0^*(w_i;\beta)$, as defined in Follmann (2006), is the expected value of the risk for a binary outcome Y over the distribution of a normal surrogate S given the BIV, W. Using the equality proven in the appendix of Gilbert and Hudgens (2008), the integration is reduced to a simple formula. The equality and likelihood are outlined below. The contribution to the likelihood for $i\epsilon P(I)$ is given by:

$$risk_0^*(w;\beta) = E[\Phi(\beta_0 + \beta_2 S(1))] = \int \Phi(\beta_0 + \beta_2 s) dF_{S(1)|W}(s)\cdot$$

Using the trick $E[\Phi(a+U)] = \Phi a + \mu/\sqrt{(1+\sigma^2)}$ with $E[U] = \mu$ and $Var[U] = \sigma^2$ we find that

$$risk_0^*(w_i;\beta) = E[\Phi\beta_0 + \beta_2 S_i(1)] = \Phi\left\{\frac{\beta_0 + \beta_2\mu^*(w_i)}{\sqrt{1+\beta_2^2 * \sigma^{*2}(w)}}\right\}.$$

Under the assumption that S(1) and W are bivariate normal, $S(1)|W = w$ is normal with mean $\mu^*(w) = \mu_S + \rho_{SW} * \sigma_S/\sigma_W(w-\mu_W)$ and variance $\sigma^{*2}(w) = \sigma_S^2(1-\rho^2)$. As these are unknown, Follmann suggests the use of the empirical estimates from the validation sample, those subjects in vaccine arm with both W and S(1) measured. In the case of BIV+CPV full sampling this will be all the vaccine recipients.

General Two-Phase Sampling

Gilbert and Hudgens (2008) extend the fully parametric EML of Follmann (2006) to two-phase sampling of a BIP. The Phase I covariates are a set of baseline covariates, X, measured on all subjects. The Phase II covariate is the BIP, W, that is only measured for cases, those infected, and a randomly selected cohort of uninfected controls. Normally the number of controls are directly proportional to the number of cases, e.g., a 5:1 control:case sampling, within a treatment arm. This extends the EML method of Follmann (2006) to include the nuisance parameters $\nu = (F_{S(1)|X,W}, F^{W|X})$, where $F^{W|X}$ is the conditional CDF of W given the baseline covariates X and $F_{S(1)|X,W}$ is the conditional CDF of S(1) given X,W. Gilbert and Hudgens (2008) allow for BIP, not just BIV, by including W in the model for outcome, as well as for other baseline covariates in the risk model.

The X and W and the X and W interactions with treatment Z can also be included in the model for risk. However, Gilbert and Hudgens (2008) show that one of the interactions must be assumed to be zero in order to identify the estimand. Adding CPV allows this assumption to be relaxed. The Gilbert and Hudgens (2008) A4-P links risk to the covariates that include X and W, or just X, depending on the sampling. They extend the $risk_z(s_1;\beta)$ notation and introduce $risk_z(s_1, w, x;\beta)$. For subjects with S(1) measured the likelihood contribution in the binary clinical endpoint case is given by:

$$risk_z(S(1), X, W;\beta)^Y \times (1 - risk_z(S(1), W, X;\beta))^{(1-Y)}.$$

As Gilbert and Hudgens (2008) only deal with the two-phase sampling of W while S(1) is measured for all vaccine recipients and no CPV is preformed, the above likelihood contribution would only

be possible for vaccine recipients. For subjects with W measured but $S(1)$ missing, the likelihood contribution is given by:

$$\left(\int risk_z(s_1, W, X; \beta) dF_{S(1)|X,W}(s_1|X, W)\right)^Y \times \left(1 - \int risk_z(s_1, W, X; \beta) dF_{S(1)|X,W}(s_1|X, W)\right)^{(1-Y)}.$$

For subjects with S(1) and W missing the likelihood contribution is given by the double integral,

$$\left(\int\int risk_z(s_1, w, X; \beta) dF_{S(1)|W,X}(s_1|w, X) dF^{W|X}(w|X)\right)^Y$$
$$\times \left(1 - \int\int risk_z(s_1, w, X; \beta) dF_{S(1)|W,X}(s_1|w, X) dF^{W|X}(w|X)\right)^{(1-Y)}.$$

The nuisance parameters $\nu = (F_{S(1)|X,W}, F^{W|X})$ are estimated in vaccine recipients that have W and X measured independent of β. The estimation of these distributions can be weighted to account for bias sampling of W within the vaccine arm.

A different type of two-phase sampling for BIP+CPV augmented trials can be accomplished by treating the BIP as the Phase I or baseline covariate and S(1)/S^C as the Phase II measurements. If sampling is performed in a case-control manner in both arms, as was done in Gilbert et al. (2011b), all vaccine cases have S(1) measured and a random cohort of vaccine controls also have S(1) measured. Infected placebo recipients clearly cannot have CPV taken, but a random sample of placebo controls proportional to the number of placebo cases having CPV. As was found in Gilbert et al. (2011b), sub-sampling can reduce power significantly regardless of the proportion sampled.

The likelihood contributions are as in Follmann (2006), with all subjects who do not have S(1) measured via CPV or otherwise contributing $risk_0^*(w; \beta)$ or $1 - risk_0^*(w; \beta)$. The nuisance parameter, $\nu = (F_{S(1)|W})$, is estimated in vaccine recipients with S(1) and W measured. Inverse probability weights (IPW) are used in the estimation of the moments of S(1) and W to account for the two-phase sampling scheme within vaccine recipients; the IPWs are one and one over the proportion of controls sampled for the cases and controls, respectively. In Gilbert et al. (2011b) the CPV sampled in this manner did not increase efficiency or power over BIP full sampling alone, when there was a strong correlation between W and S. Refinement of the Follmann (2006) method to improve efficiency under CPV two-phase sampling with full BIP was investigated by Dr. Huang. As mentioned above, she found that the pseudoscore method eliminates this paradox.

Binary Outcome Non-parametric EML: SoP Evaluation

Gilbert and Hudgens (2008) introduce a non-parametric form of EML estimation for a categorical or binned potential surrogate $S(1)$ and BIP, W, with J and K levels, respectively. It is a special case of the binary outcome parametric EML method, subsection 4.3.1. With the assumption of a non-parametric risk model and non-parametric estimation of $F_{S(1)|W}$, as well as A1 through A3 and CB Gilbert and Hudgens (2008) model risk non-parametrically by:

$$risk_z(j,k;\beta) = \beta_{zj} + \beta'_k.$$

This defines the risk model by the $J * K$ levels of S(1) and W. Gilbert and Hudgens (2008) assume β'_k is independent of study arm and constrain the parameters such that $0 \leq \beta_{zj} + \beta'_k \leq 1$ for all z, j, k and $\sum_{k=1}^{K} \beta'_k = 0$. Then the observed likelihood is given by,

$$\begin{aligned}
\log L(\beta,\nu) &= \sum_i \left\{ \sum_{z,y} I[\delta_i = 1, Y_i = 0, Z_i = z, Y_i^\tau = 1] \times \log \left\{ \sum_{j,k} \{\beta_{zj} + \beta'_k\}^y \{1 - \beta_{zj} - \beta'_k\}^{1-y} \nu_{jk} \right\} \right. \\
&+ \sum_{k,y} I[\delta_i = 0, Y_i = 0, Z_i = 0, W_i = k, Y_i^\tau = 1] \times \log \left\{ \sum_{j,k} \{\beta_{zj} + \beta'_k\}^y \{1 - \beta_{zj} - \beta'_k\}^{1-y} \nu_{jk} \right\} \\
&+ \left. \sum_{z,y} I[\delta_i = 0, Y_i = 0, Z_i = z, Y_i^\tau = 1] \times \log \left\{ \sum_{j,k} \{\beta_{zj} + \beta'_k\}^y \{1 - \beta_{zj} - \beta'_k\}^{1-y} \nu_{jk} \right\} \right\},
\end{aligned}$$

where δ is the indicator that S(1) is observed.

Here $\nu_{jk} = P(S(1) = j, W = k | Y^\tau = 1)$ is estimated non-parametrically in the vaccine recipients via:

$$\begin{aligned}
\nu_{jk} &= (n_1(j,k)/n_1)\mathrm{AR} + (n_0(j,k)/n_0)(1 - \mathrm{AR}) \qquad \text{Where} \\
\mathrm{AR} &= \sum_{i=1}^{n} Z_i Y_i^\tau I[Y_i = y] / \sum_{i=1}^{n} Z_i Y_i^\tau, \\
n_y(j,k) &= \sum_{i=1}^{n} Z_i Y_i^\tau (1 - \delta_i) I[Y_i = y, S_i = j, W_i = k], \\
n_y &= \sum_{i=1}^{n} Z_i Y_i^\tau (1 - \delta_i) I[Y_i = y] \qquad \text{and} \\
\nu_j &= \sum_k \nu_{jk}.
\end{aligned}$$

Gilbert and Hudgens (2008) show that $\hat{\mathrm{F}}^{S(1)|W}(j|k) = \sum_{i=1}^{j} \hat{\nu}_{ik} / \sum_{i=1}^{J} \hat{\nu}_{ik}$ and $\hat{\mathrm{F}}^{S(1)}(j) = \sum_{i=1}^{j} \hat{\nu}_i$

are consistent estimators, making the non-parametric estimated likelihood to be solved:

$$\begin{aligned}\log L(\beta,\hat{\nu}) &= \sum_i \Bigg\{ \sum_{z,y} I[\delta_i = 1, Y_i = 0, Z_i = z, Y_i^\tau = 1] \times log\{\sum_{j,k} \{\beta_{zj} + \beta'_k\}^y \{1 - \beta_{zj} - \beta'_k\}^{1-y} \hat{\nu}_{jk}\} \\ &+ \sum_{k,y} I[\delta_i = 0, Y_i = 0, Z_i = 0, W_i = k, Y_i^\tau = 1] \times log\{\sum_{j,k} \{\beta_{zj} + \beta'_k\}^y \{1 - \beta_{zj} - \beta'_k\}^{1-y} \hat{\nu}_{jk}\} \\ &+ \sum_{z,y} I[\delta_i = 0, Y_i = 0, Z_i = z, Y_i^\tau = 1] \times log\{\sum_{j,k} \{\beta_{zj} + \beta'_k\}^y \{1 - \beta_{zj} - \beta'_k\}^{1-y} \hat{\nu}_{jk}\} \Bigg\}.\end{aligned}$$

This method can be used with continuous $S(1)$ and W by binning into categorical variables. Simulations in Gilbert and Hudgens (2008) show the method has reasonable power and correct size in a HIV vaccine trial setting.

Discrete Time-to-Event Outcome EML

Using a similar EML method as applied in Follmann (2006) and Gilbert and Hudgens (2008), Qin et al. (2008) introduce an estimation method for the VE estimand using the Cox model framework and allowing for discrete time-to-event clinical endpoints. The discrete time-to-event setting requires different notation than the binary outcome case, however, the same assumptions are helpful in identifying the risk model. Qin et al. (2008) assume A1 through A3 with the addition of the assumption of random censoring. Let T(z) be the potential time-to-event under z and t_k be the time at the beginning of visit window k. If an event occurs between t_{k-1} and t_k under z, T(z) = t_k. Let $X(z) = min\{T(z), C(z)\}$, where C(z) is the potential censoring time under z.

In Qin et al. (2008) the estimand of interest $VE(s_1)$ is based on the hazard ratio:

$$VE(s_1) = 1 - \frac{P(T(1) = t_k | T(1) \geq t_{k-1}, S(1) = s_1, Y^\tau(1) = Y^\tau(0) = 1)}{P(T(0) = t_k | T(0) \geq t_{k-1}, S(1) = s_1, Y^\tau(1) = Y^\tau(0) = 1)}.$$

This is not a causal estimand due to conditioning on different risk sets. This is, however, an interesting estimand as it treats $S(1)$ as a baseline variable. Unless one conditions on risk sets based on both $T(0) \geq t_{k-1}$ and $T(1) \geq t_{k-1}$ in $risk_0$ and $risk_1$ (Hernán, 2010) which to our knowledge has never been done in the principal stratification framework, the estimands based on hazards are not causal.

Using the Cox framework, the risk is linked to the discrete partial likelihood, via the discrete cumulative hazard function, $\Lambda(t)$, given by,

$$d\Lambda(t_k; Z = z, Y(t), S(1) = s_1, Q = q; \beta) = exp(\beta_1 z + \beta_2 s_1 + \beta_3 s_1 z + \beta_4 q) d\Lambda_0(t_k) = exp(M'\beta) d\Lambda_0(t_k),$$

where Q represents a set of baseline covariates measured on everyone. The A4-P here is semi-parametric as the baseline hazard is not characterized. Using this link and assumption about the form of the clinical outcome T, $VE(s_1)$ can be represented as $exp(\beta_1 + \beta_3 s_1)$. Qin et al. (2008) suggests the null and alternative hypotheses $\beta_3 = 0$ and $\beta_3 < 0$ as a good test for some surrogate value. The basis for this test is the concept of causal sufficiency, although in this case it is not causal. They estimate $VE(s_1)$ via EML, taking expected values of the likelihood when $S(1)$ is missing.

Qin et al. (2008) allow for two-phase sampling of the BIP and CPV by integrating over an assumed distribution of $F_{S(1)|W,Q}$ for those subjects missing $S(1)$ but having the BIP and over $F_{S(1)|Q}$ for those missing both $S(1)$ and W. This is similar to the method used in Gilbert and Hudgens (2008), in that it requires a set of baseline covariates measured on everyone as the first-phase sample and treats the BIP as a second-phase measurement.

Qin et al. (2008) is the only published work in the SoP literature to allow for a time-to-event clinical endpoint. A EM-type algorithm process was also described in Qin et al. (2008), outlined below, that allowed for continuous time clinical endpoints but it assume proportional hazards.

Binary Outcome Semi-parametric EML

Huang and Gilbert (2011) introduces a semi-parametric EML estimation method. Starting with the same estimated likelihood as Gilbert and Hudgens (2008) and Follmann (2006):

$$\widehat{\mathrm{L}}(\beta) \equiv \mathrm{L}(\beta, \hat{\theta}) \prod_{i \in \epsilon=1} \mathrm{P}_\beta(Y_i|S_i, W_i) \prod_{i \in \epsilon=0} \hat{P}(Y_i|Z_i, W_i),$$

where $\hat{P}(Y_i|Z_i, W_i) = \int \mathrm{P}_\beta(Y_i|s, W_i) d\mathrm{F}^{S(1)|Z_i, W_i}(s)$ and assuming a generalized linear model for risk. The glm for risk is given by:

$$\begin{aligned} risk_z(S(1), W) &\equiv \Pr(Y(z) = 1 | Y^\tau(1) = Y^\tau(0) = 1, S(1) = s_1, W = w) \\ &= g(\beta_0 + \beta_1 * z + \beta_2 * s_1 + \beta_3 * z * s_1 + \beta_4 * w + \beta_5 * w * z) \cdot \end{aligned}$$

Calling this model for risk assumption A4 and assuming A1 through A6 and CB hold, Huang and Gilbert (2011) estimate risk semi-parametrically.

Rather than using a parametric plug-in estimate $F_{S(1)|W_i}$, Huang and Gilbert (2011) fit $F_{S(1)|W}$ semi-parametrically using the residuals from the location-scale model of Heagerty and Pepe (1999)

in the vaccine recipients with both S(1) and W measured. Let the number of subjects in the validation sample number n_V and membership in which be indicated by $\epsilon = 1$. Then $F_{S(1)|W}$ is given by:

$$F_{S(1)|W} \sim F[\{s_1 - \mu(w)\}/\sigma(w)] = F(\varsigma),$$

where F is the CDF of univariate residual ς and $\mu(w)$ and $\sigma(w)$ are the location and scale parameters. Using the residuals from the location scale model, Huang and Gilbert (2011) fit $F_{S(1)|W}$ by:

$$F_{S(1)|W} = P(S(1) \leq s_1|W) = P\left\{\varsigma \leq \frac{s_1 - \mu(w)}{\sigma(w)}\right\} = F\left\{\frac{s_1 - \mu(w)}{\sigma(w)}\right\}$$

Therefore, $F_{S(1)|W}$ can be estimated by fitting the location-scale parameters $\mu(w)$ and $\sigma(w)$. Huang and Gilbert (2011) show that by assuming $\mu(w)$ and $\sigma(w)$ are parametric forms of W, $\mu_w = \gamma' w$ and $log\{\sigma_w\} = \eta' w$ one can estimate them by solving the equations,

$$\sum_{k=1}^{n_V} \frac{w_k(s_{(1,k)} - \gamma' w_k)}{\sigma(w_k)^2} = 0$$

and

$$\sum_{k=1}^{n_V} \frac{w_k\{(s_{(1,k)} - \gamma' w_k)^2 - \sigma(w_k)^2\}}{\sigma(w_k)^2} = 0.$$

This yields a series of residuals, ς_k where k refers to the kth member of the validation sample. Using the estimate for $F_{S(1)|W}$ derived in Huang and Gilbert (2011) use the estimates γ' and η' to impute the individual missing S(1) values:

$$S^*_{i,k}(1) = \hat{\gamma'} w_i + \exp(\hat{\eta}' w_i)\varsigma_k.$$

There are K total imputations for each missing S(1) value. For an individual i with $\epsilon_i = 0$ the estimated likelihood contribution is given by:

$$\hat{P}(Y_i|Z_i, W_i) = \int risk_{z_i}(S(1), W_i) d\mathrm{F}^{S(1)|W_i}(s) = \frac{1}{n_v} \sum_{k=1}^{n_v} risk_{z_i}(S(1) = S^*_{i,k}(1), W_i).$$

Using this estimate of $\hat{P}(Y_i|Z_i, W_i)$ in the estimated likelihood, it is stated in Huang and Gilbert (2011) that an EM algorithm can be used to estimate consistently the risk parameters β given assumptions A1 through A6.

For a given set of β estimates each imputed value has an associated weight, which is derived from taking the score of the above likelihood. The weight associated with $S^*_{i,k}(1)$ for individual i,

$w_{i,k}$, is given by,

$$w_{i,k} = \frac{P(Y_i = 1|S^*_{i,k}(1), W_i)^{Y_i} \times \{1 - P(Y_i = 1|S^*_{i,k}(1), W_i)\}^{(1-Y_i)}}{\sum_{k=1}^{n_V} P(Y_i = 1|S^*_{i,k}(1), W_i)^{Y_i} \times \{1 - P(Y_i = 1|S^*_{i,k}(1), W_i)\}^{(1-Y_i)}},$$

or more generally $w_{i,k}$ is the likelihood contribution for a given imputed value of $S(1)$ over the sum of all the imputed likelihood contributions for that individual.

Huang and Gilbert (2011) suggest the EM-algorithm to maximize the estimated likelihood. Using the weights based on a set of β^0 starting values, fit a weighted GLM for $P(Y = 1|S(1), W) = risk_1(s_1, W; \beta) = g(\beta_0 + \beta_1 z + \beta_2 s_1 + \beta_3 z * s_1)$ to the augmented data set and update the β estimates. Then recalculate the associated weights given the updated β estimates, the expectation step, and refit the weighted GLM, the maximization step, repeating until the β values converge.

One of the main advantages of this method as outlined in Huang and Gilbert (2011) is its easy application to multiple biomarkers by estimating the location-scale parameters for each biomarker separately and using them to estimate their joint distribution, $F^{S_1(1),\ldots,S_n(1)|W}$, via:

$$P[\{\varsigma_1 \leq \frac{s_1 - \hat{\mu_1}(w)\}}{\hat{\sigma_1}(w)}, \ldots, \varsigma_n \leq \frac{s_n - \hat{\mu_n}(w)\}}{\hat{\sigma_n}(w)}]$$

This method also allows for a more flexible modeling of the potential surrogate, reducing the number of assumptions needed for identification. Bootstrap inference is still needed, as this is a EML and infected placebo recipients have zero probability of being in the validation set.

Huang and Gilbert (2011) illustrate their method under assumptions CB and A1 through A6 with all subjects having the BIP, W, measured and all uninfected placebo recipients receiving CPV. This method is easily extended to two-phase sampling of CPV and S(1). To account for the two-phase sampling of S(1) in the vaccines, one can weight both the location-scale estimating equations and the contribution in calculating the CDF of the baseline residuals with the inverse probability of being in the validation sample. For those uninfected placebo recipients who did not receive CPV, the same imputation method as stated for the infected placebo recipients can be used to estimate their likelihood contribution. To accommodate two-phase sampling of the BIP, W, the assumed form of $F_{S(1)|W}$ needs to be switched to a Breslow and Wellner (2007) model, a weighted, semi-parametric likelihood.

Discrete Time-to-Event Outcome Calibration-Based EM

This method briefly outlined in Qin et al. (2008) allows for continuous and discrete time-to-event clinical endpoints. However, the approximate EM-type estimation proposed can only reliably estimate the Cox model parameters from,

$$d\Lambda(t_k; Z = z, Y(t), S(1) = s_1, Q = q) = exp(\beta_1 z + \beta_2 s_1 + \beta_3 s_1 z + \beta_4 q) d\Lambda_0(t_k)$$

in the case of BIP-alone design. Using a continuous BIP, W, Qin et al. (2008) outline the procedure for using regression calibration to impute missing S(1) values for those subjects with W measured and an EM-type algorithm based on full likelihood for those subjects without W measured. Assuming that S(1) missingness and censoring of T(z) does not depend on S(1), they give the observed log-likelihood for the BIP-alone design:

$$\begin{aligned}
\ell(\beta, \alpha, \Lambda_0) &= \sum_{i \epsilon IC} \{\delta_i (M_i' \beta) - \Lambda_0(X_i) exp(M_i' \beta)\} \\
&+ \sum_{i \epsilon \bar{IC}, IB} log \left\{ \int exp \left\{ \delta_i (M_i' \beta) - \Lambda_0(X_i) exp(M_i' \beta) \right\} dF_{S(1)|Z_i, W_i, Q_i}(s_1) \right\} \\
&+ \sum_{i \epsilon \bar{IC}, \bar{IB}} log \left\{ \int exp \left\{ \delta_i (M_i' \beta) - \Lambda_0(X_i) exp(M_i' \beta) \right\} dF_{S(1)|Z_i, Q_i}(s_1) \right\}.
\end{aligned}$$

This likelihood can be solved using the EM algorithm (Chen and Little, 1999). Qin et al. (2008) modify the EM process by imputing missing $S_i(1)$ by the expected value of S(1) given W for those who have W measured. Assuming a parametric form of S(1) given W, $E(S(1)|W) = g(\beta_i, \theta)$, $E(S(1)|W)$ is estimated by $\hat{E}(S(1)|W) = g(\beta, \hat{\theta})$. Imputed values are then treated as known in the EM steps that follow. Using either starting or the last estimated values for $(\beta, \Lambda_0(X), \alpha, P_{kli}, \theta)$, Qin et al. (2008) calculate the conditional expectations under $F_{S(1)|Z,W,Q_d,\delta}$. Where P_{kli} denotes the probability mass of observed values S(1) at discrete levels of Z and Q_d. Where Q_d are the discrete members of the Phase I covariate set, Q_c are the continuous. The symbol α denotes the unknown nuisance parameters in the distribution of $F^{Q_c|S(1),Z,Q_d,X,\delta}$. Values are updated by solving the resulting expected score equation. The process is repeated until convergence.

The imputation of the missing $S(1)$ makes this procedure approximate and is referred to as the Approximate Calibration-Based EM (ACEM) in Qin et al. (2008). This procedure can account for continuous time and is computationally faster than the EML method for discrete time. Due to the

imputation the missing $S(1)$ estimates of $E(S(1)|W)$ by $E(S(1)|W, X, \delta)$ this method only works well in rare event settings. This was shown in Prentice (1982) and restated by Qin et al. (2008).

4.4 Bayesian Method

Li et al. (2010) develop a Bayesian imputation estimation method for evaluating the correlation between treatment effects on a binary clinical endpoint and treatment effects on a binary potential surrogate. To our knowledge, this is the only paper in the literature that does not assume CB in the formulation of their estimation method. Huang and Gilbert (2011) do provide instruction on how to reformulate their methods without assumption of CB. Li et al. (2010) impute missing $S(0)$ as well as missing $S(1)$. Considering all the counterfactual measures Li et al. (2010) classify subjects into groups by their potential outcomes $(Y(0), Y(1))$ and their potential surrogate measurements, $(S(0), S(1))$. There are 16 possible counterfactual probabilities which (Li et al., 2010) reduce to nine by assuming no harm of treatment and only positive effect on the surrogate by treatment. Thus, $S(1) \geq S(0)$ and $Y(1) \leq Y(0)$. Table 4.4 outlines the probabilities from the counterfactual model under these assumptions.

Table 4.1: Table of counterfactual probabilities

	$(Y(0), Y(1))$		
$(S(0), S(1))$	(1,1)	(1,0)	(0,0)
(0,0)	p_{11}	p_{12}	p_{13}
(0,1)	p_{21}	p_{22}	p_{23}
(1,1)	p_{31}	p_{32}	p_{33}

(Li et al., 2010)

Li et al. (2010) define r_z to be the number of patients in the $Z = z$ group $\{z = 0, 1\}$ and $r = r_0 + r_1$; r_{zst} is the number of patients for each of the combinations of Z, S and T. Using these, the observed

likelihood is defined as:

$$\begin{aligned} L_{obs} &= (p_{11} + P_{12} + P_{21} + P_{22})^{r000}(p_{13} + p_{23})^{r001}(p_{31} + p_{32})^{r010}p_{33}^{r011} \\ & \quad p_{11}^{r100}(p_{12} + p_{13})^{r101}(p_{21} + P_{31})^{r110}(P_{22} + p_{23} + p_{32} + p_{33})^{r111} \end{aligned}$$

Li et al. (2010) define n_{jk} to be the cell count corresponding to the counterfactual probability in the jth row and kth column where $j, k = 1, 2, 3$, with n_{jk}^z for treatment group z. Thus, the complete data likelihood is:

$$L_{com} = p_{11}^{n_{11}} p_{12}^{n_{12}} p_{13}^{n_{13}} p_{21}^{n_{21}} p_{22}^{n_{22}} p_{23}^{n_{23}} p_{31}^{n_{31}} p_{32}^{n_{32}} p_{33}^{n_{33}}$$

Li et al. (2010) assume that $E(n_{jk}^z) = \mu_{jk}$ and specify a model for μ_{jk} given by:

$$log\mu_{jk} = \lambda + \lambda_{jS} + \lambda_{jT} + \lambda_{jk}.$$

This is log-linear model for n_{jk}^z, where λ_{jS} and λ_{jT} denote the row and column effects and λ_{jk} their interaction. Enforcing the assumptions given above, ($\lambda_{2S} = \lambda_{2T} = \lambda_{j2} = \lambda_{2k} = 0$) and the following log odds ratios are given by the four corners of table 4.4:

$$\begin{aligned} log(OR_1) &= log((\mu_{11}\mu_{22})/(\mu_{12}\mu_{21})) = \lambda_{11}, \\ log(OR_2) &= log((\mu_{12}\mu_{23})/(\mu_{13}\mu_{22})) = -\lambda_{13}, \\ log(OR_3) &= log((\mu_{21}\mu_{32})/(\mu_{22}\mu_{31})) = -\lambda_{31}, \\ log(OR_4) &= log((\mu_{22}\mu_{33})/(\mu_{23}\mu_{32})) = \lambda_{33}. \end{aligned}$$

The positive association between S and Y implies that λ_{11} and λ_{33} are positive and that λ_{13} and λ_{31} are negative. Exploiting this fact, the counterfactual probabilities can be expressed via the λ parameters as:

$$p_{jk} = \frac{\exp(\lambda_{jS} + \lambda_{jS} + \lambda_{jS})}{\sum_j \sum_k \exp(\lambda_{jS} + \lambda_{jS} + \lambda_{jS})}.$$

Li et al. (2010) estimate this via a Bayesian approach. They treat the unobserved potential outcomes as missing and use Little and Rubin (2002) data augmentation and a Metropolis-Hastings algorithm for fitting of the model, assuming Gamma priors for the coefficients. The lengthy discussion of prior consideration and sensitivity in Li et al. (2010) illustrates the computational complexity of this method. However, it is shown to be a viable estimation technique in the binary-outcome, binary-potential-surrogate setting.

4.5 Pseudoscore Method for Discrete Outcome and Potential Surrogate

Wolfson (2009) and Huang et al. (2012a) extend the work on two-phase sampling of Chatterjee et al. (2003) to the couterfactual framework. The main advantages of the pseudoscore approach is that closed-form variance expressions can be obtained for the joint risks in terms of observed quantities. Also, as was shown in Huang et al. (2012a), the Pseudoscore method's exploitation of the randomness assumption leads to increased efficiency over the EML methods in many settings. There are three types of subjects whose level of the potential surrogate under vaccination could possibly be observed, which we denote by $\delta = 1$. Those groups are: uninfected vaccine recipients, infected vaccine recipients and uninfected placebo recipients. The pseudoscore method requires three assumptions beyond A1 through A6, as outlined in Wolfson (2009):

- Ps1: $\delta \perp S(1)|Z, Y, W$ sampling of δ is independent of S(1) conditional on the observed data
- Ps2: $\int_y \phi(y, Z, W)dy > 0$ For all ϕ in the neighborhood of the true ϕ_0, where $\phi(y, Z, W) = P(\delta = 1|Y = y, Z = z, W = w)$.
- Ps3: $f_\beta(Y|Z = z, S(1) = s_1, W = w) > 0$ For almost all observed data in the neighborhood of the true β_0.

Assumption Ps1 will most likely hold in clinical trials, as sampling should only depend on $S(1)$ via outcome, the ability to measure $S(1)$, $T > \tau$ and the observed baseline covariates W. Assumption Ps2 can only be true in the presence of CPV, implying the need for Assumptions A5 and A6, Section 4.2.1. Assumption Ps3 relies not only on Assumption A4-P holding, but also on there being positive risk at all times. Thus, when there is a binary outcome such as Y there can be no perfect prediction of Y given S(1), W, and Z. Unlike in Wolfson (2009), where A3 is relaxed to monotonicity, with the assumption of full subject level equal risk until τ there is no need to include the sensitivity parameters in the formulation of risk. Given Assumption A1 though A6 and assumption Ps1 through Ps3, the likelihood takes the form:

$$L(\beta) = \prod_i f_\beta(Y_i|Z_i, S_i(1), W_i, T > \tau, [Z_iY_i + Z_i + (1 - Z_i)(1 - Y_i)]\delta_i S_i(1).$$

This gives the standard likelihood contributions from those with $\delta_i = 1$ either in the placebo or vaccine arms. The likelihood contribution for those without $S(1)$ measured is given by,

$$\int_{s_1} f_\beta(y|Z = z, s, W = w)dF(s|Z, W),$$

where $F(s|Z, W) = P(S(1) \leq s|Z = z, T > \tau, W = w)$. Then we can write the score function as:

$$\text{Score} = (S_1 + S_2) + (S_3 + S_4) = 0,$$

where S_k indicates the group: vaccine recipients with S(1) measured, S_1, Placeboees with S(1) measured, S_2, and vaccine recipients and placebo recipients without $S(1)$ measured, S_3 and S_4. Score contributions are standard for $k = \{1, 2\}$

$$S_k = \frac{\partial L_k/\partial\beta}{L_k}.$$

The score contribution for vaccine recipients without S(1) measured is given by:

$$S_3 = \frac{\int_s \frac{\partial f_\beta(y|1,s,W)}{\partial\beta} dF(s|1, W)}{\int_s f_\beta(y|1, s, w)dF(s|1, W)}.$$

Similarly, for S_4 we have the same formula with $Z = 1$ replaced with $Z = 0$. Estimation of $F(s|Z, W)$ is complicated by the biased sampling. Chatterjee et al. (2003) suggests an estimator, the argument for which is given in Wolfson (2009) and repeated here; let

$P_\delta(Z = z, S(1) = s_1, W = w) = P(\delta = 1|Z = z, T > \tau, S(1) = s_1, W = w)$.

Then, by Bayes' Theorem: $P_\delta(Z = z, S(1) = s_1, W = w)$

$$\begin{aligned}
&= \frac{dP(S(1) \leq s_1, T > \tau|Z = z, \delta = 1, W = w)P(Z = z, W = w, \delta = 1)}{P(Z = z, T > \tau, S(1) = s_1, W = w)} \\
&= \frac{P(S(1) \leq s_1|Z = z, T > \tau, W = w, \delta = 1)P(T > \tau|Z = z, W = w, \delta = 1)P(Z = z, W = w, \delta = 1)}{dF(s_1|Z = z, W = w)P(T > \tau, W = w)} \\
&= \frac{P(S(1) \leq s_1|Z = z, T > \tau, \delta = 1, W = w)P(\delta = 1|Z = z, T.\tau, W = w)}{dF(S(1)|Z = z, W = w)}.
\end{aligned}$$

Rearranging, we obtain

$$\begin{aligned}
dF(s_1|Z = z, W = w) &= \frac{DP(S(1) \leq s_1|Z = z, T > \tau, \delta = 1, W = w)P(\delta = 1|Z = z, T > \tau, W = w)}{P_\delta(Z = z, S(1) = s_1, W = w)} \\
&\equiv \frac{dF^*(s_1|Z = z, W = w)P(\delta = 1|Z = z, T > \tau, W = w)}{P_\delta(Z = z, S(1) = s_1, W = w)},
\end{aligned}$$

provided $P_\delta(Z = z, S(1) = s_1, W = w) > 0$, almost surely. Then, given Ps1:

$$\begin{aligned} & P_\delta(Z = z, S(1) = s_1, W = w) \\ = & \int_y P(\delta = 1|y, Z = z, T > \tau, S(1) = s_1, W = w) P(y|Z = z, T > \tau, S(1) = s_1, W = w) dy \\ = & \int_y P(\delta = 1|y, Z = z, T > \tau, W = w) f_\beta(y|Z = z, S(1) = s_1, W = w) dy \\ = & \int_y \phi(y, Z = z, W = w) f_\beta(y|Z = z, S(1) = s_1, W = w) dy\cdot \end{aligned}$$

If ϕ and β are the true values by Ps2 and Ps3, $P_\delta(Z, S(1), W) > 0$. Then $P_\delta(Z, S(1), W)$ can be estimated via:

$$\widehat{P}_\delta(Z = z, S(1) = s_1, W = w) = \int_y \widehat{\phi}(y, Z, W) f_{\widehat{\beta}}(y|Z, S(1), W) dy\cdot$$

Then we can estimate S_3 by S_3^{Ps} where S_3^{Ps} is given by,

$$S_3^{Ps} = \frac{\int_s \frac{\partial f_\beta(y|1,s,W)}{\partial \beta} \frac{dF^*(s|1,W)}{P_\delta(1,s,W)}}{\int_s f_\beta(y|1, s, w) \frac{dF^*(s|1,W)}{P_\delta(1,s,W)}}.$$

Similarly, for S_4 we use S_4^{Ps}. The pseudoscore estimator is given by,

$$Sc^{Ps}(\beta, F^*, \phi) = (S_1 + S_2) + (S_3^{Ps} + S_4^{Ps}) = 0,$$

and this can be solved via Newton-Raphson. Wolfson (2009) follows the proof of Chatterjee et al. (2003) and states that when Y, W and $S(1)$ are all discrete, β^{Ps} is asymptotically normal with closed form variance Ω; where Ω is a sandwich variance of the same form given in Chapter 2.

The extension to the counterfactual framework obtained in Wolfson (2009) assume A1 though A3 which implies that $f_\beta(y|Z, S(1), W) = f_\beta(Y(z) = y|S(1), W, T > \tau)$. Wolfson (2009) leaves all other derivations as they are given in Chatterjee et al. (2003). Wolfson (2009) finds that the method is unbiased and markedly more powerful and efficient than the EML in the same setting. We explore these findings for our extension of this method.

In the submitted work Huang et al. (2012a) develop the pseudoscore method for surrogate evaluation for binary clinical outcomes. They state that assumptions A1 through A3 also imply that $F(S(1)|Z, W) = F(S(1)|W)$. Thus, all derivations for the estimate of $F(S(1)|W)$ should not depend on Z. The Huang et al. (2012a) pseudoscore method then integrates over z as well as y in the calculation of $P_\delta(S(1) = s_1, W = w)$. Huang et al. (2012a) find that this adaptation of the method improves

efficiency. In our extension of this method to time-to-event data with possible time varying effects we do not consider this extension. This is of future research interest.

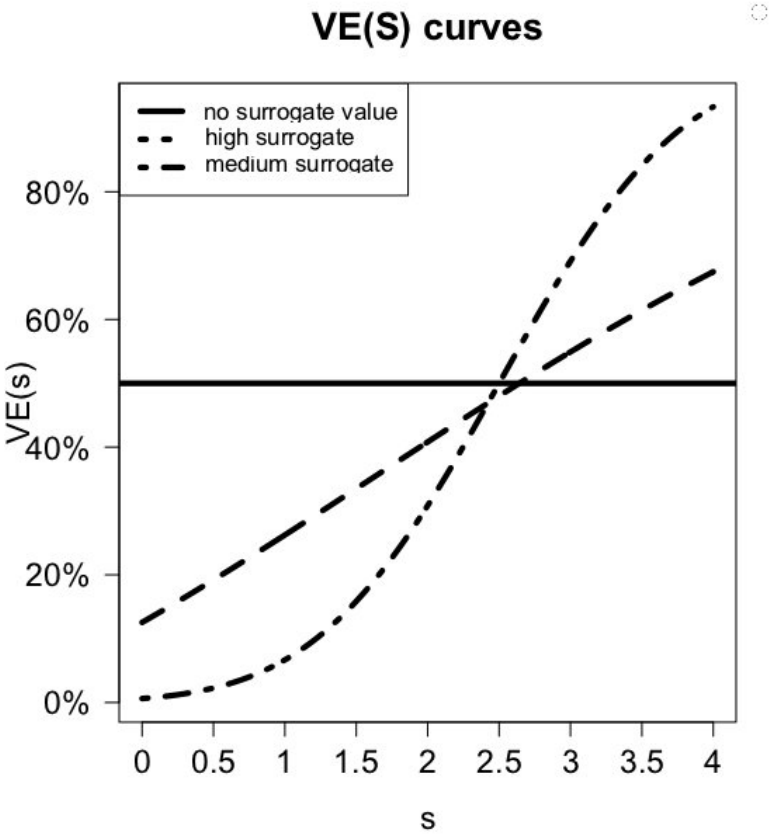

Figure 4.1: For CB with S(0)=0, S has no (horizontal solid line), modest (dashed line) and high (dash-dotted) surrogate value based on VE(s)

Chapter 5

TIME-DEPENDENT SOP ESTIMANDS

We introduce time-dependent SoP estimands using continuous time-to-event clinical outcomes. This could be adapted for discrete time-to-event outcomes, but we do not investigate this here. The notation for a continuous time-to-event setting differs from the binary outcome notation. As in the binary clinical endpoint setting, let Z denote the type of treatment we are testing in a randomized trial, ($Z \epsilon \{0, 1\}$), $Z = 1$ treatment, $Z = 0$ placebo. Let $T(z)$ be the potential time-to-event outcome under treatment z, for $z = (0, 1)$. Let $S(z)$ be the potential biomarker value under treatment z measured at a fixed time-point τ after treatment. Let $C(z)$ be the potential censoring time under z, and let $X(z) = \min(T(z), C(z))$. Let $Y(z)$ indicate that $T(z) = X(z)$. Let δ be the indicator that $S(1)$ is observed. If $X_i < \tau$ then $S = *$ is undefined and we exclude the subject from evaluation; the population of interest are those subjects at risk at time τ. Let W be a baseline variable measured on all or some sample of the subjects prior to randomization. W will serve as our BIP. Let Q be a set of baseline variables measured on all subjects. The full set of potential and observed outcomes and potential and observed biomarker measurements for each individual i, $\{S_i(1), S_i(0), T_i(1), T_i(0), C_i(1), C_1(0), Y_i(1), Y_i(0), W_i, Q_i, \delta\}$ are assumed to be iid. Let $F_{S(z)}$ denote the distribution of $S(z)$, and $F_{S(z)|W}$ the conditional distribution of $S(z)$ given W. Let $F^{-1}_{S(1)}$ be the quantile function of $S(1)$, with $c_1(\nu)$ denoting the νth quantile of $S(1)$. Similarly, $F^{-1}_{S(1)|W}$ and $c_{(1,w)}(\nu)$ are the quantile function and the quantile value conditional on W. Let μ_S, μ_W, σ^2_S σ^2_W denote the first and centered-second moments of $S(1)$ and W, and ρ^2_{SW} denote the correlation between $S(1)$ and W.

5.1 Time-dependent Risk

We propose to extend the potential surrogate-dependent risk $risk_z(s_1)$ to time dependence. This extension requires the inclusion of time in the definition of risk, $risk_z(t|s_1)$. There are many ways to define risk in the continuous time-to-event setting. One may define $risk_z(t|s_1)$ based on cumulative

risk at time t,

$$risk_z^{CDF}(t|S_1) \equiv Pr(T(z) \leq t|S(1) = s_1, T(1) > \tau, T(0) > \tau),$$

or based on the hazard function,

$$risk_z^{HZ}(t|s_1) = \lambda_z(t|s_1).$$

We write $risk_z^{HZ}(t|s_1)$ here for simplicity of notation, the risk model can include other baseline covariates include the BIP, W. Defining risk in this way allows one to consider surrogate-dependent risk at a particular timepoint after measurement of the potential surrogate. Characteristics of the surrogate and clinical endpoint relationship can then be looked at with reference to time and not over the full amount of follow-up. This allows us to investigate potential surrogates that may have been missed due to declining associations between the treatment effect on the surrogate and treatment effect on the outcome. Trials that find no vaccine efficacy at the close of event follow-up due to rapid waning of treatment effects may still be valuable resources for surrogate evaluation, if the time dependence can be characterized and the surrogate quality considered accounts for that time dependence.

5.2 Time-dependent Vaccine Efficacy (VE)

We extend the concept of surrogate-depend VE, introduced in Gilbert and Hudgens (2008), to include time dependence. We define time-dependent and surrogate-dependent VE in terms of the time-dependent risks, given above, as:

$$VE(t|s_1) = 1 - risk_1(t|s_1)/risk_0(t|s_1),$$

where $risk_z(t|s_1)$ could be any of the forms of time-dependent risk; $VE(t|s_1)$ can vary in both time and $S(1)$. For a useful SoP, $VE(t|s_1)$ will vary greatly in s_1 for at least some time t. This is the time-varying adaptation of the relaxed definition of a useful SoP used in Gilbert et al. (2011b) and Wolfson and Gilbert (2010). Changes in VE over time imply that VE is either changing over follow-up time, independent of the potential surrogate, or is changing over follow-up time due to time dependence in potential surrogate. Both forms of time dependence are simultaneously possible and both need to be modeled in order to fully characterize the time-variation of interest in this setting. We discuss one way to accomplish this in Chapter 6.

5.3 Time-dependent Summary Statistics Based on Causal Effect Predictiveness (CEP) Curve

As was first suggested in Gilbert and Hudgens (2008), the comparison between the risk estimates over the treatment arms is the causal estimand of interest for surrogate evaluation. Let us define $\Delta(t|s_1)$ to be one of those comparisons at a fixed time t; some examples include, $\Delta(t|s_1) = risk_0(t|s_1) - risk_1(t|s_1)$, $\Delta(t|s_1) = \log(risk_1(t|s_1)/risk_0(t|s_1))$ and $VE(t|s_1)$. These are all comparisons of risk between the treatment arms conditional on the potential SoP.

Let $Y^t(z)$ be the indicator that $T(z) \leq t$. Let $D(t) = Y^t(0) - Y^t(1)$ be the individual treatment effect on the clinical endpoint at time or before time t. This effect is what we are attempting to predict with all the $\Delta(t|s_1)$. Using these, the time-dependent sensitivity and specificity for the comparison of risk between the treatment arms conditional of S(1), $\Delta(t|s_1)$, be defined following Heagerty et al. (2000) by:

$$\begin{aligned}
\text{Sensitivity}(t|c) &= P(\Delta(t|s_1) > c|D(t) = 1) = TPR(t|c), \\
\text{Specificity}(t|c) &= P(\Delta(t|s_1) \leq c|D(t) = 0) = 1 - FPR(t|c),
\end{aligned}$$

As noted in Huang and Gilbert (2011) for the time-independent classification accuracy measures, the classification accuracy measures of Sensitivity$(t|c)$ and Specificity$(t|c)$ are based on unobservable data and therefore cannot be estimated empirically. However, Huang et al. (2012b) show that in our setting, if we assume monotonicity $Y^t(0) \geq Y^t(1)$ of treatment effect, they can be estimated parametrically.

Adapting Huang et al. (2012b) $TPR(t|c)$ and $FPR(t|c)$ can be defined via the model for risk by:

$$\begin{aligned}
TPR(t|c) &= \frac{P(D(t) = 1, \mathrm{I}\{\Delta(t|s_1) > c\})}{P(D(t) = 1)} \\
&= \frac{E\{(D(t) = 1)\mathrm{I}\{\Delta(t|s_1) > c\})\}}{E\{D(t) = 1\}} \\
&= \frac{E\{E\{D(t) = 1|S(1) = s_1\}\}\mathrm{I}\{\Delta(t|s_1) > c\})\}}{E\{E\{D(t) = 1|S(1) = s_1\}\}} \\
&= \frac{E\{\Delta(t|s_1)\}\mathrm{I}\{\Delta(t|s_1) > c\})\}}{E\{\Delta(t|s_1)\}}.
\end{aligned}$$

Similarly $FPR(t|c)$ can be defined by:

$$FPR(t|c) = \frac{E\{1 - \Delta(t|s_1)\}\mathrm{I}\{\Delta(t|s_1) > c\})\}}{E\{1 - \Delta(t|s_1)\}}.$$

Again adapting Huang et al. (2012b), one can estimate these accuracy measures by:

$$\widehat{TPR}(t|c) = \frac{\int \hat{\Delta}(t|s_1)\mathrm{I}\{\hat{\Delta}(t|s_1) > c\}d\hat{F}_{S(1)}(s_1)}{\int \hat{\Delta}(t|s_1)d\hat{F}_{S(1)}(s_1)}$$

$$\widehat{FPR}(t|c) = \frac{\int 1-\hat{\Delta}(t|s_1)\mathrm{I}\{\hat{\Delta}(t|s_1) > c\}d\hat{F}_{S(1)}(s_1)}{\int 1-\hat{\Delta}(t|s_1)d\hat{F}_{S(1)}(s_1)}$$

There have been many summary statistics for comparison of candidate SoP and SoP evaluation suggested in the literature. We discuss the extension of one such statistic, the STG of Huang and Gilbert (2011) and introduce to the SoP evaluation framework several summary statistics from the biomarker evaluation literature.

5.4 Time-dependent Standardized Total Gain (STG)

The novel use of standardized total gain to compare biomarkers as potential SoP (Huang and Gilbert, 2011) can further be extended to allow for time dependence. We must first define as did Huang and Gilbert (2011), $\rho_z(t) = Pr(Y^t(z) = 1) = E(Y^t(z)) = Pr(T(z) \leq t)$. Then following Gilbert and Hudgens (2008) and Huang and Gilbert (2011) let $R^t(v)$ be the vth quantile of $\Delta(t|s_1)$, it can be shown that if $\Delta(t|s_1) = risk_0(t|s_1) - risk_1(t|s_1)$, the risk difference for risk based on the CDF, the area under the $R^t(v)$ versus v curve is equal to the difference in prevalence of potential clinical outcomes between the two treatment arms before a fixed time t, $\rho_0(t) - \rho_1(t)$.

Using this, time-dependent total gain for a fixed timepoint can be defined for the CEP of $risk^{CDF}$ difference as:

$$TG(t) = \int_0^1 |R^t(v) - (\rho_0(t) - \rho_1(t))|dv,$$

Huang and Gilbert (2011) suggest a standardized version of TG for a binary outcome and we adapt this to the time-to-event setting by:

$$STG(t) = TG(t)/[2(\rho_0(t) - \rho_1(t))\{1 - \rho_0(t) + \rho_1(t)\}].$$

$STG(t)$ has a clinically relevant interpretation based on the classification accuracy measures, if one assumes no harm by treatment, $Y^t(0) \geq Y^t(1)$, given by:

$$STG(t) = max_c\{\text{Sensitivity}(t|c) + \text{Specificity}(t|c)\} - 1.$$

A modification of the proof of this form of $STG(t)$ given in Web Appendix A of Huang and Gilbert (2011) follows.

Proof. Let F^t, F^t_D, and $F^t_{\overline{D}}$ denote the CDF of $\Delta(t|s_1)$ for a given t in the general population, and the population with $D(t) = 1$ and $D(t) = 0$ respectively. Let f^t, f^t_D, $f^t_{\overline{D}}$ be the corresponding density functions. Denote $ROC(t|c) = F^t_D\{F^t_{\overline{D}}(c)\}$. Denote the vth quantile of $\Delta(t|s_1)$ to be $R^t(v)$ with $r = {F^t}^{-1}(v)$ and $F^t_{\overline{D}}(r) = 1 - c$, thus making $r = {F^t}^{-1}_{\overline{D}}(1 - c)$. Let LR denote the likelihood ratio function: $LR(r) = f^t_D(r)/f^t_{\overline{D}}(r)$. Then $R^t(v)$ can be defined by:

$$\begin{aligned} R^t(v) &= P\{D(t) = 1|\Delta(t|s_1) = r\} = \frac{P(D(t) = 1)LR(r)}{P(D(t) = 1)LR(r) + P(D(t) = 0)} \\ &= \frac{P(D(t) = 1)LR\{{F^t_{\overline{D}}}^{-1}(1 - c)\}}{P(D(t) = 1)LR\{{F^t_{\overline{D}}}^{-1}(1 - c)\} + P(D(t) = 0)} \\ &= \frac{P(D(t) = 1)ROC'(t|c)}{P(D(t) = 1)ROC'(t|c) + P(D(t) = 0)} \end{aligned}$$

and,

$$\begin{aligned} v &= F^t(r) = P(D(t) = 0)F^t_{\overline{D}}(r) + P(D(t) = 1)F^t_D(r) \\ &= P(D(t) = 0)(1 - c) + P(D(t) = 1)F^t_D\{{F^t_{\overline{D}}}^{-1}(1 - c)\} \\ &= P(D(t) = 0)(1 - c) + P(D = 1)\{1 - ROC(t|c)\} \\ &= 1 - P(D(t) = 0)c - P(D(t) = 1)ROC(t|c). \end{aligned}$$

Then figure $R^t(v)$ versus v is given by $P(D(t) = 1)ROC'(t|c)/P(D(t) = 1)ROC'(t|c) + P(D(t) = 0)$ versus $1 - P(D = 0)c - P(D = 1)ROC(t|c)$. Similarly to TG in Huang and Gilbert (2011), one can define $TG(t)$ by:

$$\begin{aligned} TG(t) &= 2P(D(t) = 1)P(D(t) = 0)sup_c\{\text{ROC}(t|c) - c\} \quad \Rightarrow \\ STG(t) &= max_c\{\text{Sensitivity}(t|c) + \text{Specificity}(t|c)\} - 1. \end{aligned}$$

□

As stated above, the assumption of no individual active harm by treatment, or monotonicity in individual treatment effect in cases like the Step trial, where we more safely assume no treatment effect or harm, is needed to connect the STG(t) to these accuracy measures. This may seem a

stronger assumption in the time-to-event setting than in the binary setting. We believe, depending on the time, t, of interest, it may actually be a weaker assumption. In the binary setting the assumption of no active harm, is for the full length of the trial. $Y^t(0) \geq Y^t(1)$ for any t less than the full amount of follow-up seems a weaker statement.

This connection with the accuracy measures suggests that a STG(t)+1 of 2 described the prefect SoP, and that based on the standard rule of thumb of a desired 80% sensitivity and specificity we would like to have a STG(t) of at least 0.6, translating to a maximum sensitivity($t|c$) and specificity($t|c$) of 0.8 each.

Although STG(t) is linked to the accuracy measures and can therefore suggest the usefulness of a particular biomarker as an SoP, inference on the STG(t) is primarily for comparison between potential SoP rather than for the evaluation of a biomarker as an SoP. Bootstrap percentile CI will almost never contain zero as STG(t) is theoretically bounded by zero.

Therefore, the STG(t) is more interesting as a mode of comparison between potential biomarkers and P-values for significant differences between the STG(t) of different potential SoP are of greatest interest. Confidence intervals that do not overlap the CI of the other potential SoP provide some evidence that one SoP is better than another, however as this does not account for potential correlation between the STG values. Therefore, P-values or CI for the difference in STG(t) are the preferred mode of inference for comparison.

We use plug-in estimators for $STG(t)$ and $TG(t)$ given by,

$$\begin{aligned} \widehat{TG}(t) &= \int_0^1 |\hat{R}^t(v) - (\hat{\rho}_0(t) - \hat{\rho}_1(t))|dv) \quad \text{and} \\ \widehat{STG}(t) &= \widehat{TG}(t)/[2(\hat{\rho}_0(t) - \hat{\rho}_1(t))\{1 - \hat{\rho}_0(t) + \hat{\rho}_1(t)\}. \end{aligned}$$

5.5 Time-dependent CEP-based Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

One may wish to consider how well the risk model is predicting protection or lack of protection separately. For this, the concepts of positive predictive value and negative predictive value are useful. Time dependent PPV and NPV were defined for a risk model $f(x)$ as $PPV(t|v) = Pr(Y^t = 1|f(x) \geq c)$ and $NPV(t|v) = Pr(Y^t = 0|f(x) \leq c)$ in Heagerty et al. (2000).

This concept can be extended to our setting using, a $\Delta(t|s_1)$ of risk difference, where risk is based

on the CDF, and $D(t) = Y^t(0) - Y^t(1)$. Defining $PPV(t|v) = Pr(D(t) = 1|\Delta(t|s_1) \geq c)$ and $NPV(t|v) = Pr(D(t) = 0|\Delta(t|s_1) \leq c)$. If we assume no active harm, or monotonicity in vaccine effect, $PPV(t|v)$ is the probability of protection given that $\Delta(t|s_1)$ is greater than or equal to c. Similarly, if we assume no active harm, $NPV(t|v)$ is the probability of no effect of vaccine given that $\Delta(t|s_1)$ is less than or equal to c. These can also be defined and estimated via the quantile curve $R^t(v)$, as defined above, adapted from Gu and Pepe (2011) by,

$$\begin{aligned} PPV(t|v) &= \int_v^1 R^t(\mu)d\mu/(1-v) \\ NPV(t|v) &= 1 - \int_0^v R^t(\mu)d\mu/(v), \end{aligned}$$

The values of $PPV(t|v)$ and $NPV(t|v)$ at a particular v and t can be used as summary statistics. They differ from the time-dependent and covariate-specific $PPV_z^{'}$, outlined below, as they are based on the quantiles of a risk difference rather than the quantiles of the surrogate itself. Because $PPV(t|v)$ and $NPV(t|v)$ are model based, they can be used to evaluate different risk models, compare potential SoPs, or for SoP evaluation when compared to the clinical outcome prevalence difference over the trial arms.

Positive predictive value, $PPV(t|c)$ and $NPV(t|v)$ have an appealing clinical interpretation given in Gu and Pepe (2011) when we assume monotonicity of vaccine activity. By applying Bayes theorem we have:

$$\begin{aligned} PPV(t|v) &= \frac{(\rho_0(t) - \rho_1(t))}{\{1 - \rho_0(t) + \rho_1(t)\}} \frac{TPR(t|v)}{FPR(t|v)} = \frac{(\rho_0(t) - \rho_1(t))}{\{1 - \rho_0(t) + \rho_1(t)\}} \frac{\text{Sensitivity}(t|v)}{1 - \text{Specificity}(t|v)} \\ NPV(t|v) &= \frac{\{1 - \rho_0(t) + \rho_1(t)\}}{(\rho_0(t) - \rho_1(t))} \frac{1 - FPR(t|v)}{1 - TPR(t|v)} = \frac{\{1 - \rho_0(t) + \rho_1(t)\}}{(\rho_0(t) - \rho_1(t))} \frac{\text{Specificity}(t|v)}{1 - \text{Sensitivity}(t|v)}, \end{aligned}$$

where $\rho_1(t) = Pr(Y^t(z) = 1)$. We use the plug-in estimator for $PPV(t|c)$ and $NPV(t|c)$ given by:

$$\begin{aligned} \widehat{PPV}(t|v) &= \int_v^1 \hat{R}^t(\mu)d\mu/(1-v) \\ 1 - \widehat{NPV}(t|v) &= \int_0^v \hat{R}^t(\mu)d\mu/(v). \end{aligned}$$

Particular points on the $PPV(t|v)$ and $NPV(t|v)$ curves can be compared via confidence interval to the difference in prevalence of potential clinical outcomes, $(\rho_0(t) - \rho_1(t))$ for the same time t. For example, if $(\rho_0(t) - \rho_1(t))$ at year t is not contained in the CI for $\widehat{PPV}(t|0.8)$, this suggests that there is evidence to reject the null that there is not variation in predictive power of the model over

the potential surrogate. Said another way, this is evidence to support that there is higher probability of protection when the model for risk difference is greater than the 80th percentile of predicted risk difference. Most simply, there is evidence of increased probability of protection given that the model predicts protection. A SoP with causal necessity will have a $NPV(t|v)$ of one for some low v, but any $\widehat{NPV}(t|v)$ confidence interval that does not cover $1 - (\rho_0(t) - \rho_1(t))$ provides evidence that the risk model is predictive and therefore the candidate biomarker has some value as a SoP.

5.6 Time-dependent CEP-based Partial Total Gain (pTG)

The definition of $PPV(t|v)$ and the suggested comparison to $(\rho_0(t) - \rho_1(t))$ leads one to the question if these two concepts can be combined into a single summary statistic similar to STG. We consider the partial total gain (pTG) of Sachs (2011) for this purpose. A simple and intuitive definition pTG, given $PPV(t|v)$, might be,

$$pTG(t|v) = \int_v^1 |R^t(\mu) - (\rho_0(t) - \rho_1(t))| d\mu/(1 - v). \tag{5.1}$$

The pTG is defined more flexibly for a binary endpoint in the time-independent biomarker framework in Sachs (2011) over the lower tail $(0, v)$ and upper tail $(d, 1)$ of the quantile curve, $R(u)$, for a given risk model by,

$$pTG(B) = \frac{1}{v\rho + (1 - d)(1 - \rho)} \int_B |R(u) - \rho| \, d\, u,$$

where $B = \{(0, v) \bigcup (d, 1)\}$ and ρ is prevalence. The $pTG(B)$ can be standardized so that the measure takes values between 0 and 1. This can be extended to the causal framework for a time-to-event outcome by again using the quantile curve of the risk difference, $R^t(u)$. The standardized pTG in our setting is given by,

$$pTG(t|B) = pTG(t|d, v) = \frac{1}{v(\rho_0(t) - \rho_1(t)) + (1 - d)(1 - (\rho_0(t) + \rho_1(t))} \int_B |R^t(u) - (\rho_0(t) - \rho_1(t))| \, d\, u\cdot$$

The thresholds are quantiles of risk difference $v = R^t(c)$ and $d = R^t(q)$; if one sets $v = 0$ we arrive at the standardized version of equation 5.1. As given in Sachs (2011), if we assume that $v < (\rho_0(t) - \rho_1(t))$, $d > (\rho_0(t) - \rho_1(t))$ and there is no individual active harm by treatment, we can define the unstandardized $pTG(t|B)$ in terms of $PPV(t|q)$ and $NPV(t|c)$. When $R^t(d) = q$ and our

CEP of interest is risk difference this derivation is given by:

$$
\begin{aligned}
& \{v(\rho_0(t) - \rho_1(t)) + (1-d)(1 - \rho_0(t) + \rho_1(t))\} \cdot pTG(t|B), \\
= & \int_0^v |R^t(u) - (\rho_0(t) - \rho_1(t))|\, du + \int_d^1 |R^t(u) - (\rho_0(t) - \rho_1(t))|\, du \\
= & \int_0^v (\rho_0(t) - \rho_1(t)) - R^t(u)\, du + \int_d^1 R^t(u) - (\rho_0(t) - \rho_1(t))\, du \\
= & \; v(\rho_0(t) - \rho_1(t)) - v\{1 - NPV(t|c)\} + (1-v)PPV(t|q) - (1-d)(\rho_0(t) - \rho_1(t)) \\
= & \; v\{(\rho_0(t) - \rho_1(t)) - (1 - NPV(t|c))\} + (1-d)\{PPV(t|q) - (\rho_0(t) - \rho_1(t))\}.
\end{aligned}
$$

The $pTG(t|B)$ is of interest in the surrogate evaluation framework for comparison of potential SoP quality in cases where STG(t) is not useful or is too similar. Unlike the $TG(t)$ and $STG(t)$ which are summaries over the full quantile curve of the CEP, the $pTG(t|B)$ allows one to summarize a surrogate on a particular range of the risk difference. Finding a range on which the potential surrogate is of high-quality is the main goal of specific SoP comparison. Thus, the $pTG(t|B)$ could be a more interesting summary statistic than those that consider the entire range of risk difference.

Figure 5.1 illustrates the quantile functions for risk difference of two different potential SoP, a continuous biomarker S_i and a discrete biomarker S_j, which have the same total gain, 0.1. With percentile thresholds of 0.75 and 0.1, the partial total gains of S_i and S_j are different (0.06 for S_i and 0.1 for S_j). Both the risk-difference curves plotted in Figure 5.1 have area under the curve 0.2, indicating that the clinical outcome prevalence difference between arms is 0.2. The discrete potential SoP classifies the groups based on its risk model more clearly, but the STG does not illustrate this, and the difference is difficult to determine based on the figure alone.

Similarly to STG(c), inference on $pTG(t|d, v)$, is most useful for comparison of candidate SoP and risk models, rather than for evaluation of specific SoP. Confidence intervals will almost never include zero, suggesting that any candidate SoP has some partial surrogate value; CI should rather be compared to those of other candidate SoP. When sets of CI within the same trial for different candidate SoP do not overlap this provides some evidence that one of the SoP is better at classifying vaccine effect groups, with higher $pTG(t|d, v)$ implying better classification. However, just as with STG(t) P-values for the difference between two $\widehat{pTG}(t|d, v)$ being different than zero are the preferred means of inference.

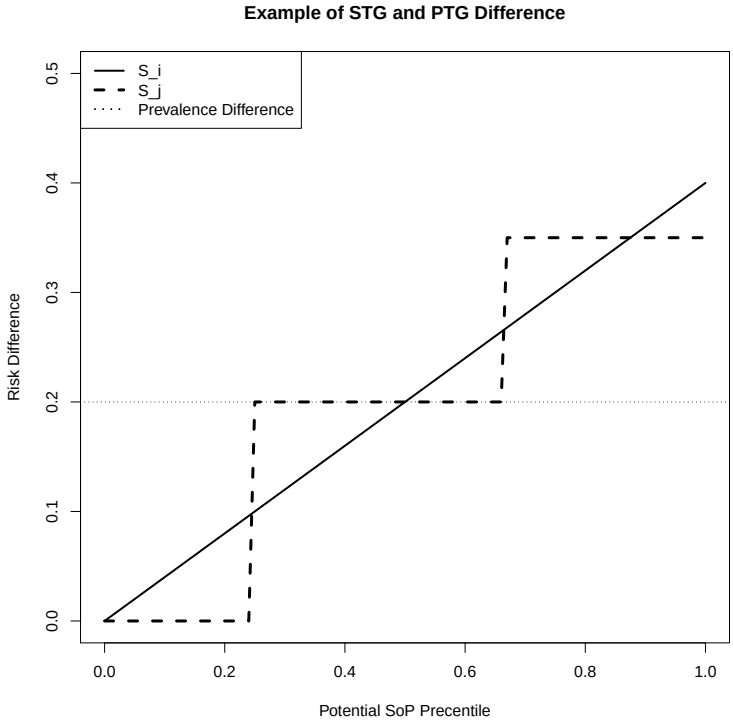

Figure 5.1: The potential SoPs depicted have the same STG(1.5)=0.1, but the step function of the difference, for the discrete potential SoP S_j, has a higher estimated $pTG(1.5|0.75, 0.1)$ of 0.1, while it is 0.06 for S_i. Figure adapted from Sachs (2011)

We use the plug-in estimator for $pTG(t|d, v)$ when the CEP is risk difference this is given by:

$$\widehat{pTG}(t|B) = \frac{1}{v(\widehat{\rho}_0(t) - \widehat{\rho}_1(t)) + (1-d)(1 - (\widehat{\rho}_0(t) + \widehat{\rho}_1(t))} \int_B |\widehat{R}^t(u) - (\widehat{\rho}_0(t) - \widehat{\rho}_1(t))| \, d\, u.$$

5.7 Time-dependent and Covariate-specific PPV

If we instead wish to consider risk at thresholds of the surrogate itself, we can use PPV as a measure of risk within a treatment group z. For this purpose, the time-dependent and surrogate-dependent PPV_x curve of Zheng et al. (2008) and Zheng et al. (2010) is desirable. The PPV_x is a useful depiction of the quality of a diagnostic biomarker because it compares predictive value of a diagnostic test at various thresholds. As thresholds are commonly used in practice for validated surrogates (Plotkin, 2008), a modified version of the PPV_x is appealing for SoP quality evaluation and comparison. It is

often of interest to quantify the ability of the surrogate to classify risk above a threshold, which can easily be assessed using the concept of PPV.

We modify the meaning of the time-dependent and covariate-specific PPV of Zheng et al. (2010), $PPV_z(t,v) = P\{T \leq t | S \geq c_z(v), Z = z\}$, to be the probability of survival above a given threshold of $S(1)$ and call it $PPV_z^{'}$. Given Assumption A2, $c_z(v) = F^{-1}_{S(1)|Z}(v) = F^{-1}_{S(1)}(v) = c(v)$. We therefore define our time-dependent and treatment-specific $PPV_z^{'}$ as:

$$PPV_z^{'}(t|v) = P\{T(z) > t | S(1) \geq c(v)\}.$$

This version of $PPV_z^{'}(t|v)$ is interesting and possibly more easily understood than the risk difference based, $PPV(t|c)$ outlined above, as the thresholds are for the candidate surrogate itself and not the risk contrast model; the probabilities are still model based. Via modification of the derivation given in Moskowitz and Pepe (2004a),

$$PPV_z^{'}(t|v) = (1-v)^{-1} \int_{c(v)}^{\infty} S_z(t|s_1) dF_{S(1)}(s_1),$$

where $S_z(t|s_1)$ is $P\{T(z) > t | S(1) = s_1, T(0) > \tau, T(1) > \tau\}$ for $t \geq t_0$. The plug-in estimator for $PPV_z^{'}$ is given by:

$$\widehat{PPV}_z^{'}(t|v) = (1-v)^{-1} \int_{c(v)}^{\infty} \widehat{S}_z(t|s_1) d\widehat{F}_{S(1)}(s_1)\cdot$$

This can be estimated using the estimated survival function and estimated $F_{S(1)}$ distribution from any of the estimation methods outlined in Chapter 6. Although Assumption A2 implies $c_z(v) = c(v)$, the $S(1)$ threshold may not be independent of all baseline covariates. For example the threshold may vary with the BIP, W. One could easily extend these estimands and estimators to include other covariates by applying Zheng et al. (2010) directly. Although this form of estimator is not investigated here, the estimator is given by:

$$\widehat{PPV}_z^{'}(t|v,w) = (1-v)^{-1} \int_{c_w(v)}^{\infty} \widehat{S}_z(t|s_1,w) d\widehat{F}_{S(1)|W}(s_1)\cdot$$

The interest in $PPV_z^{'}$ for specific SoP evaluation is in comparisons over z. Comparisons over z using the $PPV_z^{'}$ are causal, as they are based on the survival function and not the hazard. A meaningful comparison of $PPV_z^{'}$ for potential surrogate evaluation is the VE above a given threshold of $S(1)$ defined by:

$$VE(t|v+) = 1 - RR(t|v+) \equiv 1 - \frac{1 - PPV_1^{'}(t|v)}{1 - PPV_0^{'}(t|v)}.$$

Specific SoP are used as targets in Phase I trials and $VE(t|v+)$ can be used to estimate the VE before a given follow-up time if all vaccinated individuals achieved at least the threshold value of S. Figures depicting $VE(t|v+)$ versus v are of interest for surrogate quality description and optimal threshold determination. Steps in the curve suggest points where thresholds could be considered, as they indicate rapid gains or losses in VE above a given threshold; smooth curves suggest that there is not point at which the VE jumps, thus suggesting that a threshold may not be a useful view of the candidate SoP. Surrogate quality can be tested via a Wald test of the null $VE(t|v_1+) - VE(t|v_2+) = 0$ where $v_1 \neq v_2$.

Finding a threshold of the potential surrogate above which there is high VE is of interest for surrogate evaluation and vaccine improvement in Phase I and IIa trials, but it is not necessary that this threshold be based on the quantiles of the potential surrogate. The $PPV_z^{'}$ can be determined for fixed cutoffs of $S(1)$ by changing the limits of integration from $c(v)$ to the fixed cutoff, c. This version of the $PPV_z^{'}$ may be advantageous for evaluation of potential surrogates with a defined protective level of immune response.

Basing the threshold based on the quantiles of $S(1)$ standardizes the curves for comparison between biomarkers. For example, for a desired level of VE, η, and a candidate surrogate S^j, let $v^{j,\eta}$ be the lowest percentile of S^j such that $VE_j(t|v^{j,X}+) = \eta$. If this is lower than $v^{i,\eta}$ for candidate surrogate S^i it means that more of the population obtains the desired VE as measured by S^j than S^i. This is only one way to look at the curves and comparisons based on the full $VE(t|v+)$ curve of each candidate must always be considered. Lower $\hat{v}^{k,X}$ does not imply that VE has greater variation over the candidate surrogate; lower $\hat{v}^{k,X}$ will often imply a flatter VE curve, making it a worse SoP by our definition.

Although we discuss $VE(t|v+)$ in terms of a fixed timepoint of follow-up, t, this is not required. There may be different optimal thresholds over different times for the same candidate surrogate; a high quality surrogate at 3 years after vaccination may not be a high quality surrogate at 5 years. The desired longevity of surrogate quality should be considered when determining the best surrogate among a group of candidates.

The $PPV_z^{'}$ curve is introduced here via the time-dependent Weibull model, however it is easy to see how a time-independent version of the $PPV_z^{'}$ could be estimated via the Cox model or parametric exponential model. The $PPV_z^{'}$ curve could also be used for binary clinical endpoints by modification

of the PPV of Moskowitz and Pepe (2004b), to fit the surrogate evaluation framework.

Chapter 6

IDENTIFICATION AND ESTIMATION OF TIME-DEPENDENT SOP ESTIMANDS

To estimate the extended SoP estimands, we need a model that incorporates time and a technique to estimate the parameters of that model. We introduce a model that allows for two types of time-dependence based on a continuous time-to-event clinical endpoint and modify three existing estimation methods to accommodate this model.

6.1 Identification and Data Sampling

As outlined in section 4.2.1 Follmann (2006) developed three augmented trial designs: closeout placebo vaccination (CPV), baseline irrelevant variable (BIV), which is a special case of the baseline immunogenicity prediction (BIP) of Gilbert and Hudgens (2008), and their combination (BIP + CPV). We use these same augmented trial designs in the time-to-event setting to identify time-dependent risk estimands of interest. Just as in the binary endpoint setting, these trial designs can be modified to reduce measurement cost with case-control sampling of $S(1)$, S^C and W.

The BIP trial design can be extended to two-phase sampling by leveraging other baseline covariates following the same procedure outlined in subsection 4.3.1 (Gilbert and Hudgens, 2008). Two-phase sampling of the BIP can also be used without the aid of baseline covariates and this is the technique we use in the Zoster vaccine data example. All methods developed in this dissertation can also be extended to some type of two-phase sampling of $S(1)/S^C$. Commonly, the number of uninfected placebo recipients and vaccine recipients that have S^C or $S(1)$ measured is proportional to the number of cases in each arm and W is measured for all subjects.

Under two-phase sampling of $S(1)/S^C$, there are measurements of W taken in placebo recipients who have S^C measured. We introduce a sampling scheme under which all vaccine recipients have the BIP measured, but only those placebo recipients that do not have CPV have W measured. We refer to this sampling as fill sampling of BIP. Under 1:5 case:control sampling of $S(1)$ and S^C there

are far fewer measurements of W taken under fill sampling of the BIP than under full sampling of BIP. Reducing the number of measurements taken, can help reduce trial cost and all subjects missing $S(1)$ still have a BIP measurement to use in imputation.

If W is not used in the risk model fill sampling will give the exact same estimates as full sampling of W for all EML methods given below. However, BIP measurements in all subjects can be useful in a SoP analysis even when they do not contribute to the estimation of the likelihood. When considering fill sampling versus full sampling of the BIP one should also consider the usefulness of the extra W measurements in validating the CPV assumptions or the increase in precision form including W in the risk model and weight that against the cost reduction of full sampling.

Just as in the binary outcome setting, we will need to make some set of assumptions about the data in order to identify the risk estimands. Following Qin et al. (2008), the standard A1-A3 assumptions reduce the number of missing potential outcomes and define the average risks for comparison. We divide the A3 assumption of Qin et al. (2008) into assumptions A3 and A7, separating random censoring and risk prior to potential SoP measurement. We assume the following:

- A1: Stable unit treatment value assumption (SUTVA) and consistency
- A2: Ignorable treatment assignment
- A3: Equal individual clinical risk up to time τ, $T(1) < \tau$ if and only if $T(0) < \tau$
- A7: Censoring is random; $T(z) \perp C(z)$ for $z = \{0, 1\}$.

Assumptions A1 through A3 have been used and explained previously in the literature (Gilbert and Hudgens, 2008; Gilbert et al., 2008; Qin et al., 2008) and above. We denoted Assumption A7 to allow for consistency with the numbering convention followed in most of SoP literature. Assumption A3 is an untestable assumption that can be violated in some trials. We continue to make Assumption A3 here because it is plausible for our motivating examples and it aids in identifiability. Wolfson and Gilbert (2010) and Wolfson (2009) relaxed A3 and it is of future interest to relax A3 for the proposed methods. Assumptions A1 though A3 imply that the conditioning sets for the marginal risk estimand for $T(z)$ given $\{S(1) = s_1, T(1) < \tau, T(0) < \tau\}$, is equivalent to $\{Z =$

$z, S(1) = s_1, T(1) < \tau, T(0) < \tau\}$ if one is assigned $Z = z$. This is helpful for identification of the risk estimands.

6.2 Weibull Model

Risk based on the hazard can be linked to the unknown parameter sets γ, β and the variables Z, $S(1)$, W, and a set of other baseline variables Q, using a Weibull structural risk model. The covariate linked hazard for this Weibull model is given by:

$$\begin{aligned} risk_z^{HZ}(t|s_1, w, q; \gamma, \beta) &= \lambda_z(t|s1, w, q) \\ &= \frac{\exp(\beta_{z0} + \beta_{z1} s_1)}{\exp(\gamma_{z0} + \gamma_{z1} s_1 + \gamma_{z2} w + \gamma_{z3} q)} \\ &\times \left(\frac{t}{\exp(\gamma_{z0} + \gamma_{z1} s_1 + \gamma_{z2} w + \gamma_{z3} q)} \right)^{(\exp(\beta_{z0} + \beta_{z1} s_1) - 1)}. \end{aligned}$$

The survival function, $S_z(t|s_1, w, q)$, can also be linked to the covariates:

$$\begin{aligned} risk_z^{CDF}(t|s_1, w, q; \gamma, \beta) &\equiv 1 - S_z(t|s_1, w, q) \\ &= 1 - \exp\left\{ - \left(\frac{t}{\exp(\gamma_{z0} + \gamma_{z1} s_1 + \gamma_{z2} w + \gamma_{z3} q)} \right)^{\exp(\beta_{z0} + \beta_{z1} s_1)} \right\}. \end{aligned}$$

Although we have linked risk to γ, β via the hazard function and survival function, the same estimates of γ, β could also be linked to the cumulative risk at time t (CDF) or the cumulative hazard. In the time-dependent setting it is more convenient to define the parametric assumption of the risk form, $g(\cdot)$ as given in Gilbert and Hudgens (2008), based on both the survivor and hazard functions. For this reason, the Weibull PDF for right-censored data will serve as our $g(\cdot)$. Let the function $g(\cdot)$ be defined by:

$$g_z(t|s_1, w, q, y; \gamma, \beta) = \lambda_z(t|s_1, w, q)^y \times S_z(t|s_1, w, q) \cdot \tag{6.1}$$

We refer to this model for outcome as our assumption 4 parametric, A4-P as in Gilbert and Hudgens (2008). Assumption A4-P can be tested in trials that perform CPV, without CPV the assumption is untestable. An important consequence of using this Weibull model is that $VE(t|s_1)$ based on the hazard will always be monotonic in t, for any fixed level of s_1. We believe this to be a strength of this model for evaluation of SoP in vaccine trials. However, as pointed out in Chapter 2, the hazard at differing levels of s_1 may change over time differently.

6.2.1 *Identifiability*

Theorem 3. Global identifiability of the Saturated Weibull, Model 6.1
There are no two distinct sets of parameter values $\{\gamma, \beta\}$ such that the distribution of T is the same as defined by the PDF $g_z(t|s_1, w, q, y; \gamma, \beta)$.

Proof. Identifiability of the saturated Weibull model:
Let $W_s(\beta, \gamma)$ denote the saturated Weibull Model 6.1 and $I_w(.)$ denote the Fisher information for the model with respect to $\{\beta, \gamma\}$. For an arbitrary set of points $\theta_0 = \{\beta = \mathbf{B}, \gamma = \mathbf{G}\}$; $\mathrm{I}_w(\theta_0)$ is non-singular: calculations done via Mathematica. Then by Theorem 5 and Corollary 6 of Dasgupta et al. (2007) $W_s(\mathbf{B}, \mathbf{G})$ is locally identifiable at θ_0. Therefore, $W_s(\beta, \gamma)$ is globally identifiable as θ_0 is arbitrary. □

6.2.2 *Testing for Surrogate Value Under the Weibull Model*

Constant Shape Parameter Model

One main test of interest is determining if we should use the more complex model or not. The hypothesis of interest is the null:

$H0_1$: Constant Shape Parameter: equivalent to the null $\{(\beta_{10} - \beta_{00}) = \beta_{01} = (\beta_{11} - \beta_{01}) = 0\}$

If we fail to reject null hypothesis $\mathrm{H0}_1$, we use a less complex model for inference that characterizes the scale parameter with covariates but only includes a constant shape parameter. In testing all but the constant term of the shape for $\mathrm{H0}_1$, the time-dependence of risk associated with $S(1)$ is also tested. Therefore, it is possible that hazard based VE is proportional in time while we reject $\mathrm{H0}_1$. When VE is truly proportional in time but risk is time-dependent in $S(1)$, we found no great efficiency loss nor increase in bias associated with fitting the fully-characterized shape. The time-dependent VE model is still the suggested model for inference in these cases, as inference for $H0_2$ should be based on points along the $VE(t|s_1)$ curve for fixed time. The null hypotheses $H0_3$ and $H0_4$, described below, can be used to determine if the time-dependence is completely driven by the time-dependence of risk in $S(1)$ rather than time-dependence of VE.

The ideal $\mathrm{H0}_1$ would be $VE(t|s_1) = VE(s_1)$, based on the testable null hypothesis $\{(\beta_{10} - \beta_{00}) = (\beta_{11} - \beta_{01}) = 0\}$. However this test would lead us to a less complex model that included a β_{01} term

in shape, thus making the less complex model depend on $S(1)$ in multiple ways. It is not yet known if this model would be unbiased for truly Exponential distributions of time and therefore we stay with the $H0_1$:constant shape parameter test. Selected simulations of the testable null hypothesis $\{(\beta_{10}-\beta_{00}) = (\beta_{11}-\beta_{01}) = 0\}$ were run for power comparison, results not displayed, and found to be almost identical to the power for the joint Wald test of $\{(\beta_{10}-\beta_{00}) = \beta_{01} = (\beta_{11}-\beta_{01}) = 0\}$, further discussion of this can be found in the simulation results section.

Although it is possible that the hazard based risk is also time independent, models allowing for a constant shape parameter did not suffer from great efficiency loss when T was truly distributed exponential. We call this the constant shape parameter Weibull, under which the risk based on the hazard is given by:

$$\begin{aligned} risk^*_z(t|s_1,w,q;\beta^*,\gamma^*) &= \lambda^*_z(t|s1,w,q) \\ &= \frac{\beta^*}{\exp\left(\gamma^*_{z0}+\gamma^*_{z1}s_1+\gamma^*_{z2}w+\gamma^*_{z3}q\right)} \times \left(\frac{t}{\exp\left(\gamma^*_{z0}+\gamma^*_{z1}s_1+\gamma^*_{z2}w+\gamma^*_{z3}q\right)}\right)^{(\beta^*-1)}. \end{aligned}$$

Although the hazard based risks may be time-dependent, the hazard based VE is not. The parameter β^* has the same value for either treatment. It is easy to see if one takes the log of $risk^*_z(t|s_1;\beta^*,\gamma^*)$,

$$\begin{aligned} \log(risk^*_z(t|s_1,w,q;\beta^*,\gamma^*)) &= \log(\beta^*)-\gamma^*_{z0}-\gamma^*_{z1}s_1-\gamma^*_{z2}w-\gamma^*_{z3}q \\ &+ \log(t)+\beta^*\log(t)-\beta^*(\gamma^*_{z0}+\gamma^*_{z1}s_1+\gamma^*_{z2}w+\gamma^*_{z3}q) \end{aligned}$$

making log of VE(s_1),

$$\begin{aligned} \log(VE^*(s_1)) &= -\log(\beta^*)+\gamma^*_{10}+\gamma^*_{11}s_1+\gamma^*_{12}w+\gamma^*_{13}q-\log(t)-\beta^*\log(t)+\beta^*(\gamma^*_{10}+\gamma^*_{11}s_1+\gamma^*_{12}w+\gamma^*_{13}q) \\ &+ \log(\beta^*)-\gamma^*_{00}-\gamma^*_{01}s_1-\gamma^*_{02}w-\gamma^*_{03}q+\log(t)+\beta^*\log(t)-\beta^*(\gamma^*_{00}+\gamma^*_{01}s_1+\gamma^*_{02}w+\gamma^*_{03}q) \\ &= (1+\beta^*)\times\{\gamma^*_{10}-\gamma^*_{00}+(\gamma^*_{11}-\gamma^*_{01})*s_1+(\gamma^*_{12}-\gamma^*_{02})*w+(\gamma^*_{13}-\gamma^*_{03})*q\}\cdot \end{aligned}$$

VE is time-independent, but not free of β^*. The $g^*(.)$ link function for outcome, we will call A4-P-null, is the constant shape parameter Weibull PDF allowing for right-censoring:

$$g^*_z(t|s_1,w,q,y;\gamma^*,\beta^*) = \lambda^*_z(t|s1,w,q)^y \times S^*_z(t|s1,w,q),\cdot \tag{6.2}$$

The survival function is given by $S_z(t|s_1) = \exp\left\{-\left(\frac{t}{\exp(\gamma^*_{z0}+\gamma^*_{z1}s_1+\gamma^*_{z2}w+\gamma^*_{z3}q)}\right)^{\beta^*}\right\}$.

Under the constant shape parameter Weibull, the scale coefficients have clear interpretation. The coefficient contrasts of interest are $\gamma^*_{10} - \gamma^*_{00}$ and $\gamma^*_{11} - \gamma^*_{01}$. These can be interpreted as the VE unassociated with the vaccine effect on the potential SoP and the VE that is associated with the vaccine effect on the SoP, respectively. Inference on surrogate quality can be acquired via Wald test based on the null $\gamma^*_{11} - \gamma^*_{01} = 0$ in this setting, as the interpretation is not complicated by time-dependence. However, tests can also be based on comparisons of points along the $VE(s_1)$ curve. We will refer to all null hypotheses of VE being independent of $S(1)$ as $H0_2$ and in the time-independent setting this can be written as:

$$H0_2 : VE(s_1) = VE.$$

Saturated Weibull Model

If the data support rejection of null hypothesis $H0_1$, we must parametrize the shape in order to determine what is causing the time dependence of VE. Thus we arrive at the saturated model described in detail in section 6.2. In this setting, analytical tests for surrogate value become more complicated, and there is more than one way to look at the coefficients. As the model depends on S(1) and Z in multiple ways, tests based on the null $\gamma_{11} - \gamma_{01} = 0$ are no longer useful in evaluating surrogate value. The most comprehensive way to evaluate surrogate value in this setting is by plotting the estimated $VE(t|s_1)$ for the range of s_1 values for several different timepoints of interest. As well, one can plot VE(s_1, t) for a range of timepoints, t, $t > \tau$ and less than the longest follow-up time, for several different levels of s_1. This will be a clear visual indication of the surrogate value of S.

The null hypothesis $H0_2$ in the time-dependent setting is denoted by:

$$H0_2 : VE(t|s_1) = VE(t).$$

Inference for this null can be based on Wald tests of two different points on the $VE(t|s_1)$ being the same, for some fixed time t. This can be tested via the null, $VE(t|s_{1,i}) - VE(t|s_{1,j}) = 0$ for a fixed time t and $s_{1,i} \neq s_{1,j}$. Larger $s_{1,i}$ and small $s_{1,j}$ relative to the range of S(1) are good starting points for this test, however any two points along the $VE(t|s_1)$ that differ significantly for fixed t provide evidence of SoP value. We also considered a joint Wald test of the null $\{(\beta_{11} - \beta_{01}) = (\gamma_{11} - \gamma_{01}) = 0\}$ as this would also be a testable null for $H0_2$ in the time-dependent setting. We found that this test was not more powerful in a selected set of the simulation scenarios than was the suggested testable

null $VE(t|s_{1,i}) - VE(t|s_{1,j}) = 0$; data not shown. Thus, we continue to suggest the testable null $VE(t|s_{1,i}) - VE(t|s_{1,j}) = 0$, particularly when EML estimation is used, Section 6.3. There may be reason to use this the testable null hypothesis, $\{(\beta_{11}-\beta_{01}) = (\gamma_{11}-\gamma_{01}) = 0\}$, when using pseudoscore estimation, discussion below.

Inference on SoP value can also be based on the confidence intervals of the $PPV(t|v)$ and $NPV(t|v)$ estimates, as outlined in Chapter 5, or on a comparison of points along the $VE(t|v+)$ curve. This comparison can be tested via the null, $VE(t|v_1+) - VE(t|v_2+) = 0$ for a fixed time t and $v_1 \neq v_2$.

The novel parameterization of the Weibull given above allows us not only to test the null hypothesis $H0_1$, but also to investigate what is driving the time dependence of VE. The interaction term of the shape parameter for vaccine and the potential surrogate, $(\beta_{11} - \beta_{01})$, is of interest to determine if the time-variation in VE differs over the levels of the surrogate. We call this time-variation in the quality of the surrogate, as the surrogate becomes a worse predictor of VE over time when this term is greater than zero when there is a positive association between treatment effects on the surrogate and treatment effects on the outcome and treatment has a positive effect on outcome. This can be evaluated via a Wald test based on the null $\beta_{11} - \beta_{01} = 0$ and we denote this as null hypothesis $H0_3$. One may also be interested in testing for time-dependent VE that is unassociated with the potential surrogate; this can be evaluated via a Wald test based on the null, $\beta_{10} - \beta_{00} = 0$. We denote this test as null hypothesis $H0_4$. When $\beta_{10} - \beta_{00} > 0$ and VE is positive, this implies there is waning in VE over time. Due to the complicated meaning of the parameters, it is always best to consult the figures prior to attempting to determine what a significantly time-varying effect means in your data.

6.2.3 Procedure Using the Weibull model for SoP Evaluation

- Step 1: Fit the saturated Weibull, Model 6.1.

- Step 2: Test $H0_1$: Shape parameter constant, $\{(\beta_{10} - \beta_{00}) = \beta_{01} = (\beta_{11} - \beta_{01}) = 0\}$

- Step 3: If the data support null hypothesis $H0_1$, fit the constant shape parameter Weibull, Model 6.2. Use estimates from this model for figures and inference on surrogate quality, $H0_2$:$VE(s_1) = VE$. If the data support rejection of $H0_1$ use the saturated model estimates for

figures and inference on surrogate quality, $H0_2$:$VE(t|s_1) = VE(t)$.

In order to test these hypotheses and compare potential specific SoP, we need a way to estimate the coefficients from the Weibull models. As outlined in Chapter 4, there have been many suggested estimation methods for the risk estimands in the binary outcome setting. We propose extensions to three of these methods of estimation to accommodate the Weibull model.

6.3 Parametric EML

We propose to extend the parametric estimation method of Follmann (2006) and Gilbert and Hudgens (2008) to accommodate the Weibull model. Just as in Follmann (2006), we use the EML of Pepe and Fleming (1991) in order to account for the missing potential surrogate data. Using assumed parametric model, Model 6.1, the observed likelihood is given by:

$$\mathrm{L}(\beta, \gamma, \nu) \equiv \prod_i f(T_i|Z_i, S_i(1), W_i, Q_i, Y_i, \delta_i; \gamma, \beta) \quad \text{where} \tag{6.3}$$

$f(T_i|Z_i, S_i(1), W_i, Q_i, Y_i, \delta_i; \gamma\beta)$ is defined as:

$$\begin{aligned} f(T|Z, S(1), W, Q, Y, \delta; \gamma, \beta) &= \{g_z(t|s_1, w, q, y; \gamma, \beta,)\}^{\delta} \\ &\times \left\{ \int g_z(t|s, w, q, y, \gamma, \beta) dF_{S(1)|W}(s) \right\}^{(1-\delta)}. \end{aligned}$$

For EML, we estimate $F_{S(1)|W}$ independently of γ, β, and treat it as a nuisance parameter ν. For the fully-parametric EML we assume a parametric form of $S(1)$ and W, which induces a conditional distribution $S(1)|W$. For example if we assume $S(1)$ and W are bivariate normal, the distribution of $S(1)|W$ is given by:

$$F_{S(1)|W}(s) = \frac{1}{(\sqrt{(}2\pi(1-\rho^2_{SW})\sigma^2_S)} \; e^{-\left(\frac{(s-(\mu_S + (\frac{\sigma_S}{\sigma_W}) * \rho_S W(w-\mu_W)))^2}{(2*\sigma^2_S(1-\rho^2_{SW}))} \right)}.$$

In Subsection 6.3.1 we outline alternative plausible parametric models for $S(1)|W$. The parametric model should be selected based on the trial data and can be tailored to the particular type of $S(1)$ and W data observed (Gilbert and Hudgens, 2008). Continuing with our bivariate normal example, we need a way to estimate the unknown moments of $S(1)$ and W. As suggested in Follmann (2006) we fit the model to the data via maximum likelihood in the vaccine recipients that have both $S(1)$

and W. We restrict to this group for fitting the model because it is either unbiased, in the case of full sampling of BIP and S(1), or can be weighted to account for the bias sampling Breslow and Wellner (2007), in the case of case-control sampling of S(1).

Using a consistent estimate of $F_{S(1)|W}$ and plugging back into 6.3 we can estimate the coefficients of interest. As given in Pepe and Fleming (1991), any consistent estimate of $F_{S(1)|W}$ will give consistent estimates of coefficients of interest via optimization of 6.3 given that the observed likelihood is identified.

Theorem 4. Identifiability of the observed Likelihood under [CB], [A1] through [A7], CPV or BIP If CB, Assumptions A1 through A7 hold and the trial is augmented via CPV, BIP or both, then both the marginal and joint risk estimands are identifiable given the observable data.

Proof. Following the proof in Gilbert and Hudgens (2008), if CB and assumptions A1 through A3 plus A7 hold and a BIP exists and is measured, marginal risk for vaccine recipients $risk_1(t|s_1)$ is identified and equivalent to the joint risk given the observed data, provided there are observed events in the vaccine group as mentioned in Cox (1972) and complete data likelihood is identified, Theorem 3. Similarly, $risk_0(t|s_1)$ is identified provided there are observed events in the placebo group. By following a similar proof of Proposition 3 of Wolfson (2009) the risks are also identified if CPV is performed. □

To illustrate the method more concretely, we continue with our example assuming our trial has BIP+CPV full sampling. Then only infected placebo recipients, group denoted P_1, do not have a direct measurement of either $S(1)$ or S^C and their likelihood contribution L_{01} is given by:

$$L_{01}(\beta,\gamma,\hat{\nu}|t,w,q,y) \quad = \quad \prod_{i\epsilon P_1}\int_s \exp\left\{-\left(\frac{t_i}{\exp(\gamma_{00}+\gamma_{01}*s+\gamma_{02}w_i+\gamma_{03}q_i)}\right)^{\exp(\beta_{00}+\beta_{01}*s)}\right\} d\hat{F}_{S(1)|W=w}(s)\cdot$$

The estimated likelihood for this example given by:

$$\mathrm{L}(\beta,\gamma,\hat{\nu}) = \prod_{i\epsilon\overline{P_1}} g_z(t|s_1,w,q,y,\gamma\beta,)\times L_{01}(\beta,\gamma|t,w,q,y)\cdot$$

The general likelihood for possible sub-sampling of S^C, $S(1)$ is given by:

$$\mathrm{L}(\beta,\gamma,\hat{\nu}) = \prod_i g_z(t_i|s_{i,1},w_i,q_i,y_i;\gamma\beta,)^{\delta_i}\times\prod_i\int g_z(t|s,w,q,y,\gamma,\beta)dF_{S(1)|W}(s)^{(1-\delta_i)}.$$

Due to the zero probability of observing $S(1)$ in the infected placebo recipients, there is not an expression for the asymptotic distribution of the estimated β and γ sets. For this reason, bootstrap is the suggested mode of inference. The likelihood does not have a closed form and numerical integration is necessary for optimization. Due to the computational intensity of the procedure one may wish to run fewer bootstrap samples than would be advisable for forming bootstrap SE. Inverted bootstrap CI tests may be desirable over bootstrap SE. Inverted bootstrap CI tests consist of ordering the bootstrap statistics and then using the percentiles of the tail values to test the null at the correct level. An example of this would be running fifty bootstrap samples, ordering the statistics of interest, then rejecting the null hypothesis that an estimand is equal to zero for the alternative that it is greater than zero if the mean of the second and third smallest statistics is greater than zero.

The method is easily extended to two-phase sampling of CPV/$S(1)$. The distribution of $S(1)|W$ can be estimated using weighted likelihood to account for the biased sampling using inverse probability weights, based on the probability of a particular vaccinee being selected to the validation sample. This is demonstrated in the simulations. Two-phase sampling of W with or without the aid of additional baseline covariates can also be accommodated, although this is not demonstrated in the main simulations. One method for accounting for sub-sampling of W is outlined in Gilbert and Hudgens (2008), it leverages other baseline covariates measured on all subjects. The Zoster real-data example has a sub-sampled BIP and no CPV, and in the fully parametric EML analysis of that data we estimate the likelihood for subjects missing both the BIP and $S(1)$ by integrating over $F_{S(1)}$ rather than $F_{S(1)|W}$. Table 6.1 describes the various sampling scenarios that could be used and the way the method would need to be modified.

6.3.1 Modeling of S(1)

There are many different parametric models that can be used for $S(1)$ and W, and each will allow for different data characteristics. One can allow for censored S(1) values by assuming that $S(1)|W$ is censored normal, with left censoring of values below some constant, ζ. An example of where this may be useful is in Luminex immune assays. The limit of detection for the Luminex assay is the line below which the measurements are outside the range of plausible positive response values. In CB, all $S(0)$ will be equal to the limit of detection, ζ, and $S(1) < \zeta$ will be considered unreliable and

Table 6.1: Sampling scenarios in SoP evaluation for EML methods

BIP	CPV	S(1)	$\hat{L}(\beta, \gamma)$ changes
Full	Full	Full	Original Scenario
Fill sampling	Full	Full	No change
Fill or Full	Two-phase	Two-phase	IPW used in $F_{S(1)\|W}$ estimation*
Two-phase	Full,Two-phase or None**	Full or Two-phase	IPW used in $F_{S(1)\|W}$ estimation* when no CPV or BIP for a subject use the double integral method*** or integrate over the unconditional distribution of $S(1)$ estimated with IPW
None	Full or Two-phase	Full or Two-phase	IPW used in $F_{S(1)\|W}$ estimation* when no CPV integrate over the unconditional distribution of $S(1)$ estimated with IPW

$*$ follow Gilbert et al. (2011b) subsection 4.3.1

$**$ in the case of BIP only, no CPV, interactions of both the BIP, W, and Z and the other baseline covariates, Q, and Z cannot be modeled at the same time.

$***$ follow Gilbert and Hudgens (2008) subsection 4.3.1

set to ζ. Modeling $S(1)|W$ as censored normal accounts for this truncation, as was done in Gilbert and Hudgens (2008).

There are assays where the $S(1)$ distribution will have a high density of zeros. An example of a biomarker measurement that follows this pattern is the breadth of immune response to vaccine insert. For breadth, subjects will have some count of reactions, zero to the total number of tested peptides. One could accommodate the distribution of $S(1)$ in this case by assuming $S(1)|W$ is distributed zero-inflated Poisson.

These are only two examples of plausible distribution types of potential surrogates. As numerical integration is most likely necessary in the proposed method regardless of the $S(1)|W$ distribution choice, the decision of how to model $S(1)|W$ should be based on prior knowledge of the biomarker or on investigation of the current data. The proposed method can be modified to allow for non-parametric and semi-parametric estimation of $S(1)|W$. We explore this further in the extension of Huang and Gilbert (2011) in section 6.4.

6.4 Semi-parametric EML

We propose to extend the semi-parametric estimation method of Huang and Gilbert (2011) to a continuous time-to-event clinical endpoint. We again start from the observed likelihood 6.3. Just as in Huang and Gilbert (2011), rather than assuming a parametric form of $F_{S(1)|W}$, we assume the semi-parametric location-scale model of Heagerty and Pepe (1999) for S(1) given W. We use the residuals from fitting this model in the vaccine recipients with both S(1) and W measured to estimate $F_{S(1)|W}$ via:

$$F_{S(1)|W} \sim F[\{s_1 - \mu(w)\}/\sigma(w)] = F(\varsigma),$$

where F is the CDF of univariate residual ς and $\mu(w)$ and $\sigma(w)$ are the location and scale parameters, respectively. Just as in Huang and Gilbert (2011) we use the residuals to estimate the conditional distribution of $S(1)|W$, given by:

$$F_{S(1)|W} = P(S(1) \leq s_1|W) = P\left\{\varsigma \leq \frac{s_1 - \mu(w)}{\sigma(w)}\right\}.$$

Therefore, $F_{S(1)|W}$ can be estimated in the validation sample via the location-scale parameters $\mu(w)$ and $\sigma(w)$. Huang and Gilbert (2011) show that by assuming μ and σ are parametric forms of W,

$\mu(w) = \gamma' w$ and $log\{\sigma(w)\} = \eta' w$ one can estimate them by solving the equations:

$$\sum_{k=1}^{n_V} \frac{w_k(s_{(1,k)} - \gamma' w_k)}{\sigma(w_k)^2} = 0$$

and

$$\sum_{k=1}^{n_V} \frac{w_k\{(s_{(1,k)} - \gamma' w_k)^2 - \sigma(w_k)^2\}}{\sigma(w_k)^2} = 0\cdot$$

Where n_V is the set of validation subjects, these are vaccine recipients with both $S(1)$ and W measured. This yields a series of residuals, ς_k where k refers to the kth member of the validation sample. Just as in Huang and Gilbert (2011), we use the estimates γ' and η' to impute missing S(1) values:

$$S^*_{i,k}(1) = \hat{\gamma'} w_i + \exp(\hat{\eta}' w_i)\varsigma_k$$

There are k total imputations used for each missing $S(1)$ value. We can then use these imputed values to estimate $\left\{\int g_z(t_i|s, w_i, q_i, y_i, \gamma, \beta) dF_{S(1)|W}(s)\right\}$ by the empirical integral:

$$\left(\frac{1}{n_V}\right)\sum_{k}^{n_V} g_z(t_i|S^*_{i,k}(1), w_i, q_i, y_i, \gamma, \beta)\cdot$$

This gives us a general estimated log likelihood of:

$$\mathrm{l}(\beta, \gamma, \hat{\nu}) = \sum_i log(g_z(t_i|s_{i,1}, w_i, q_i, y_i; \gamma\beta,))*\delta_i + \sum_i \left\{\left(\frac{1}{n_V}\right)\sum_{k}^{n_V} log(g_z(t_i|S^*_{i,k}(1), w_i, q_i, y_i, \gamma, \beta))\right\}*(\delta_i - 1)\cdot$$

Unlike Huang and Gilbert (2011), we do not use the EM-algorithm to solve the estimated likelihood. We optimize the estimated likelihood directly. This makes the semi-parametric method directly comparable in algorithm to the fully parametric method above. We are merely replacing the parametric form of $S(1)|W$ with a location-scale form and the numeric integral with an empirical integral over those distributions, respectively.

We investigate the properties of this method in simulations and real data example from the Step trial. We do not investigate the extension to two-phase sampling of the BIP, W, in the main simulations; we consider an adaptation of the method allowing for two-phase sampling of the BIP in the real data example for the ZEST trial, Chapter 7. We do consider case-control sampled $S^C/S(1)$ data in the simulations. To account for the bias sampling of $S(1)$, when case:control sampling is used in the vaccine recipients, Huang and Gilbert (2011) suggests the use of inverse sampling probabilities in both the fitting of the location-scale model and in the EM algorithm.

As we are not using the EM algorithm, we apply the weights in fitting of the location-scale model and the fitting of the imputed likelihood contributions. Those with $S(1)$ measured have their likelihood contribution unweighted; each imputed likelihood contribution is weighted by the inverse probability of measurement for the $S(1)$ giving rise to that imputation. Thus, the likelihood contribution for subjects missing $S(1)$ is an inverse probability weighted average of their imputed likelihood contributions.

This method is easily applied to multiple biomarkers even in the time-to-event clinical endpoint framework, by fitting a location-scale model to each biomarker separately and using the residuals to estimate their joint distribution, $F^{S_1(1),...,S_j(1)|W}$. Although we do not investigate multiple biomarker combinations in this dissertation, it is a future goal. It was brought to our attention by Dr. Huang that one may need to use the EM when multiple biomarkers are being investigated as the complexity of the model may cause direct optimization to fail.

6.5 Pseudoscore Method

As was first pointed out in Wolfson (2009), the pseudoscore method of Chatterjee et al. (2003) can be used for SoP evaluation. The main appeal being that it has a closed form variance under vaccine trial conditions, infected placebo recipients having zero probability of having $S(1)$ or S^C. Huang et al. (2012a) point out that the pseudoscore estimation also eliminates the CPV paradox, whereby CPV sub-sampling actually decreases power over BIP alone. The CPV paradox first found in (Gilbert et al., 2011b), can be seen in the results from the EML estimation methods. Elimination of the paradox is useful for SoP evaluation, as an appropriate BIP may not always exist and when utilized appropriately CPV can improve efficiency. CPV is also one way to test the outcome modeling assumption A4-P. Huang et al. (2012a) also finds the pseudoscore method to have increased efficiency in comparison to the EML methods.

Pseudoscore estimation in the time-dependent setting can be used for SoP evaluation with only slight modification from the method described in Chapter 2 for CoR evaluation. First, the assumptions of Chatterjee et al. (2003), as pointed out in Wolfson (2009), require that CPV be performed for at least some sub-sample of the placebo group. Without CPV one of the main assumptions of Chatterjee et al. (2003), Assumption A, would be violated. We restate these assumptions here for

clarity in terms of an SoP, $S(1)$.

- Ps1: $\int_t \phi(t, Z, W)dt > 0$ for all ϕ in the neighborhood of the true ϕ_0;
  where $\phi(t, Z, W) = P(\delta = 1|T = t, Z = z, W = W)$. This states that there is positive expected value of being in the second phase sampling with respect to outcome and baseline variables Z and W.

- Ps2: $g_z(t|s_1, z, w, y, \beta, \gamma) > 0$ for almost all observed data in the neighborhood of the true β_0 and γ_0; strictly positive value given the assumption of the parametric model for outcome T.

- Ps3: $P(\delta = 1|T, S(1), Z, W) = P(\delta = 1|T, Z, W) = \phi(T, Z, W)$, meaning $S(1)$ is missing at random, (MAR).

We break-out the baseline variable of treatment randomization, Z, for clarity, as Z could not have been include in the W vector in the CoR evaluation case as CoR are only investigated in vaccine recipients. However, this is not an adaptation of CoR evaluation PS assumptions given above, Z is just an additional variable in the W vector.

Assumption Ps3 will most likely hold as sampling should only depend on $S(1)$ via outcome, the ability to measure $S(1)$, $X > \tau$, and the observed baseline measurements, W. Assumption A, as mentioned above, can only be true in the presence of CPV implying the need for assumptions A5 and A6. Assumption Ps2 relies not only on A4-P being the correct specification of model, but also on there being positive risk at all times. In the time-dependent setting this implies that time must not exceed the maximum true event time. Given these assumptions along with assumptions A1 through A7, we again consider the observed likelihood 6.3. Unlike the other methods where we attempt to estimate $F_{S(1)|W}$ unbiasedly and consistently in a validation sample, following Chatterjee et al. (2003) we use all the $S(1)$ or S^C measurements and attempt to account for the biased sampling.

We consider the estimation of $F_{S(1)|W}$ via the biased sample as did Wolfson (2009). Using the observed $F(S(1)|Z, W, \delta = 1)$ which we will denote similarly to Chatterjee et al. (2003) as $F^*(S(1)|Z, W)$, let

$$q_z^{\phi}(S(1), W\beta, \gamma) \equiv P(\delta|S(1), Z = z, W) = \int \phi(t, z, W) g_z(t|S(1), W, Y, \beta, \gamma)dt.$$

Assumptions Ps1 and Ps2 above ensure that $q_z^{\phi}(S(1), X, W, \beta, \gamma) > 0$ almost surely. Using this, we can define $F(S(1)|Z, W)$ from the observable data by:

$$F(S(1)|Z,W) = \frac{P(S(1) \leq s_1|W,Z,\delta=1)P(\delta=1|Z,W)}{P(\delta|S(1)=s_1,Z,W)} \equiv \frac{F^*(S(1)|Z,W)P(\delta=1|Z,W)}{P(\delta|S(1)=s_1,Z,W)}.$$

Taking the score of the observed likelihood we have:

$$\begin{aligned} S_z(\beta,\gamma;F^*,\phi) &= \frac{\partial logL(\beta,\gamma;F)}{\partial(\beta,\gamma)} = \sum_{i\epsilon v} S_{\beta,\gamma}(T_i|S(1)_i,z_i,W_i,Y_i) \\ &+ \sum_{i\epsilon\bar{v}} \frac{\int S_{\beta,\gamma}(T_i|s_1,z_i,W_i,Y_i)h_z^{\phi}(T_i|s_1,W_i,Y_i,\beta,\gamma)dF^*(s_1|z,W)}{\int h_z^{\phi}(T_i|s_1,W_i,Y_i,\beta,\gamma)dF^*(s_1|z,W)}, \end{aligned}$$

where

$$h_z^{\phi}(t|s_1,z,w,y,\beta,\gamma) = \frac{g_z(t|s_1,w,y,\beta,\gamma)}{q_z^{\phi}(s_1,w,\beta,\gamma)}.$$

We denote the piece of the score for those with $S(1)/S^C$ measured as $S_{\beta,\gamma}(T_i|s(1)_i,Z_i,W_i,Y_i)$, and the piece of the score for those missing $S(1)$ as $S_{\beta,\gamma,F^*}(T_i|Z_i,W_i,Y_i)$ given by:

$$S_{\beta,\gamma,F^*}(T_i|Z_i,W_i,Y_i) \equiv \frac{\int S_{\beta,\gamma}(T_i|s_1,Z_i,W_i,Y_i))h_z^{\phi}(T_i|s_1,W_i,Y_i;\beta,\gamma)dF^*(s_1|z,W)}{\int h_z^{\phi}(T_i|s_1,W_i,Y_i;\beta,\gamma)dF^*(x|z,W)}.$$

Using the empirical estimate of $F^*(S(1)|Z,W)$,

$$F_N(s_1|z,w) = \frac{\sum_i I_{[S(1)\leq s_1,Z=z,W=w,\delta=1]}}{\sum_i I_{[W=w,Z=z,\delta=1]}},$$

and plugging into the score S_{β,γ,F^*}, one can write the pseudoscore estimating equation as:

$$\begin{aligned} S_{Ps}(\beta,\gamma;F_N,\phi) &= \sum_{i\epsilon v} S_{\beta,\gamma}(T_i|S(1)_i,Z_i,W_i,Y_i) \\ &+ \sum_{j\epsilon\bar{v}}\sum_{i\epsilon v} \frac{S_{\beta.\gamma}(T_j|S(1)_i,Z_j,W_j,Y_j)h_{z_j}^{\phi}(T_j|S(1)_i,W_j,Y_j,\beta,\gamma)I_{[z_j=Z_i,W_j=W_i]}}{\sum_{l\epsilon v} h_{z_j}^{\phi}(T_j|S(1)_l,W_j,Y_j,\beta,\gamma)I_{[z_j=Z_l,W_j=W_l]}}. \end{aligned}$$

These estimating equations can be solved via Newton-Raphson algorithm, arriving at the pseudoscore estimates β^{Ps}, γ^{Ps}. Again, the iterative reweighting algorithm could instead be used. As the algorithm is given above in Chapter 2, Section 2.2.3, List 2.2.3 we will not repeat it here. The weights in this case are given by:

$$w_{ij}(\beta^c,\gamma^c) = \frac{h_{z_j}^{\hat{\phi}}(T_j|S(1)_i,W_j,Y_j,\beta^c,\gamma^c)}{\sum_{l\epsilon v_{z_j,w_j}} h_{z_j}^{\hat{\phi}}(T_j|s(1)_l,W_j,Y_j,\beta^c,\gamma^c)}.$$

Just as in the CoR analysis case, no software implements a glm for our Weibull model; and there are again at least two ways to fit our model in Step 5. We decide upon the iteratively reweighted solutions to the score, just as in the CoR pseudoscore solutions, and leave the investigation of the use of an equivalent linear-transformation model to future research.

Calculation of $h_z^{\phi}(t|s_1, z, w, y, \beta, \gamma)$ can again be considered to be under case-control sampling. As $S(1)/S^C$ sub-sampling will not depend directly on the time-to-event, but rather on whether or not an event was observed prior to the close of the trial. As defined above, let Y to be the indicator that an event was observed before the close of the trial.

We can partition $h_z^{\phi}(t|s_1, w, y, \beta, \gamma)$ by levels of T defined by $Y(c)$. Given $\delta \perp T|Y$ we have,

$$\begin{aligned} q_z^{\phi}(S(1), W, \beta, \gamma) &\equiv \int \phi(t, Z, W) g_z(t|S(1), W, Y, \beta, \gamma) dt \\ &= \phi(Y(c) = 0, Z, W)(1 - G_z(c|S(1), W, Y, \beta, \gamma)) + \phi(Y = 1, Z, W) G_z(c|S(1), W, Y, \beta, \gamma) \\ &= \phi(0, Z, W)(1 - G_z(c|S(1), W, Y, \beta, \gamma)) + \phi(1, Z, W) G_z(c|S(1), W, Y, \beta, \gamma) \end{aligned}$$

where $G_z(c|X, W, Y, \beta, \gamma)$ is the parametric distribution function of T given $S(1)$, W for z arm subjects evaluated at the administrative censoring time c. This removes the need for numeric integration at each iteration.

By Theorem 2 and Theorem 3 and using the same argument given for validation of the conditions of Theorem 2 in the CoR evaluation method based on the Weibull, it can be stated:

- a. The pseudoscore estimating equations $S_{Ps}(\beta, \gamma; F_N, \phi) = 0$ have a unique, consistent sequence of solutions, $\{\hat{\theta}_N^{Ps}\}_{N \geq 1}$

- b.
$$\sqrt{N}(\hat{\theta}_N^{Ps} - \theta_0) = -\Psi_{\theta}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} g^0(T_i|S(1)_i, W_i, Y_i, Z_i, \delta_i) + o_p(1);$$
where $g^0(T|S(1), W, Y, Z, \delta) = \delta\{S_{0,\beta_0,\gamma_0}(t|s_1, z, w, y) + a(s_1, z, w)\} + (1 - \delta) S_{0,\beta_0,\gamma_0;F_0}(t|z, w, y)$ and subscript 0 denotes that both the model and the parameters in the model are the truth

- c. If $var_0 g^0(T|S(1), W, Y, Z, \delta) < \infty$, then $\sqrt{N}(\hat{\theta}_N^{Ps} - \theta_0) \rightarrow_d N(0, \Omega)$, where Ω is defined by

the sandwich formula,

$$\Omega = [\Psi_\theta(\theta_0, F_0^*)]^{-1} var_0 g^0(T|S(1), W, Y, Z, \delta)[\Psi_\theta^t(\theta_0, F_0^*)]^{-1}.$$

We again use the alternative form of $a(t, w)$ in the estimation of the variance. As given in Chapter 2 this form is,

$$a(s_1, z_k, w_k) = E^0_{t|s_1,w,z,\delta=1}\left\{\frac{1-\phi_0(t, z_k, w_k)}{\phi_0(t, z_k, w_k)}\left[S_{0,\beta_0,\gamma_0}(t|s_1, z_k, w_k, y) - S_{0,\beta_0,\gamma_0;F_{k,0}}(t|z_k, w_k, y)\right]\right\} I_{[z=z_k,w=w_k]}.$$

Where K is the number of levels of W cross Z with w_k and z_k being the values at the kth level $v_k = \{j : \delta = 1, w_j = w_k, z_j = z_k\}$, $\overline{v_k} = \{j : \delta = 0, w_j = w_k, z_j = z_k\}$ and $F_{k,0}$ is the kth conditional distribution of $X|W = w_k, Z = z_k$. The alternative form of $a(s_1, z_k, w_k)$ can be estimated by:

$$\hat{a}(s_1, z, w) = \sum_{i\epsilon\overline{v_k}} \frac{g_{z_k}(T_j|x, w_k, Y_j; \hat{\beta}, \hat{\gamma})}{\sum_{i\epsilon v_k} g_{z_k}(T_j|X_i, w_k, Y_j; \hat{\beta}, \hat{\gamma})}\left[S_{\hat{\beta},\hat{\gamma}}(T_j|s_1, z_k, w_k, Y_j) - S_{\hat{\beta},\hat{\gamma};\hat{F}}(T_j|z_k, w_k, Y_j)\right] I_{[z=z_k,w=w_k]}.$$

Then we can estimate $Var_0 g^0(T, X, W, Y, Z, \delta)$ by:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\delta_i\left\{S_{\hat{\beta},\hat{\gamma}}(T_i|S(1)_i, Z_i, W_i, Y_i) + \hat{a}(T_i, Z_i, W_i)\right\} + (1-\delta_i)S_{\hat{\beta},\hat{\gamma};\hat{F}}(T_i|Z_i, W_i, Y_i)\right]^{\otimes 2}.$$

The estimator for Ω is given by:

$$\Omega = [\Psi_\theta(\hat{\theta}, \hat{F}^*)]^{-1}\widehat{Var}(g(T, X, W, Y, Z, \delta))[\Psi_\theta^t(\hat{\theta}, \hat{F}^*)]^{-1}.$$

As is suggested in Section 6.2.2, when there is time-dependence in VE the tests for null hypothesis $H0_2 : VE(t|s_1) = VE(t)$ are best tested via comparison of points on the VE curve that differ in $S(1)$ value. Therefore, in order to test the null $VE(t|s_{1,k}) - VE(t|s_{1,}j) = 0$ for $s_{1,k} \neq s_{1,}j$, we need to determine the form of the variance of points on the VE curve. Let VE_{θ_0}, VE'_{θ_0} be VE and the derivative of VE with respect to $\theta = \{\beta, \gamma\}$, evaluated at θ_0, for particular fixed time, t and value of the potential SoP $S(1)$; then by the Delta method and Theorem 2, we can state that:

$$\sqrt{N}(VE_{\hat{\theta}_N^{Ps}} - VE_{\theta_0}) \rightarrow_d N(0, \Omega[VE'_{\theta_0}]^2).$$

Thus, $VE(t|s_{1,k}) - VE(t|s_{1,j})$ has asymptotic variance given by:

$$\Omega[VE'_{\theta_0}(t|s_{1,k})]^2 + \Omega[VE'_{\theta_0}(t|s_{1,j})]^2 - 2\Omega[VE'_{\theta_0}(t|s_{1,j})VE'_{\theta_0}(t|s_{1,k})]\cdot$$

One could also use the functional Delta method to determine the form of the variance for the infinite dimensional VE curve. This, along with determining the form of the variance for the of the various summary statistics suggested above, is of future research interest.

We investigate the performance of the pseudoscore estimator in the simulations using Monte Carlo standard error. Although we have proven the asymptotic properties of the pseudoscore estimates, we do not implement them here due to time restrictions and the desire to provide an analysis of the recently acquired ZEST data. Since CPV is required for pseudoscore estimation, there are not yet any real data examples with which to illustrate the pseudoscore method. CPV is a planned augmentation of the proposed trial of Gilbert et al. (2011b).

The pseudoscore method easily extends to two-phase sampling of $S^C/S(1)$ provided S^C is sampled for at least some subjects. The concept of fill sampling or any sub-sampling of the baseline variables is not possible with pseudoscore method. As such, the pseudoscore method, although more efficient, may not be as useful in practice as the EML methods.

6.6 Simulation Setting and Results

We consider simulation studies to investigate bias, type 1 error and power for our method under different surrogate quality levels and different sampling scenarios. We follow the novel trial design outlined in Gilbert et al. (2011b) for their 1:1 randomized two-arm trial with 2000 subjects per treatment-arm. Suppose the conditional time-to-event endpoint, T given $S(1)$ and Z, follows a Weibull model and that $\{S(1), W\}$ follows a bivariate normal model with correlation ρ_{WS}. Information lost to drop out is MCAR, and occurs at a rate of 5% per year. Event times are censored at 3 years of follow-up at which time the trials have 50% VE on average, with an average of 104 vaccine group infections per treatment arm and 208 placebo group infections.

We investigate three time-independent Weibull models for T given $S(1)$ and Z that give three different SoP quality levels, a high quality surrogate, a marginal quality surrogate and useless surrogate. These three scenarios are used to investigate the method's bias, type 1 error of $H0_1$ and power and type 1 error of $H0_2$, in the time-independent setting. We also consider two surrogate levels under

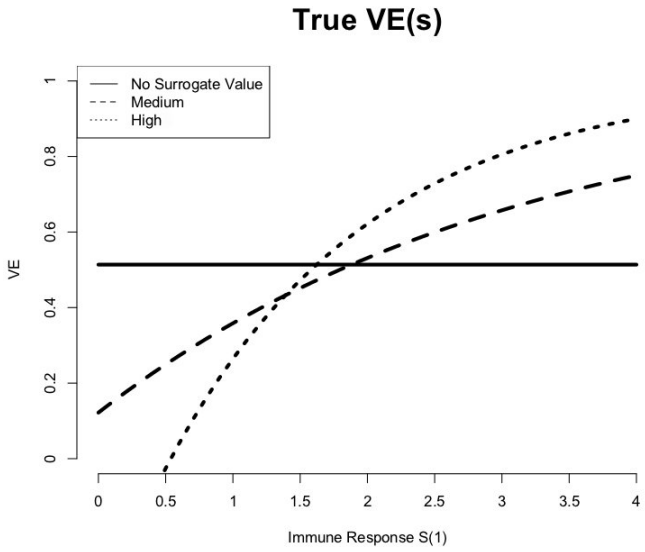

Figure 6.1: The three levels of surrogate value with no time dependence used in the simulation studies

a time-dependent Weibull model for T given $S(1)$ and Z with time-dependence in VE independent of $S(1)$. We investigate a high-quality surrogate and a marginal-quality surrogate under this type of time-dependence. Finally, we consider a high-quality surrogate and a marginal-quality surrogate under a time-dependent Weibull model for T, given $S(1)$ and Z where there is time dependence in VE that is both dependent and independent of surrogate quality. The four time-dependent scenarios are used to investigate the methods bias and power to reject $H0_1$-$H0_4$ in the presence of waning.

Figures 6.1 to 6.5 depict the true VE curves for the seven different surrogate-quality levels studied. Figure 6.1 displays the high, medium and useless quality time-independent surrogates used

in the simulations. The greater the variation in $VE(s)$ over the range of S the better the surrogate. Figures 6.3 and 6.2 display a high quality and the medium quality surrogate with waning in $VE(t|s_1)$ that is not associated with surrogate quality, which can be seen by the decreasing $VE(t|s_1)$ in the left panel of each figure. Both have similar amounts of slight waning over follow-up time; we can expected their rejection of $H0_1$ to be similar. As vaccines are not expected to wane extremely rapidly, this was considered a realistic amount of waning to consider in the simulations.

Figures 6.5 and 6.4 display a high quality and the medium quality surrogate with waning in $VE(t|s_1)$ that is both associated and unassociated with surrogate quality. This can be seen by the decreasing $VE(t|s_1)$ in the left panel of each figure and the convergence of some of the lines in the same panel as time goes on. This implies that not only is $VE(t|s_1)$ decreasing as time goes on, the surrogate is less able to differentiate between groups differing in VE. We would expect to reject $H0_1$ more often for these scenarios than for the VE waning alone. Also, the medium quality surrogate in this case is better than the medium quality surrogate in either of the medium surrogates considered, as such, we would expect to reject $H0_2$ more often in this scenario.

For each of the Weibull models for T we consider several different types of case-control sampling of $S(1)$, S^c and W, and different levels of ρ_{WS} to investigate the effects on bias and power. We only display results from 6 different types of case-control sampling of $S(1)/S^c$ and W all for $\rho_{WS} = 0.8$. For the pseudoscore method, the ρ_{WS} is actually between S(1) and a binning of the truly continuous W, which we refer to as Wdis. To make the bias and power comparable to the other two methods, we made the $\rho_{Wdis,S} = 0.8$, which required an increased $\rho_{W,S}$ of 0.89 for a 4 bin Wdis based on the quantiles of W. This increase in $\rho_{W,S}$ should be kept in mind when comparing the method.

The six case-control sampling scenarios include full sampling of both $S(1)/S^c$ and W, full sampling of $S(1)/S^c$ and fill sampling of W, full sampling of $S(1)$ and no CPV (S^c) with fill sampling of W, 5:1 control:case sampling of CPV full sampling $S(1)$ and fill sampling of W, 5:1 control:case sampling of $S(1)$ and S^c with fill sampling of W, 5:1 control:case sampling of $S(1)$ no CPV and fill sampling of W. Tables 6.2 to 6.7 below display the bias and power associated with each method.

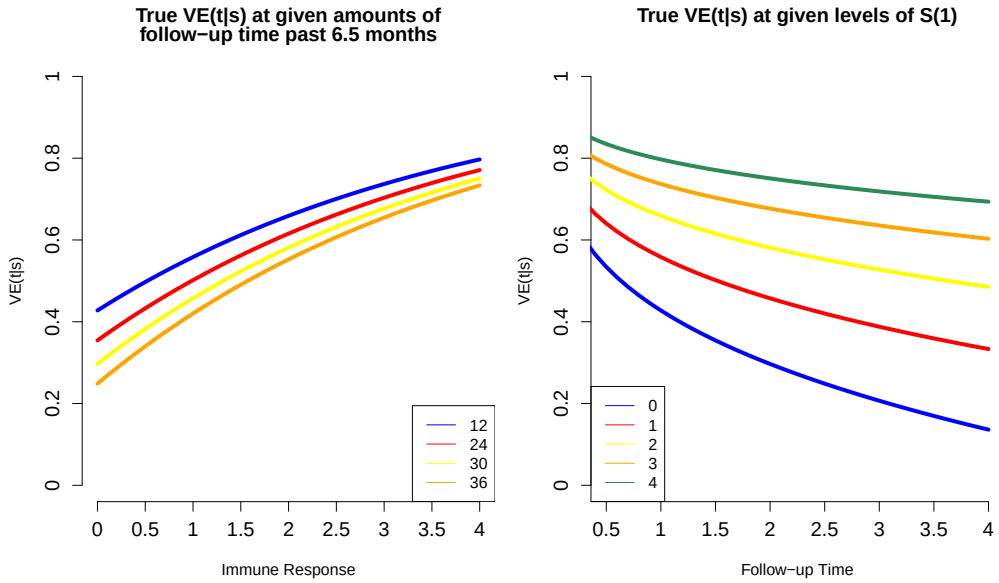

Figure 6.2: True VE for the medium surrogate with a small amount of waning in VE used in the simulation studies

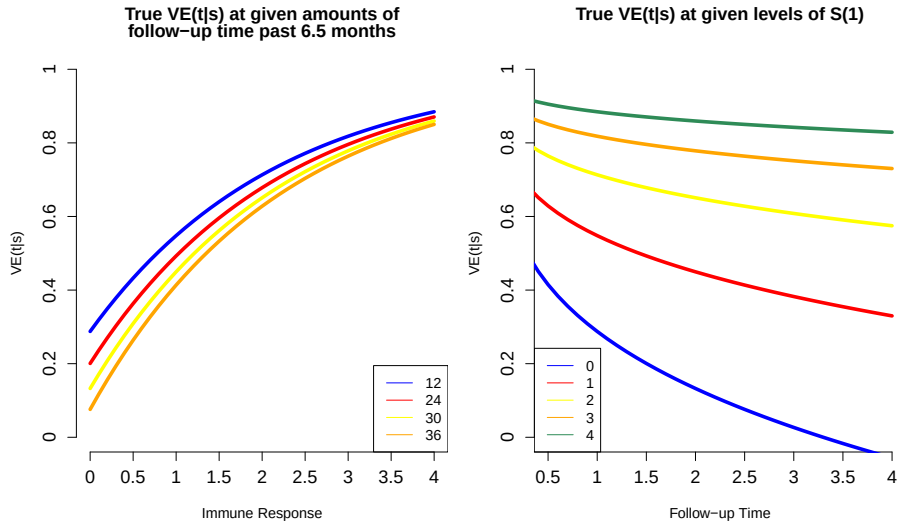

Figure 6.3: True VE for the high surrogate with a small amount of waning in VE used in the simulation studies

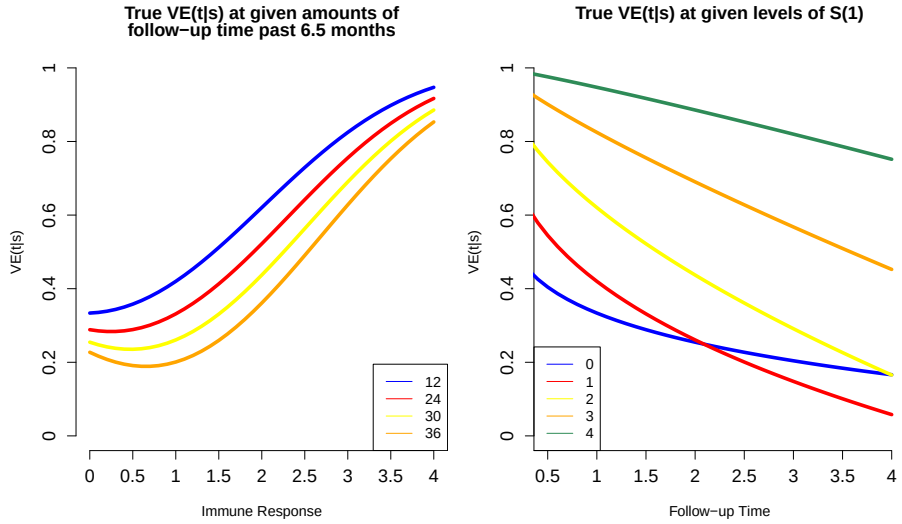

Figure 6.4: True VE for the medium surrogate with a small amount of waning in VE and in surrogate quality used in the simulation studies

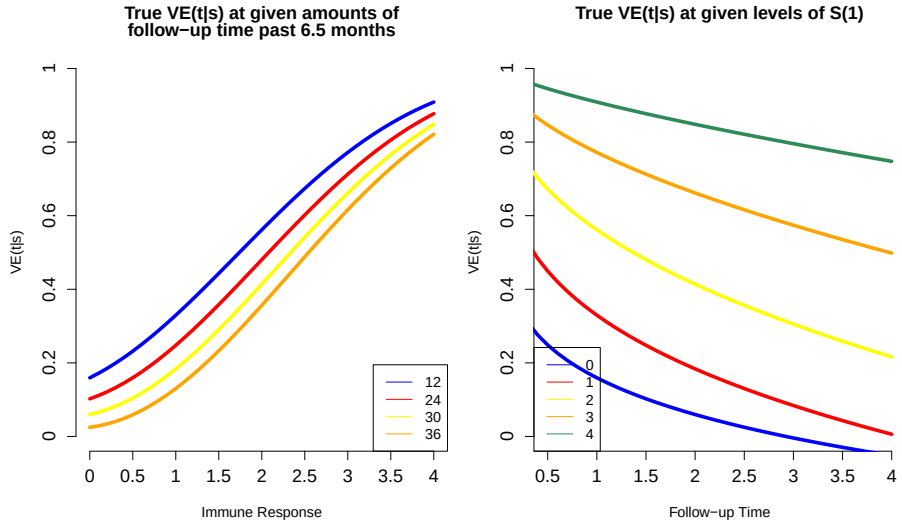

Figure 6.5: True VE for the high surrogate with a small amount of waning in VE and in surrogate quality used in the simulation studies

6.6.1 Results

Tables 6.2, 6.4 and 6.6 display the percent bias for various points on the $VE(t|s_1)$ and $VE(t|v+)$ curves for each of the 7 surrogate types and 6 sampling scenarios. As coefficients were selected to give the $VE(t|s_1)$ curves particular visual qualities, the bias of the curves are of greater ultimate importance than the bias of the coefficients. We find that all Weibull estimation methods have satisfactory performance in terms of minimal bias of the $VE(t|s_1)$ and $VE(t|v+)$ curves.

We display in Tables 6.3, 6.5 and 6.7 the results from various tests of $H0_1$ and $H0_2$ for each of the the 7 surrogate types and 6 sampling scenarios; Monte Carlo standard errors are used. A small number of bootstrap SE were calculated and found to generally similar, but in some cases slightly larger than the Monte Carlo SE. For comparison of power to other tests of $H0_1$, we display results from a test of proportional hazards using a Cox model containing treatment alone based on the Schoenfeld residuals (Grambsch and Therneau, 1994). As these results will not change based on the sampling scheme, we only display them once for each of the 7 surrogate quality scenarios, for each of the methods. We also display the results of two different tests of $H0_1$ for all of the surrogate quality levels and two of the sampling scenario for each of the methods. We find that in the full sampling scenario the Cox-based PH test has lower power to reject $H0_1$ than either of the Weibull model based tests and that the joint Wald test of $H0_1$ $\{(\beta_{10} - \beta_{00}) = \beta_{01} = (\beta_{11} - \beta_{01}) = 0\}$ generally has the highest power to reject $H0_1$. In the case of no surrogate value no time-dependence the nominal type 1 error is 12% for the fully parametric EML. For either of the other two methods the joint Wald generally has better power and correct type one error; in all cases the high quality surrogate with waning in VE alone rejects $H0_1$ more often under the sum based test. We display the sum based test for all other sampling scenarios to ensure correct type 1 error.

As mentioned above our $H0_1$:constant shape parameter was not the ideal test for time-dependence. We ideally wished to test the null $VE(t|s) = VE(s)$ under the title of $H0_1$ via the testable hypothesis $\{(\beta_{10} - \beta_{00}) = (\beta_{11} - \beta_{01}) = 0\}$ we found that this test had almost identical power to the of the testable hypothesis $\{(\beta_{10} - \beta_{00}) = \beta_{01} = (\beta_{11} - \beta_{01}) = 0\}$ in our scenarios. Results not displayed.

We also display in Tables 6.3, 6.5 and 6.7 the results for two tests for $H0_2$, for all of the surrogate types and two of the sampling scenarios. We find that the null $(\gamma^*_{11} - \gamma^*_{01} = 0)$ based Wald test of the $H0_2$ is well powered and correctly sized in the time-independent scenarios for all methods, but

tests based on the null ($\gamma_{11} - \gamma_{01} = 0$) is extremely underpowered in the time-dependent scenarios. In contrast, the ($VE(3|4) - VE(3|1) = 0$) null based test of $H0_2$ is correctly sized and well powered for all scenarios and all methods. The results of this test are also displayed for the rest of the 42 scenarios.

The simulation results of $H0_3$, based on the Wald test based on the null ($\beta_{11} - \beta_{01} = 0$) suggest the test has correct type 1 error and satisfactory power to detect the alternative ($\beta_{11} - \beta_{01} > 0$) for all methods. Based on the simulation results for null $H0_4, \beta_{10} - \beta_{00} = 0$, we find this test has correct type 1 error and satisfactory power to detect the alternative $\beta_{10} - \beta_{00} > 0$ under all estimation methods. Given the small amount of waning in the simulations, the lower power of rejection is expected. As time-dependency is increased in simulations, the power to reject null hypotheses $H0_3$ and $H0_4$ increases as expected.

As expected, power to reject $H0_1$ and $H0_2$ by any of the tests declines from full sampling power when two-phase sampling is used for $S(1)$ in the vaccine recipients. The CPV sub-sampling paradox first in Gilbert et al. (2011b) is present in both of the EML methods. The pseudoscore method corrects this paradox by accounting for the bias in the CPV sampling and increasing efficiency. The pseudoscore method also seems to be much less effected by sub-sampling of S^C. This was also found in Huang et al. (2012a).

Aside from the ability to detect and characterize time dependence, all of the methods seem to have less power loss associated with sub-sampling of $S(1)/S^C$ to detect surrogate value over their binary endpoint counterparts in the same setting Gilbert et al. (2011b). We are encouraged to see that the necessary complexity of the model did not reduce power or increase bias in comparison to the binary methods.

Table 6.2: Percent Bias: two-arm trial for given sampling of W, S^C and $S(1)$; for W and $S(1)$ correlation (0.8) parametric EML

Null	fill sampling W, full sampling S^C and $S(1)$							fill sampling W, 5:1 S^C and full $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	0.10	0.10	-0.03	0.90	0.90	0.14	0.80	-0.50	-0.40	-0.80	0.20	0.20	0.20	0.80
$\widehat{VE}(1\|4)$	0.30	0.40	-0.10	0.10	0.40	0.04	0.00	0.10	-0.60	0.10	-0.30	-0.50	0.10	0.20
$\widehat{VE}(2.5\|4)$	0.30	0.40	-0.10	-0.10	0.70	-0.06	0.00	0.10	-0.60	0.10	-1.30	-0.90	-0.80	0.10
$\widehat{VE}(2.5\|0.8+)$	0.49	0.67	-0.21	-0.30	0.90	-0.54	-0.27	0.00	-0.86	-0.26	-0.81	-0.98	-0.47	-0.98
Null	fill sampling W, no S^C and full $S(1)$							fill sampling of W, full S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|1)$	-0.40	-0.50	-0.70	0.20	0.75	0.40	1.00	-0.40	-0.50	-0.70	-0.10	0.90	0.30	-0.10
$\widehat{VE}(1\|4)$	0.80	0.10	0.00	-0.50	0.84	0.30	0.30	1.30	-0.10	0.10	-1.00	0.40	0.20	0.30
$\widehat{VE}(2.5\|4)$	0.80	0.10	0.00	-0.60	-0.25	-0.90	0.20	1.30	-0.10	0.10	-1.50	0.70	-1.00	-0.10
$\widehat{VE}(2.5\|0.8+)$	0.84	-0.13	-0.13	-1.53	-0.78	-0.72	-0.71	-0.28	-0.37	-0.17	-1.54	-0.59	-0.66	-0.35
Null	fill sampling W, 5:1 S^C and $S(1)$							fill sampling W, no S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	-0.50	-0.70	-0.70	0.00	0.69	0.30	0.85	-0.60	-0.50	-0.71	-0.10	0.80	0.35	-0.12
$\widehat{VE}(1\|4)$	0.10	-0.60	0.00	-0.60	0.79	0.00	0.36	1.01	-0.11	0.10	-0.90	0.35	0.24	0.29
$\widehat{VE}(2.5\|4)$	0.10	-0.60	0.00	-1.10	-0.26	-1.60	0.79	1.01	-0.11	0.10	-1.10	0.68	-0.85	-0.11
$\widehat{VE}(2.5\|0.8+)$	-0.74	-1.18	-0.27	-0.88	-1.05	-1.04	-1.03	-0.25	-0.30	-0.22	-1.54	-0.27	-0.39	-0.53

Table 6.3: Proportion of Rejections: two-arm trial for given sampling of W, S^C and $S(1)$; W, $S(1)$ correlation (0.8) parametric EML

Null	fill sampling W, full sampling S^C and $S(1)$							fill sampling W, 5:1 S^C and full $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
PH [a]	0.05	0.04	0.05	0.37	0.42	0.53	0.59	-	-	-	-	-	-	-
$H0_1$ [b]	0.12	0.05	0.08	0.34	0.42	0.78	0.71	0.11	0.05	0.09	0.37	0.42	0.73	0.70
$H0_1$ [c]	0.04	0.05	0.07	0.51	0.60	0.71	0.70	0.05	0.04	0.07	0.49	0.58	0.68	0.65
$H0_2$ [d]	0.04	0.77	0.99	0.10	0.28	0.05	0.12	0.04	0.77	0.99	0.09	0.28	0.05	0.14
$H0_2$ [e]	0.06	0.65	0.99	0.49	0.93	0.90	0.99	0.06	0.63	0.99	0.35	0.89	0.87	0.99
$H0_3$ [f]	0.04	0.06	0.05	0.06	0.06	0.51	0.30	0.03	0.06	0.06	0.07	0.05	0.41	0.26
$H0_4$ [g]	0.04	0.04	0.05	0.36	0.59	0.22	0.41	0.05	0.03	0.05	0.28	0.43	0.20	0.38
Null	fill sampling W, no S^C and full $S(1)$							fill sampling of W, full S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$H0_1$ [c]	0.04	0.04	0.07	0.50	0.61	0.70	0.69	0.06	0.04	0.07	0.36	0.50	0.66	0.65
$H0_2$ [e]	0.06	0.65	0.99	0.49	0.92	0.90	0.99	0.05	0.60	0.99	0.30	0.88	0.87	0.99
$H0_3$ [f]	0.03	0.07	0.06	0.06	0.07	0.50	0.30	0.03	0.06	0.06	0.05	0.06	0.37	0.26
$H0_4$ [g]	0.04	0.04	0.05	0.35	0.59	0.22	0.40	0.05	0.04	0.05	0.26	0.40	0.20	0.32
Null	fill sampling W, 5:1 S^C and $S(1)$							fill sampling W, no S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$H0_1$ [c]	0.04	0.04	0.07	0.37	0.48	0.65	0.64	0.06	0.04	0.07	0.37	0.49	0.66	0.64
$H0_2$ [e]	0.05	0.63	0.99	0.32	0.88	0.87	0.99	0.05	0.63	0.99	0.34	0.89	0.88	0.99
$H0_3$ [f]	0.03	0.06	0.05	0.06	0.05	0.37	0.25	0.03	0.06	0.06	0.06	0.06	0.37	0.25
$H0_4$ [g]	0.04	0.03	0.06	0.26	0.41	0.22	0.32	0.05	0.04	0.05	0.26	0.41	0.21	0.33

a:Proportional hazards test based on Cox model; b:Constant Shape parameter based on joint Wald of all but the constant term of the shape parameter; c:Constant Shape parameter based on the Wald test of the sum of all but the constant term of the shape parameter being equal to zero; d:$VE(t|s_1) = VE(t)$ based on a Wald test of $\gamma_{11} - \gamma_{01} = 0$; e:$VE(t|s_1) = VE(t)$ based on a Wald test of $VE(2.5|4) - VE(2.5|1) = 0$; f:test of time dependence in surrogate quality based on a Wald test of $\beta_1 1 - \beta_0 1 = 0$; g:test of time dependence in VE unassociated with the surrogate based on a Wald test of $\beta_1 0 - \beta_0 0 = 0$

Table 6.4: Percent Bias: two-arm trial for given sampling of W, S^C and $S(1)$; for W and $S(1)$ correlation (0.8) semi-parametric EML

Estimate	fill sampling W, full sampling S^C and $S(1)$							fill sampling W, 5:1 S^C and full $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	0.40	0.30	-0.10	0.40	0.50	1.00	0.50	0.40	0.30	-0.10	0.50	0.30	0.90	0.60
$\widehat{VE}(1\|4)$	1.90	0.60	0.30	-0.10	0.10	0.50	0.20	1.60	0.40	0.30	-0.20	-0.00	0.30	0.30
$\widehat{VE}(2.5\|4)$	1.90	0.60	0.30	-0.70	0.10	-0.40	0.10	1.60	0.40	0.30	-1.00	0.10	-0.60	0.00
$\widehat{VE}(2.5\|0.8+)$	2.02	0.85	0.32	-1.60	-0.52	-0.88	1.47	0.90	0.28	-1.48	-0.47	-1.07	-0.89	
Estimate	fill sampling W, no S^C and full $S(1)$							fill sampling of W, full S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	0.40	0.20	-0.10	0.40	0.60	0.80	0.50	0.30	0.40	0.10	0.30	0.60	1.40	0.80
$\widehat{VE}(1\|4)$	1.70	0.60	0.10	-0.40	0.30	0.30	0.30	1.10	1.00	-1.70	-0.20	0.20	0.10	0.20
$\widehat{VE}(2.5\|4)$	1.70	0.60	0.10	-1.10	0.40	-1.20	0.00	1.10	1.00	0.20	-1.30	0.10	-1.20	0.10
$\widehat{VE}(2.5\|0.8+)$	1.10	0.77	0.36	-1.49	-0.86	-0.91	-0.66	1.42	1.07	0.49	-0.95	-0.38	-1.15	-0.61
Estimate	fill sampling W, 5:1 S^C and $S(1)$							fill sampling W, no S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	0.30	0.30	0.20	0.30	0.40	1.50	0.80	0.40	0.40	0.10	0.30	0.40	1.00	0.60
$\widehat{VE}(1\|4)$	1.10	0.70	0.20	-0.40	0.10	0.10	0.30	1.40	0.60	0.10	-0.40	0.20	0.20	0.30
$\widehat{VE}(2.5\|4)$	1.10	0.70	0.20	-0.80	-0.20	-0.50	0.20	1.40	0.60	0.10	-1.10	0.00	-0.60	0.10
$\widehat{VE}(2.5\|0.8+)$	1.02	0.95	0.36	-1.20	-0.39	-1.25	-0.79	0.70	1.09	0.26	-1.54	-0.50	-0.96	-0.69

Table 6.5: Proportion of Rejections: two-arm trial for given sampling of W, S^C and $S(1)$; W, $S(1)$ correlation (0.8) semi-parametric EML

Null	fill sampling W, full sampling S^C and $S(1)$							fill sampling W, 5:1 S^C and full $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
PH [a]	0.05	0.04	0.05	0.37	0.42	0.53	0.59	-	-	-	-	-	-	-
$H0_1$ [b]	0.05	0.05	0.06	0.36	0.45	0.81	0.71	0.05	0.06	0.06	0.36	0.42	0.78	0.65
$H0_1$ [c]	0.04	0.04	0.04	0.39	0.54	0.69	0.69	0.04	0.04	0.04	0.36	0.54	0.68	0.66
$H0_2$ [d]	0.05	0.72	0.99	0.09	0.25	0.04	0.10	0.05	0.69	0.99	0.09	0.22	0.04	0.10
$H0_2$ [e]	0.05	0.67	0.99	0.36	0.90	0.88	0.99	0.05	0.66	0.99	0.35	0.89	0.87	0.99
$H0_3$ [f]	0.05	0.07	0.07	0.06	0.05	0.39	0.24	0.05	0.07	0.07	0.06	0.05	0.40	0.25
$H0_4$ [g]	0.05	0.04	0.04	0.27	0.47	0.21	0.35	0.05	0.03	0.03	0.26	0.45	0.21	0.33
Null	fill sampling W, no S^C and full $S(1)$							fill sampling of W, full S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$H0_1$ [b]	0.04	0.04	0.04	0.38	0.54	0.69	0.68	0.04	0.04	0.04	0.36	0.53	0.66	0.65
$H0_2$ [e]	0.05	0.68	0.99	0.36	0.90	0.88	0.99	0.05	0.60	0.99	0.32	0.88	0.84	0.99
$H0_3$ [f]	0.06	0.07	0.07	0.06	0.04	0.39	0.24	0.05	0.06	0.06	0.07	0.05	0.36	0.24
$H0_4$ [g]	0.04	0.03	0.03	0.28	0.46	0.21	0.35	0.05	0.06	0.06	0.25	0.44	0.21	0.32
Null	fill sampling W, 5:1 S^C and $S(1)$							fill sampling W, no S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$H0_1$ [b]	0.04	0.04	0.07	0.37	0.51	0.65	0.64	0.06	0.04	0.07	0.37	0.52	0.66	0.67
$H0_2$ [e]	0.04	0.63	0.99	0.33	0.88	0.85	0.99	0.05	0.66	0.99	0.36	0.88	0.85	0.99
$H0_3$ [f]	0.03	0.06	0.05	0.06	0.05	0.37	0.25	0.03	0.06	0.06	0.06	0.06	0.37	0.25
$H0_4$ [g]	0.04	0.03	0.06	0.26	0.43	0.21	0.32	0.05	0.04	0.05	0.26	0.44	0.21	0.33

a:Proportional hazards test based on Cox model; b: Constant Shape parameter based on joint Wald of all but the constant term of the shape parameter; c:Constant Shape parameter based on the Wald test of the sum of all but the constant term of the shape parameter being equal to zero; d:$VE(t|s_1) = VE(t)$ based on a Wald test of $\gamma_{11} - \gamma_{01} = 0$; e:$VE(t|s_1) = VE(t)$ based on a Wald test of $VE(2.5|4) - VE(2.5|1) = 0$; f:test of time dependence in surrogate quality based on a Wald test of $\beta_1 1 - \beta_0 1 = 0$; g:test of time dependence in VE unassociated with the surrogate based on a Wald test of $\beta_1 0 - \beta_0 0 = 0$

Table 6.6: Percent Bias: two-arm trial for given sampling of W, S^C and $S(1)$; for W and $S(1)$ correlation (0.8) pseudoscore

Estimate	fill sampling W, full sampling S^C and $S(1)$							full sampling W, 5:1 S^C and full $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	-0.50	-0.80	-0.40	0.30	0.70	1.10	1.30	-0.50	-0.80	-0.50	0.40	0.70	1.10	0.70
$\widehat{VE}(1\|4)$	-0.10	-0.50	-0.20	0.00	0.00	0.10	0.30	-0.20	-0.50	-0.20	0.00	-0.00	0.20	0.10
$\widehat{VE}(2.5\|4)$	-0.10	-0.50	-0.20	-1.20	0.20	-0.40	0.50	-0.20	-0.50	-0.20	-1.20	0.10	-0.20	0.20
$\widehat{VE}(2.5\|0.8+)$	0.10	-0.68	-0.44	-1.13	-0.61	-0.86	-0.89	0.17	-0.71	-0.46	-1.11	-0.70	-0.84	-0.55
Estimate	full sampling W, no S^C and full $S(1)$							full sampling of W, full S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	-	-	-	-	-	-	-	-0.50	-0.80	-1.80	-0.10	-0.90	0.30	-0.20
$\widehat{VE}(1\|4)$	-	-	-	-	-	-	-	-0.20	-1.60	-2.30	-0.40	2.40	0.60	1.10
$\widehat{VE}(2.5\|4)$	-	-	-	-	-	-	-	-0.20	-1.60	-2.30	-2.30	-0.30	-2.80	0.10
$\widehat{VE}(2.5\|0.8+)$	-	-	-	-	-	-	-	-0.12	-1.59	-2.84	-2.06	-0.84	-2.20	-0.94
Estimate	full sampling W, 5:1 S^C and $S(1)$							full sampling W, no S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$\widehat{VE}(1\|2)$	-0.50	-0.70	-1.80	-0.10	-0.90	0.40	-0.20	-	-	-	-	-	-	-
$\widehat{VE}(1\|4)$	-0.30	-1.60	-2.30	-2.10	2.40	0.60	1.10	-	-	-	-	-	-	-
$\widehat{VE}(2.5\|4)$	-0.30	-1.60	-2.30	-2.40	-0.30	-2.50	0.00	-	-	-	-	-	-	-
$\widehat{VE}(2.5\|0.8+)$	-0.14	-1.67	-2.87	-2.19	-0.98	-2.96	-1.19	-	-	-	-	-	-	-

Table 6.7: Proportion of Rejections: two-arm trial for given sampling of W, S^C and $S(1)$; W, $S(1)$ correlation (0.8) pseudoscore

Null	fill sampling W, full sampling S^C and $S(1)$							full sampling W, 5:1 S^C and full $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
PH [a]	0.05	0.04	0.05	0.37	0.42	0.53	0.59	-	-	-	-	-	-	-
$H0_1$ [b]	0.05	0.05	0.05	0.34	0.44	0.79	0.70	0.05	0.05	0.06	0.34	0.44	0.78	0.70
$H0_1$ [c]	0.05	0.04	0.05	0.40	0.54	0.65	0.66	0.05	0.04	0.05	0.39	0.54	0.66	0.68
$H0_2$ [d]	0.05	0.72	0.99	0.10	0.22	0.05	0.12	0.05	0.70	0.99	0.09	0.19	0.04	0.10
$H0_2$ [e]	0.06	0.68	0.99	0.42	0.92	0.91	0.99	0.06	0.68	0.99	0.42	0.92	0.91	0.99
$H0_3$ [f]	0.05	0.07	0.06	0.07	0.05	0.37	0.24	0.05	0.07	0.06	0.07	0.05	0.37	0.24
$H0_4$ [g]	0.06	0.03	0.04	0.27	0.46	0.23	0.35	0.06	0.03	0.04	0.27	0.46	0.23	0.35
Null	full sampling W, no S^C and full $S(1)$							full sampling of W, full S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$H0_1$ [c]	-	-	-	-	-	-	-	0.05	0.04	0.05	0.39	0.53	0.64	0.67
$H0_2$ [e]	-	-	-	-	-	-	-	0.06	0.65	0.99	0.34	0.90	0.84	0.99
$H0_3$ [f]	-	-	-	-	-	-	-	0.05	0.07	0.07	0.06	0.13	0.53	0.49
$H0_4$ [g]	-	-	-	-	-	-	-	0.05	0.03	0.03	0.21	0.32	0.17	0.23
Null	full sampling W, 5:1 S^C and $S(1)$							full sampling W, no S^C and 5:1 $S(1)$						
	Time Ind			VE wane		Both wane		Time Ind			VE wane		Both wane	
	No Val	Some Val	High Val	Some	High	Some	High	No	Some	High	Some	High	Some	High
$H0_1$ [c]	0.05	0.04	0.04	0.39	0.52	0.65	0.67	-	-	-	-	-	-	-
$H0_2$ [e]	0.06	0.64	0.99	0.34	0.89	0.83	0.99	-	-	-	-	-	-	-
$H0_3$ [f]	0.05	0.07	0.10	0.08	0.13	0.53	0.49	-	-	-	-	-	-	-
$H0_4$ [g]	0.05	0.03	0.03	0.26	0.32	0.17	0.23	-	-	-	-	-	-	-

a:Proportional hazards test based on Cox model; b:Constant Shape parameter based on joint Wald of all but the constant term of the shape parameter; c:Constant Shape parameter based on the Wald test of the sum of all but the constant term of the shape parameter being equal to zero; d:$VE(t|s_1) = VE(t)$ based on a Wald test of $\gamma_{11} - \gamma_{01} = 0$; e:$VE(t|s_1) = VE(t)$ based on a Wald test of $VE(2.5|4) - VE(2.5|1) = 0$; f:test of time dependence in surrogate quality based on a Wald test of $\beta_1 1 - \beta_0 1 = 0$; g:test of time dependence in VE unassociated with the surrogate based on a Wald test of $\beta_1 0 - \beta_0 0 = 0$

Comparison of Methods

Table 6.8 display the differences between the methods at various correlation levels of the BIP and potential SoP for rejecting $H0_2$. At a high level of BIP correlation all the methods perform well, but as the correlation between W and S(1) declines so does the power in all the methods. The pseudoscore method is the least affected by the decreased correlation while the semi-parametric and parametric EML methods seem equally affected. This is most likely due to the fact that the pseudoscore method relies least on the BIP/S(1) relationship being strong as it uses the BIP only to group the S(1) values and provided there is some separation of S(1) values it seems to have reasonable power.

Table 6.8 display the differences between the methods at various correlation levels of the BIP and potential SoP for rejecting $H0_2$. Again at high correlation between BIP and S(1) all the methods perform well. However unlike for rejection of $H0_2$, all the methods decline similarly in power as the correlations declines.

Pseudoscore method failure occurs at a lower rate than do both the EML methods. Convergence failure was observed only once in all the (0.8) correlation of S(1) and W simulations, at correlation (0.5) the failure rate increased to 0.1% in both EML methods, but failure did not occur in the pseudoscore simulations. When the correlation was further reduced the failure rate increased to 0.6%, for the EML methods but was ($< 0.01\%$) in the pseudoscore method. The biases are not given in Table 6.8 or 6.9 because they are all very similar to the (0.8) correlation biases, suggesting that reduced BIP/S(1) correlation although detrimental to power does not bias the results from any of the methods.

Table 6.10 compares biases for the summary statistics suggested in Chapter 5 based on the risk difference. We find that at full sampling and (0.8) BIP/S(1) correlation that all the methods perform well in estimating the summary statistics accurately. Based on the lower correlation and sub-sampling biases are found for each method; we are confident these biases will not increase significantly under these scenarios. A small set of simulations allowing for sub-sampling of S(1) and lower correlation of BIP/S(1) support this conjecture. Power to detect difference in surrogate quality between medium and high quality surrogates under each time of time dependence were ran on a limit set of the data simulations. Findings suggest that test for differences in both STG(2.5) and

pTG(2.5) have good poor in the full sampling scenarios for all methods ranging from 90% to 60% depending of the relationship with time. It was also found that PPV(2.5—0.85) had good power to detect surrogate value as compared to the average prevalence difference between arms.

Table 6.11 displays the proportion of rejections per 1000 simulations of the STG(t) for a given high-quality SoP being significantly better than the time independent medium SoP. As expected, the power to detect differences between the time-independent high-quality SoP and the medium-quality time-independent SoP improves over time as the risk is CDF based and more events have occurred. The power is also reasonable, greater than 50% in the ideal case with the parametric EML having the best power. Both of the STGs summarizing time-dependent SoP, decline in power over time, as they are either declining in VE difference while the time-independent medium surrogate is not, or they are declining is both VE and SoP quality. The power for $pTG(t)$ to differentiate between high-quality SoP and medium-quality SoP was found to be very poor, results not shown. This is believed to not the scenarios where $pTG(t)$ is useful as all simulation scenarios use continuous SoP, $pTG(t)$ is believed to be most useful when comparing a continuous SoP and a discrete SoP. These scenarios will be investigated in greater detail the for peer reviewed publication.

The $PPV(t|v)$ has disappointingly lower power to differentiate between high-quality SoP and nothing. The lack of power is surprising given that the CI of $PPV(t|v)$ merely has to exclude the average incidence prevalence difference over trial arms to show evidence of partial SoP value. However, the same pattern of increasing power for the time-independent over time, while power decreases for both time-dependent scenarios. The lack of power may suggest that $PPV(t|v)$ may be more useful as a curve, rather than as a summary statistic. The lack of power seems to be based on the large variation over the simulations rather than a lack of estimated increased predictive power over no SoP, although this does not seems to have to same impact on the $STG(t)$ difference estimates.

Table 6.8: Comparison Over Methods and BIP Correlation $H0_2$: full sampling W, full S^C and $S(1)$

W, $S(1)$ correlation (0.8)

Method	Proportion of Rejection $H0_2$						
	Time Ind			VE wane		Both Wane	
	No Val	Some Val	High Val	Some	High	Some	High
Parametric EML	0.5	0.68	0.99	0.49	0.93	0.90	0.99
Semi-parametric EML	0.5	0.67	0.99	0.36	0.91	0.88	0.99
Pseudoscore	0.5	0.68	0.99	0.42	0.92	0.91	0.99

W, $S(1)$ correlation (0.5)

Method	Proportion of Rejection $H0_2$						
	Time Ind			VE wane		Both Wane	
	No Val	Some Val	High Val	Some	High	Some	High
Parametric EML	0.05	0.36	0.99	0.24	0.79	0.70	0.96
Semi-parametric EML	0.06	0.41	0.99	0.24	0.77	0.70	0.96
Pseudoscore	0.05	0.40	0.99	0.29	0.83	0.82	0.98

W, $S(1)$ correlation (0.25)

Method	Proportion of Rejection $H0_2$						
	Time Ind			VE wane		Both Wane	
	No Val	Some Val	High Val	Some	High	Some	High
Parametric EML	0.06	0.16	0.88	0.14	0.51	0.46	0.78
Semi-parametric EML	0.05	0.17	0.90	0.16	0.45	0.42	0.77
Pseudoscore	0.06	0.16	0.90	0.18	0.72	0.66	0.86

Table 6.9: Comparison Over Methods and BIP Correlation $H0_1$: full sampling W, full S^C and $S(1)$

W, $S(1)$ correlation (0.8)							
Method	Proportion of Rejection $H0_1$						
	Time Ind			VE wane		Both Wane	
	No Val	Some Val	High Val	Some	High	Some	High
Parametric EML	0.04	0.05	0.07	0.51	0.60	0.71	0.70
Semi-parametric EML	0.04	0.04	0.05	0.39	0.54	0.69	0.69
Pseudoscore	0.04	0.04	0.04	0.40	0.54	0.68	0.69
W, $S(1)$ correlation (0.5)							
Method	Proportion of Rejection $H0_1$						
	Time Ind			VE wane		Both Wane	
	No Val	Some Val	High Val	Some	High	Some	High
Parametric EML	0.05	0.05	0.06	0.23	0.35	0.47	0.47
Semi-parametric EML	0.05	0.05	0.05	0.27	0.39	0.48	0.50
Pseudoscore	0.05	0.04	0.05	0.32	0.39	0.47	0.49
W, $S(1)$ correlation (0.25)							
Method	Proportion of Rejection $H0_1$						
	Time Ind			VE wane		Both Wane	
	No Val	Some Val	High Val	Some	High	Some	High
Parametric EML	0.05	0.02	0.03	0.17	0.25	0.35	0.35
Semi-parametric EML	0.05	0.04	0.04	0.21	0.28	0.36	0.37
Pseudoscore	0.05	0.05	0.06	0.23	0.28	0.35	0.39

Table 6.10: Percent Bias: summary statistic comparison over methods two-arm trial; W, $S(1)$ correlation (0.8)

Estimate	full sampling W, full sampling S^C and $S(1)$, Highy Valuable Surrogate								
	Parametric EML			Semi-parametric EML			Pseudoscore		
	Time Ind	VE wane	Both wane	Time Ind	VE wane	Both wane	Time Ind	VE wane	Both wane
$\widehat{STG}(2.5)$	1.76	0.84	1.73	1.52	1.06	1.73	1.78	0.98	1.59
$\widehat{ptg}(2.5\|0.75, 0.05)$	0.07	0.11	0.19	0.07	0.12	0.19	0.05	0.10	0.16
$\widehat{PPV}(2.5\|0.75)$	0.03	0.09	0.13	0.08	0.07	0.14	0.02	0.05	0.10
$\widehat{NPV}(2.5\|0.05)$	-0.01	0.14	0.24	-0.04	0.14	0.23	-0.02	0.14	0.22
Estimate	full sampling W, 5:1 S^C and $S(1)$, Highy Valuable Surrogate								
	Parametric EML			Semi-parametric EML			Pseudoscore		
	Time Ind	VE wane	Both wane	Time Ind	VE wane	Both wane	Time Ind	VE wane	Both wane
$\widehat{STG}(2.5)$	2.28	0.84	2.55	2.12	1.30	2.03	-3.69	1.19	2.16
$\widehat{ptg}(2.5\|0.75, 0.05)$	0.08	0.11	0.27	0.10	0.13	0.20	-0.16	0.31	0.41
$\widehat{PPV}(2.5\|0.75)$	0.02	0.09	0.20	0.10	0.08	0.15	-0.06	0.34	0.48
$\widehat{NPV}(2.5\|005)$	0.11	0.14	0.29	0.05	0.23	0.35	-1.62	-0.03	-0.01

Average bias (absolute bias times 100) of summary statistic estimates over 1000 simulated experiments. All estimates in all scenarios have average bias less than 3% of the Monte Carlo standard error.

Table 6.11: Proportion of Rejections: two-arm trial full sampling of W, S^C and $S(1)$; W, $S(1)$ correlation (0.8)

	Parametric EML		Semi-parametric EML		Pseudoscore	
	$\widehat{STG}(1)$	$\widehat{STG}(2)$	$\widehat{STG}(1)$	$\widehat{STG}(2)$	$\widehat{STG}(1)$	$\widehat{STG}(2)$
Time-independent	0.516	0.526	0.414	0.434	0.568	0.590
VE Wane	0.144	0.078	0.138	0.088	0.126	0.056
Both Wane	0.136	0.088	0.198	0.104	0.184	0.090
	Parametric EML		Semi-parametric EML		Pseudoscore	
	$\widehat{STG}(1)$	$\widehat{STG}(2)$	$\widehat{STG}(1)$	$\widehat{STG}(2)$	$\widehat{STG}(1)$	$\widehat{STG}(2)$
Time-independent	0.024	0.274	0.20	0.29	0.018	0.29
VE Wane	0.156	0.012	0.156	0.08	0.15	0.08
Both Wane	0.05	0.012	0.05	0.012	0.05	0.03

Power is based on Monte Carlo SE, and comparisons are made between the time-independent SoP and the given high-quality time-dependent SoP

6.6.2 BIP and S(1) Sub-sampling: ZEST Based Simulations

The ZEST trial data have a unique sampling that was not considered in the main simulations, sub-sampling of both the BIP and S(1) with no CPV. This was not a scenario covered in the simulations above because it poses a new challenge to any of the methods. We do not pursue the pseudoscore method for application for these data as CPV not being performed violates a main assumption. We considered three extensions of the above EML methods. The first and most direct extension is to the fully parametric EML method under which subjects missing both S(1) and the BIP have their likelihood estimated by integration over $F_{S(1)}(s)$ without the aid of other baseline covariates, Q. Subjects missing both S(1) and W have their likelihood estimated via:

$$\int g_z(t|s, y, \gamma, \beta) dF_{S(1)}(s).$$

This adaptation to allow for BIP sub-sampling was suggested and tested in Gilbert et al. (2011b), where it was found to have markedly lower power full sampling of the BIP in the binary outcome case.

The second extension to the fully parametric EML method is a slight modification of a method from Gilbert and Hudgens (2008). This method leverages other baseline covariates to estimate the likelihood for those subjects missing both S(1) and W. Let Q be a set of baseline covariates and Qmod be a linear combination of those variables fit in the validation sample with a model for W given Q. Then we can apply the Gilbert and Hudgens (2008) method such that subjects missing S(1) and W have their likelihood estimated by:

$$\int \int g_z(t|s, w, q, y, \gamma, \beta) dF_{S(1)|W,Qmod}(s|w, Qmod) * dF_{W|Qmod}(w|Qmod).$$

Gilbert and Hudgens (2008) found that this method performed well using non-parametric EML for a binary outcome, we test it in simulations. Estimation of both $F_{S(1)|W,Qmod}$ and $F_{W|Qmod}$ can be accomplished via weighted maximum likelihood, with weights reflecting the inverse probability of being selected to have the dependent variable measured given you were selected to have the covariates measured.

The third extension is an adaptation of double integration method of Gilbert and Hudgens (2008) to the semi-parametric EML of Huang and Gilbert (2011). Using the same location scale model as in Huang and Gilbert (2011) for S(1) given W, we fit W given Qmod among those having W

measured. We then use this model to impute the missing Ws in the same way we used it previously to impute missing S(1). We then use the imputed W values to impute the S(1) values and take the empirical integral over both imputations sets. We weight each of location scale models by the inverse probability of selection for measurement of the outcome given measurement of the covariates. We then use the product of those weights in the empirical integral to weight the imputed likelihoods. This results in a similar method as Gilbert and Hudgens (2008) but does not require the assumption of a parametric model of W given Qmod. Also, the empirical integration removes the need for double numeric integration.

We ran four surrogate scenarios for each of the proposed method extensions. We ran two time-independent scenarios, a useless and a high quality surrogate, and two time-dependent scenarios both high-quality surrogates one with time-dependence in VE both independent and dependent of surrogate quality and one with time-dependence in VE independent of surrogate quality alone. We tailored the simulations to follow more closely the ZEST data. To match the Zoster data we set a BIP/S(1) correlation of 0.7 on average and a Qmod/BIP correlation of 0.2. This was the highest Qmod/BIP correlation we could obtain in the Zostavax dataset using a combination of all measured baseline covariates and linear regression. We assume that BIP, S(1) and Qmod are all normally distributed as is suggested by the ZEST data, using the log of titers.

Following the data example all infected vaccine recipients plus a random set of vaccinated controls had both S(1) and W measured and all other vaccine recipients had only Q measured. Therefore, the inverse probability of having S(1) measured given that you had W measured is just the inverse of the probability of having S(1) measured among the vaccine recipients. All infected placebo recipients plus a randomly selected set of placebo controls had W measured and no CPV was performed. The event rate, drop-out rate and sample size are left the same from the previous simulations for comparison of bias and power to the original methods.

The simulations suggest that our current formulation of semi-parametric method with BIP-subsampling is biased and that bias is causing an increase in the power and very high type 1 errors for $H0_3$. This method is clearly not reliable in its current form and more work is needed to determine what is causing the bias, although several attempts were made to correct the bias with weighting in the location scale models and in the use of the imputed BIP values, the bias and the strangely inflated power persist. Of the two parametric methods the one that does not leverage other baseline

covariates has the lower bias and very similar power to the one that leverages Q. The concerning case of a time-independent surrogate with high surrogate value has high bias suggesting that the method is biased in this particular case. The biases are all slightly larger than two Monte Carlo SE from the true for this case, while all other biases are well within two SE for this same method. There also seems to be increased type 1 error for null hypothesis $H0_3$ in this case.

We believe that with a poorly correlated Qmod that, although there are some power gains to be had, it may not be worth the effort or the possible increased bias. When we increased the Qmod/W correlation the bias was reduced and the power was increased, almost to the level of the original full BIP or fill sampling. Oddly, increased Qmod correlation with W did not seem to improve the bias in the semi-parametric method. This suggests that if one had a Qmod that was more correlated with the BIP in the Zoster example, the parametric EML method that leverages Qmod would clearly be preferable. However, we could not find such a Qmod, for this reason we used method one for the Zoster data analysis.

Table 6.12: BIP Sub-sampling: simulations following the ZEST clinical trial sub-sampling of W and S(1)

	Percent Bias: 5:1 Sampling W and $S(1)$ No S^C											
	Method 1 No Q				Method 2 With Q				Method 3 Semi-Para Q			
Estimate	Time Ind		VE wane	Both wane	Time Ind		VE wane	Both wane	Time Ind		VE wane	Both wane
	No Val	High Val	High	High	No	High	High	High	No	High	High	High
$\widehat{VE}(1\|2)$	-0.10	-0.20	0.90	0.70	0.20	-4.00	-1.10	0.50	0.70	6.80	-2.00	-3.10
$\widehat{VE}(1\|3)$	0.60	0.00	0.30	0.00	1.50	-4.30	2.80	2.40	-0.80	5.60	4.20	1.40
$\widehat{VE}(2.5\|3)$	0.60	0.00	0.40	-0.10	1.50	-4.30	-1.00	-1.40	-0.80	5.60	2.90	2.00
$\widehat{VE}(2.5\|+0.8)$	1.27	-0.41	0.69	0.54	1.27	-3.32	0.52	0.98	-1.63	4.83	3.83	6.01
	Proportion of Rejections: 5:1 Sampling W and $S(1)$ No S^C											
	Method 1 No Q				Method 2 With Q				Method 3 Semi-Para Q			
Null	Time Ind		VE wane	Both wane	Time Ind		VE wane	Both wane	Time Ind		VE wane	Both wane
	No Val	High Val	High	High	No	High	High	High	No	High	High	High
$H0_1$ [c]	0.04	0.04	0.45	0.57	0.05	0.08	0.82	0.95	0.05	0.06	0.32	0.64
$H0_2$ [e]	0.06	0.95	0.80	0.95	0.07	0.99	0.85	0.98	0.07	0.99	0.92	0.99
$H0_3$ [f]	0.05	0.06	0.06	0.23	0.05	0.10	0.05	0.33	0.05	0.03	0.24	0.70
$H0_4$ [g]	0.05	0.04	0.38	0.28	0.05	0.01	0.46	0.51	0.05	0.06	0.18	0.04

c:$VE(t|s_1) = VE(s_1)$ based on the Wald test of the sum of all but the constant term of the shape parameter being equal to zero; e:$VE(t|s_1) = VE(t)$ based on a Wald test of $VE(2.5|4) - VE(2.5|1) = 0$; f:test of time dependence in surrogate quality based on a Wald test of $\beta_1 1 - \beta_0 1 = 0$; g:test of time dependence in VE unassociated with the surrogate based on a Wald test of $\beta_1 0 - \beta_0 0 = 0$

Chapter 7

REAL DATA EXAMPLES

7.1 Step Trial

We illustrate our EML methods using the data from the Step HIV vaccine efficacy trial. This trial is described in Table 1.1. It was one-to-one randomized trial of 3000 subjects to receive the MRKAd5 HIV-1 Gag/Pol/Nef vaccine or placebo stratified by sex, baseline Ad5 titers above and below 18 and study site. The clinical endpoint of the Step study, HIV infection, was not found to be affected by the vaccine and there was some evidence that particular subgroups of the trial population may have had increased risk of HIV infection associated with vaccination (Buchbinder et al., 2008b).

There were 1836 male subjects that qualified for the modified intent to treat (MITT) group and of them 1821 were not infected or censored prior to the week 8 blood draw. We treat the cohort of 1821 subjects as our complete trial data; containing 906 vaccine recipients and 915 placebo recipients. From this trial population we wish to study the same three candidate SoPs as Huang and Gilbert (2011), the magnitude of pre-infection post-vaccination HIV-specific T-cell responses via Elispot to epitopes found in the vaccine insert specific to Gag, Pol and Nef proteins. Table 7.1 displays the baseline and outcome variables over the various study populations of interest.

The original protocol specified performance of immunogenicity analyses on 25% of study participants, stratified on treatment status and study site, as designated in McElrath et al. (2008) this is the stratified random sample. Several immunological sub-studies were planned and undertaken based on week 8 per-infection immune response and the sampling scheme differs for each study (McElrath et al., 2008). In the sub-study we use to obtain our potential SoP, measurements were obtained for almost all vaccinated members of the stratified random sample and almost all per-protocol infected vaccine recipients. The per-protocol sample includes all participants who received at least the first two doses of either vaccine or placebo.

Of the 906 vaccine recipients, 41 infections occurred post week 8 and prior to unblinding and of those infected 35 had the candidate SoPs measured, (85.4%). In the 865 vaccine recipients

Table 7.1: Data Summary Step Trial: cohort used in SoP analysis

Group	Vaccinees	Vaccinees with S	Placeboees
	N(%)/Mean(SD)	N(%)/Mean(SD)	N(%)/Mean(SD)
Infections after week 8	41(4.5%)	35(14.7%)	26(2.8%)
Age (years)	30.5(7.75)	30.08(7.89)	30.58(8)
Ethnicity			
Black	93(10.3%)	19(8%)	89(9.7%)
Hispanic	79(8.7%)	17(7.1%)	92(10.1%)
Mestizo/Mestiza	235(25.9%)	65(27.3%)	230(25.1%)
Multi-racial	29(3.2%)	5(2.1%)	18(2%)
Other	28(3.1%)	7(2.9%)	28(3.1%)
White	442(48.8%)	125(52.5%)	458(50.1%)
Not in North America	335(37%)	84(35.3%)	329(36%)
S measured	238(26.3%)	238(100%)	240(26.2%)
GAG	5.38(1.04)	5.38(1.04)	3.38(0.92)
POL	5.89(1.07)	5.89(1.07)	4.27(0.79)
NEF	5.31(1.05)	5.31(1.05)	3.5(0.95)

uninfected at the close of the trial, 203 of them had the candidate SoPs measured, (23.5%). There was no CPV performed. All subjects had baseline Ad5 titers measured. As no CPV was performed the pseudoscore method is not suitable for this example. We perform both semi-parametric and fully parametric analyses on each of the three candidate SoP. To account for the sub-sampling of the SoP, we weight the models for SoP given W in both methods with inverse probability weight of $1/0.235$ for uninfected and $1/0.854$ for infected subjects.

7.1.1 Parametric EML Analysis of Step

All subjects had baseline Ad5 titers measured. We will use log Ad5 titers as our BIP, W, and refer to our candidate SoP as $S1_{Gag}$, $S1_{Pol}$ and $S1_{Nef}$ which are the log of T-cell response magnitudes

specific to Gag, Pol and Nef, respectively. There was an estimated correlation between -0.3 and -0.4 for all three candidate SoP and log Ad5 titers. We use the time in years from first vaccination to infection, censoring or trial unblinding as our right-censored time-to-event measure, X, and infection prior to trial unblinding, October 17, 2007, as our event indicator, Y. If subjects were infected or censored prior to week 8, we removed them from the analysis.

We fit the saturated Weibull model using starting values of all zeros for the shape parameter components and the values seven, negative four, zero and zero for the scale parameter components of constant, vaccination, potential SoP value and the interaction of vaccination and potential SoP, respectively, as these were consistent with the observed time-to-event values in the data and the negative VE. We ran models for each of the three potential SoP separately. We assumed a bivariate normal distribution for S(1) and W; although there is evidence in the data to suggest departure from normality, there is no evidence to suggest the model is not location scale. As we obtained very similar results from the semi-parametric models, we are not overly concerned with the potential surrogates' deviation from normality causing bias.

We found no evidence of time-dependence in any of the three models; P-values of (0.381,0.325,0.344) for the models containing $S1_{Gag}$, $S1_{Pol}$, $S1_{Nef}$ respectively. There is a small number of events, prior to unblinding which reduces the power to detect time dependence significantly. This finding is supported by a test of proportional hazards using a Cox model containing treatment alone based on the Schoenfeld residuals (Grambsch and Therneau, 1994), (P-value: 0.76). In all three cases we revert to the constant shape parameter Weibull model.

We find that as expected all three models suggest that there is negative VE. Under the constant shape parameter Weibull model we test for evidence of surrogate quality it two ways, $H0_2$. First, we test surrogate values via the null $VE(s_{low}) - VE(s_{high}) = 0$, where s_{low} and s_{low} are the 95th percentile and 5th percentile of the potential surrogate of interest. We find that there is no evidence to support the log of T-cell magnitude as a partial surrogate based on the difference in the VE between the 95th and 5th of the potential surrogates for the Pol, Gag or Nef protein specific responses, (all P-values greater than 0.8). The P-values associated with Wald tests based on the null $\gamma^*_{11} - \gamma^*_{01} = 0$ are lower, (0.13, 0.08, 0.15) for Gag, Nef and Pol, respectively. This suggests weak evidence of partial surrogacy and that there may be reason to use this test over the test based on VE in time-independent scenarios.

Figure 7.1 left panel replicates Figure 1 from Huang and Gilbert (2011) displaying the quantile function of risk difference, $\{risk_1(s_1) - risk_0(s_1)\}$, with risks based on the CDF at 1.5 years of follow-up, as this was the average follow-up time in the data set and gave us the estimated 0.017 difference in infection rate matching the empirical difference in infection rate at the end of the trial. We use this direction of risk difference and the surrogate dependent relative risk, $RR(s_1) = 1 - VE(s_1)$ to reflect the negative vaccine effect. Due to the positive effect of vaccine on the hazard the positive coefficients associated with each of the potential surrogates indicates that risk difference will decrease with increased potential surrogate values and that RR will become smaller.

Although we find the same direction of effect and the same indication that risk difference may vary over $S1_{Nef}$ more than over $S1_{Gag}$, $S1_{Pol}$, we find a greater range in general for risk difference spanning from (-0.2) to (0.05) than Huang and Gilbert (2011) who found a range of (-0.04 to 0.04). Figure 7.1 right panel displays the estimated $RR(s_1)$ over the three potential surrogates. Although there is variation in the $RR(s_1)$ due to the low number of events there is very little power to detect surrogate value as well as the poor BIP, the bootstrap SE for points on the curve are high. The low precision and large SE are also found in Huang and Gilbert (2011) when using the binary outcome. The right most panel of 7.1 depicts the estimated $RR(1.5|v+) = \frac{1-PPV_1'(1.5|v)}{1-PPV_0'(1.5|v)}$ curve. This suggests that there are no clear thresholds for any of the three potential SoP.

Table 7.2: Summary Statistics: parametric EML analysis Step

	Nef		Gag		Pol	
Statistic	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
$\widehat{STG}(1.5)$	0.595	(0.011, 2.23)	0.532	(0.012,1.84)	0.44	(0.011,1.75)
$\widehat{pTG}(1.5\|0.95, 0.05)$	0.373	(0.009,1.11)	0.338	(0.029, 1.17)	0.28	(0.01, 1.42)
$\widehat{PPV}(1.5\|0.85)$	0.035	(0.02, 0.093)	0.033	(0.019, 0.064)	0.031	(0.016,0.12)
$\widehat{NPV}(1.5\|0.05)$	0.997	(0.71, 1)	0.975	(0.527, 1)	0.99	(0.58, 1)

As we assume monotonicity in treatment effect in the direction of enhancement, PPV is the probability of enhancement given that the risk model predicts enhancement. All the PPV estimates

are greater than the estimated average risk difference at 1.5 years and in all but the Pol risk model the quantile based bootstrap confidence intervals do not contain that estimate. This suggests there is weak evidence that the risk model predicts enhancement slightly better than average, for the Nef and Gag candidate SoP. In this setting, NPV is the probability of non-enhancement given the that the risk model predicts no enhancement. All the NPV estimates are similar to 0.983, one less average risk difference, and all CI cover this amount. This suggests that the risk model does not improve on the average for predicting subjects that will not have enhancement. The overlapping confidence intervals for all the candidates for both NPV and PPV suggest there is no statistical evidence to support one of the candidates is significantly better at classifying subjects.

The estimated STGs correspond to maximum sensitivity and specificity of 1.59, 1.53 and 1.44 for Nef, Gag and Pol respectively. These suggest marginal surrogate value. The comparison P-values for difference between the $STG(1.5)$ values of Pol, Gag and Nef are all greater than 0.9, just as found in Huang and Gilbert (2011). This suggests that although Nef looks like a better SoP, there is no evidence to support that difference in these data based on the $STG(1.5)$. This lack of evidence is also supported by the overlapping CI of all the candidate SoP.

A comparison of $pTG(1.5|0.95, 0.05)$ over the candidates, suggests that there is not evidence to support the superiority of any one candidate SoP in these data based on the pTG, as the CI overlap for all and all contrast P-values are > 0.9. We performed the same analysis using the semi-parametric EML method and found very similar results.

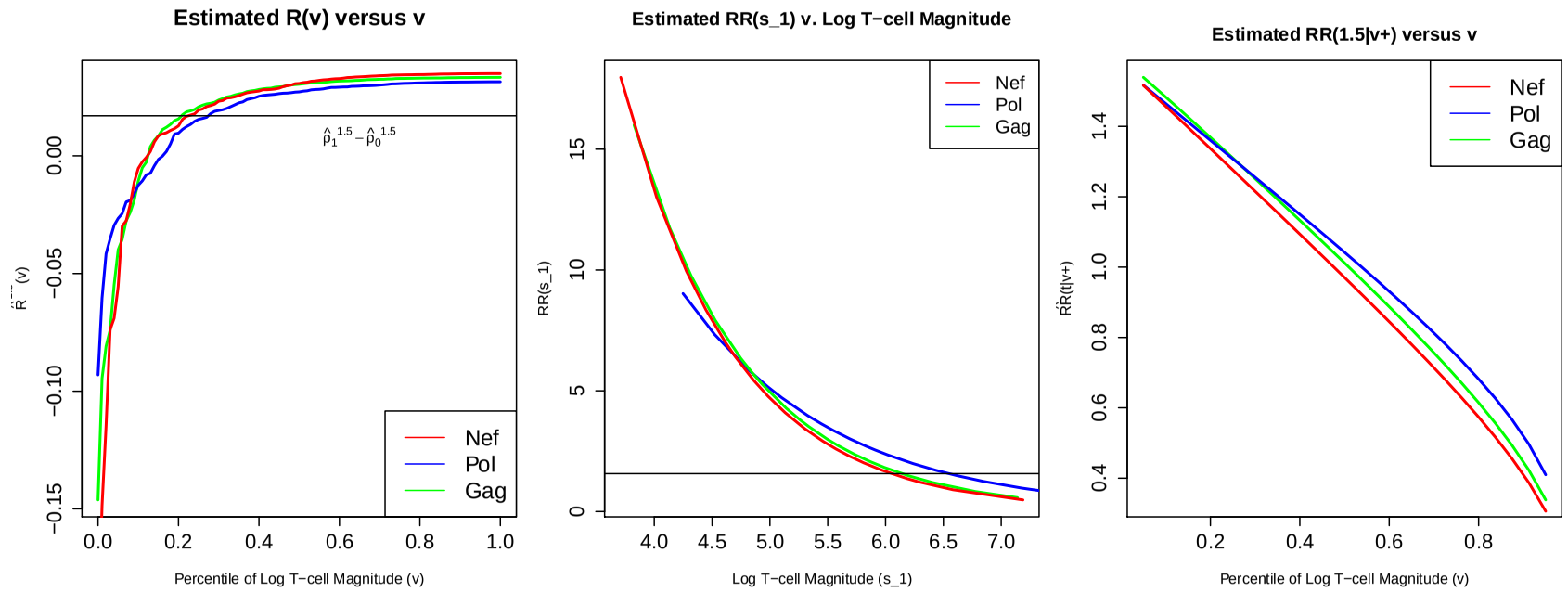

Figure 7.1: Parametric EML: Left most panel depicts the estimated quantile curve of risk difference $\{risk_1(1.5|s_1) - risk_0 1.5|s_1\}$ for each of the three potential SoP; the year 1.5 clinical outcome prevalence difference between arms is the black horizontal line. The middle panel depicts the estimated relative risk, $\widehat{RR}(1.5|s_1)$ for each of the three potential SoP; the average RR is given by the black horizontal line. The right most panel depicts the estimated relative risk at or above given thresholds of the three potential surrogates,$RR(1.5|v+) = \frac{1-PPV_1'(1.5|v)}{1-PPV_0'(1.5|v)}$.

7.1.2 Semi-parametric EML Analysis of Step

We fit the same three saturated Weibull models semi-parametrically using as starting values the estimates obtained from the parametric model, as is suggested in Huang and Gilbert (2011). In each model there were several iterations around the starting values before converging to very similar values. Again, we found no evidence of time-dependence in any of the three models; P-values of (0.370, 0.310, 0.305) for the models containing $S\,1_{Gag}$, $S\,1_{Pol}$, $S\,1_{Nef}$. We again revert to the constant shape parameter Weibull in all three cases.

We find that as expected all three models suggest that there is negative VE. Tests for $H0_2$ via testing the null $VE(s_{low}) - VE(s_{high}) = 0$, where s_{high} and s_{low} are the 95th percentile and the 5th percentile of the potential surrogate of interest yield P-values all greater 0.8 which is similar to the parametric analysis. Wald tests of the null $\gamma^*_{11} - \gamma^*_{01} = 0$ again have lower P-values, (0.34, 0.26, 0.4) for Gag, Nef and Pol, respectively. This suggests that Nef is the most likely of the three to have partial surrogate value. However, these P-values are larger than those found in the parametric analysis suggesting that the parametric analysis may have been more efficient in this setting.

Figure 7.2 left panel replicates Figure 1 from Huang and Gilbert (2011) displaying the quantile function of risk difference, $\{risk_1(s_1) - risk_0(s_1)\}$. We again find a greater range for risk difference spanning from (-0.3) to (0.1). The summary statistic findings are also very similar to those for the parametric model, they are displayed in Table 7.3.

Table 7.3: Summary Statistics: semi-parametric EML analysis Step

	Nef		Gag		Pol	
Statistic	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
$\widehat{STG}(1.5)$	0.6	(0.011, 2.23)	0.537	(0.012,1.84)	0.425	(0.011,1.75)
$\widehat{pTG}(1.5\|0.95, 0.05)$	0.383	(0.007, 1.43)	0.343	(0.011, 0.971)	0.271	(0.009, 1.12)
$\widehat{PPV}(1.5\|0.85)$	0.035	(0.019, 0.092)	0.033	(0.018,0.06)	0.031	(0.017,0.13)
$\widehat{NPV}(1.5\|0.05)$	0.998	(0.71, 1)	0.978	(0.51, 1)	0.991	(0.58, 1)

These findings add little information to those of Huang and Gilbert (2011). However, the simi-

larity of those results to these as well the two EML methods results to each other provides evidence that these methods are functioning properly.

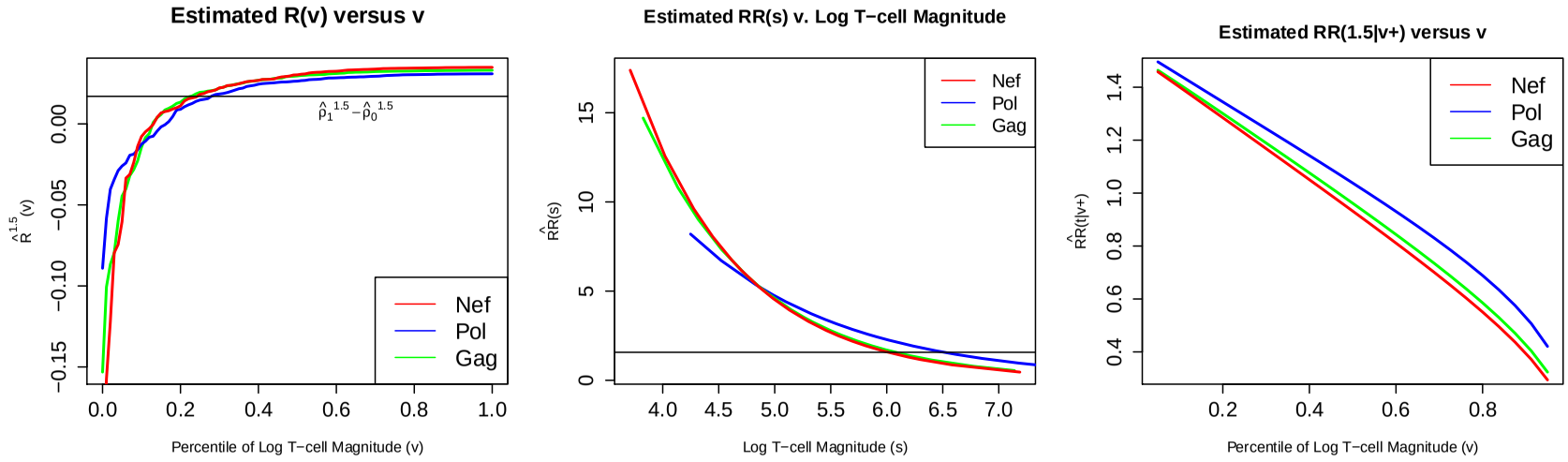

Figure 7.2: Semi-parametric EML: Left most panel depicts the estimated quantile curve of risk difference $\{risk_1(1.5|s_1) - risk_0 1.5|s_1\}$ for each of the three potential SoP; the year 1.5 clinical outcome prevalence difference between arms is the black horizontal line. The middle panel depicts the estimated relative risk, $\widehat{RR}(1.5|s_1)$ for each of the three potential SoP; the average RR is given by the black horizontal line. The right most panel depicts the estimated relative risk at or above given thresholds of the three potential surrogates,$RR(1.5|v+) = \frac{1-PPV_1'(1.5|v)}{1-PPV_0'(1.5|v)}$.

7.2 Zoster Efficacy and Safety Trial (ZEST)

Zostavax® was approved for use in people 60 years of age or greater due to the estimated 51% VE in the first Phase III trial (Oxman et al., 2005). The second Zostavax trial randomized 22,439 subjects ages 50-59 in a one-to-one ratio of vaccination and placebo. The primary analysis of the trial found an estimated 70% VE. These results were recently published in Schmader et al. (2012) This VE, although high, is only based on 129 confirmed cases.

Within the immunogenicity sub-sample (IS) of the ZEST trial, week 6 post-vaccination gpELISA antibody titers were measured for a sub-sample of vaccine recipients and day one titers were measured for a sub-sample of all subjects both using Elispot. For subjects in the IS there was an average follow-up time of approximately 1.3 years with a maximum of 1.7 years. There were 1424 females, 852 males. Table 7.4 displays a summary of the data available from the study and the distribution of the demographic variables by vaccination arm and the sampling of the potential surrogate, post-vaccination Zoster titers, and the BIP, pre-vaccination Zoster titers.

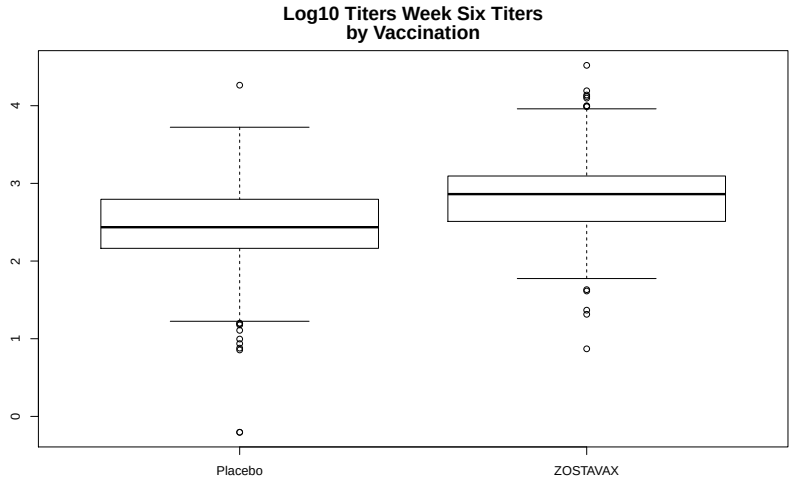

Figure 7.3: The figure depicts the boxplots for the two potential SoP log base 10 week six titers

Table 7.4: Date Summary Table: ZEST trial

Var	Vaccinees	Vaccinees with S + W	Placeboees	Placeboees with W
Var	N(%)/Mean(SD)	N(%)/Mean(SD)	N(%)/Mean(SD)	N(%)/Mean(SD)
Number of subjects	11211(100%)	1179(100%)	11228(100%)	1273(100%)
Infected	30(0.3%)	24(2%)	99(0.9%)	95(7.5%)
Age (years)	54.86(2.77)	54.9(2.78)	54.81(2.77)	54.88(2.71)
Male gender	4298(38.3%)	439(37.2%)	4256(37.9%)	465(36.5%)
Nonwhite race	623(5.6%)	68(5.8%)	627(5.6%)	70(5.5%)
Country				
BEL	210(1.9%)	24(2%)	211(1.9%)	20(1.6%)
CAN	203(1.8%)	29(2.5%)	205(1.8%)	28(2.2%)
DEU	1021(9.1%)	98(8.3%)	1024(9.1%)	102(8%)
FIN	3778(33.7%)	381(32.3%)	3777(33.6%)	421(33.1%)
USA	5999(53.5%)	647(54.9%)	6011(53.5%)	702(55.1%)
Week 6 titers measured	1179(10.5%)	1179(100%)	0(0%)	0(0%)
Week 6 titers	6.49(0.92)	6.49(0.92)	–(–)	–(–)
Day one titers measured	1218(10.9%)	1177(99.8%)	1273(11.3%)	1273(100%)
Day one titers	521.21(805.16)	525.41(810.79)	504.91(621.29)	504.91(621.29)

7.2.1 Zoster Example Method One: Not Leveraging Other Baseline Covariates

We fit the saturated Weibull model using log base 10 of week 6 post vaccination titers denoted by $S(1)_6$ and the difference between log base 10 titers at day one and log base 10 titers at week 6 dented by $S(1)_{diff}$, log of day one titers as W, vaccination as Z, confirmed outbreak as δ and the time to outbreak in years as T. We assumed that negative values of $S(1)_{diff}$ were due to measurement error and set those values to zero. We used full trial data as our complete set with the IS sample as our case-control selected BIP sub-sample. We performed 500 bootstraps. We assumed a bivariate normal distribution for the potential surrogate and the BIP, and fit the conditional model of log titers at

week six given log titers at day zero via maximum likelihood. This assumption was supported by the investigation of the qq-plots of both variables. When subjects were missing both post-vaccination titers and titers at day zero we estimated their likelihood by integrating over the marginal distribution of S(1). When subjects had day zero titers, but were missing post-vaccination titers we used the original EML parametric described in Chapter 6 to estimate their likelihood contribution. This is what we refer to as method one in the sub-sampling of BIP simulation section.

For potential SoP $S(1)_{diff}$ we found no evidence of time-dependence, rejecting $H0_1$ (P-value: 0.782), this is supported by a test of proportional hazards using a Cox model containing treatment alone based on the Schoenfeld residuals (Grambsch and Therneau, 1994). We therefore revert to the proportional hazards model allowing for a constant shape parameter. There is evidence that $S(1)_{diff}$ has surrogate value for VE. We reject $H0_2$ via $\gamma^*_{11} - \gamma^*_{01} = 0$, the scale interaction term of treatment and the potential surrogate in from the constant shape model, (P-value: <0.001). Figure 7.4 depicts the estimated $VE(s_1)$ with 95% bootstrap CI. Although the VE point estimate are slightly negative when $S(1)_{diff} = 0$, the CIs never excludes a positive VE at that level. The bootstrap CIs are larger at lower levels $S(1)_{diff}$, this may be due to the smaller amount of data support in that range. Figure 7.4 depicts a high quality SoP that may satisfy causal necessity, as there is no evidence to support rejection of $VE(s_1 = 0)0$. VE is also highly variable over the range of $S(1)_{diff}$.

For potential SoP $S(1)_6$ we found no evidence of time-dependence, rejecting $H0_1$ (P-value: 0.546). We again revert to the proportional hazards model allowing for a constant shape parameter. There is not evidence in these data to support $S(1)_6$ has surrogate value for VE. We cannot reject $H0_2$ via the null $\gamma^*_{11} - \gamma^*_{01} = 0$ (P-value: 0.201). VE based tests had no additional evidence of surrogacy. Figure 7.4 depicts the estimated $VE(s_1)$ with 95% bootstrap CI. There is no suggestion of causal necessity and there is not very much variation in VE over $S(1)_6$. Looking at the boxplots in Figure 7.3 we see that the placebo levels of log base 10 week six titers are in the range of the lower levels of $S(1)_6$ observed, therefore it is not a matter of protection at all levels, but rather a lack of causal necessity in this case.

Table 7.5: Summary Statistics: ZEST trial

	$S(1)_6$		$S(1)_{diff}$	
Statistic	Estimate	95% CI	Estimate	95% CI
$\widehat{STG}(1.5)$	0.11	(0.012, 0.379)	0.512	(0.304, 0.774)
$\widehat{pTG}(1.5\|0.85, 0.05)$	0.031	(0.002, 0.112)	0.109	(0.065, 0.156)
$\widehat{PPV}(1.5\|0.85)$	0.079	(0.01, 0.165)	0.175	(0.116, 0.242)

All statistics are based on a CEP of risk difference at the given quantile and time, 1.5 years. Risk for all measures was the cumulative risk prior to 1.5 years. $\widehat{pTG}(1.5|0.85, 0.05)$ is standardized. Bootstrap CI are quantile based.

The summary statistic $\widehat{STG}(1.5)$ has an appealing interpretation based on the classification accuracy measures. We believe this assumption to be supported in these data. Therefore, the 0.12 estimate of $STG(1.5)$ for $S(1)_6$ can be interpreted as a $max_c\{\text{Sensitivity}(t|c) + \text{Specificity}(t|c)\}$ of 1.11. This does not suggest a very strong SoP relationship. However, the 0.512 estimate of $STG(1.5)$ for $S(1)_{diff}$ can be interpreted as max Sensitivity plus Specificity of 1.512, which suggests a stronger surrogate relationship. The barely overlapping CI of $STG(1.5)$ based on $S(1)_6$ and $S(1)_{diff}$ is evidence that the difference is a better SoP than $S(1)_6$, the P-value of 0.045 for the difference in $STG(1.5)$s being zero also suggests that $S(1)_{diff}$ is a stronger SoP.

The estimated standardized $pTG(1.5|0.75, 0.05)$ of 0.109, (95% CI 0.065, 0.155) for $S(1)_{diff}$ in comparison to the estimated $pTG(1.5|0.75, 0.05)$ of 0.031, (95% CI 0.002, 0.113) for $S(1)_6$ suggests again that $S(1)_{diff}$ is a higher quality surrogate. However, as the CI overlap and the P-value for the estimated difference being zero is 0.09 there is no additional evidence given the estimated $pTG(1.5|0.75, 0.05)$s to support $S(1)_{diff}$ as a significantly better SoP.

The 0.175 estimate of $PPV(1.5|0.85)$ for $S(1)_{diff}$, (95% CI 0.116, 0.242) can be interpreted as the estimated probability of protection when the estimated risk difference is greater than or equal to 85th percentile of estimated risk difference. The estimates suggest that the vaccine will positively affect individual outcomes; this is much higher than the empirical average risk difference of 0.006

and the 95% CI excludes this value. For $S(1)_6$, the estimate of $PPV(1.5|0.85)$ of 0.079 and (95% CI 0.010, 0.165) that does not cover 0.006, again suggests increased predictive power over the average probability of protection. The $NPV(1.5|0.05)$ was high for both potential SoP due to the very low difference clinical outcome prevalence between the two arms, so it was informative for comparison. However, the CI of $S(1)_{diff}$ suggest evidence of causal necessity; these are not displayed.

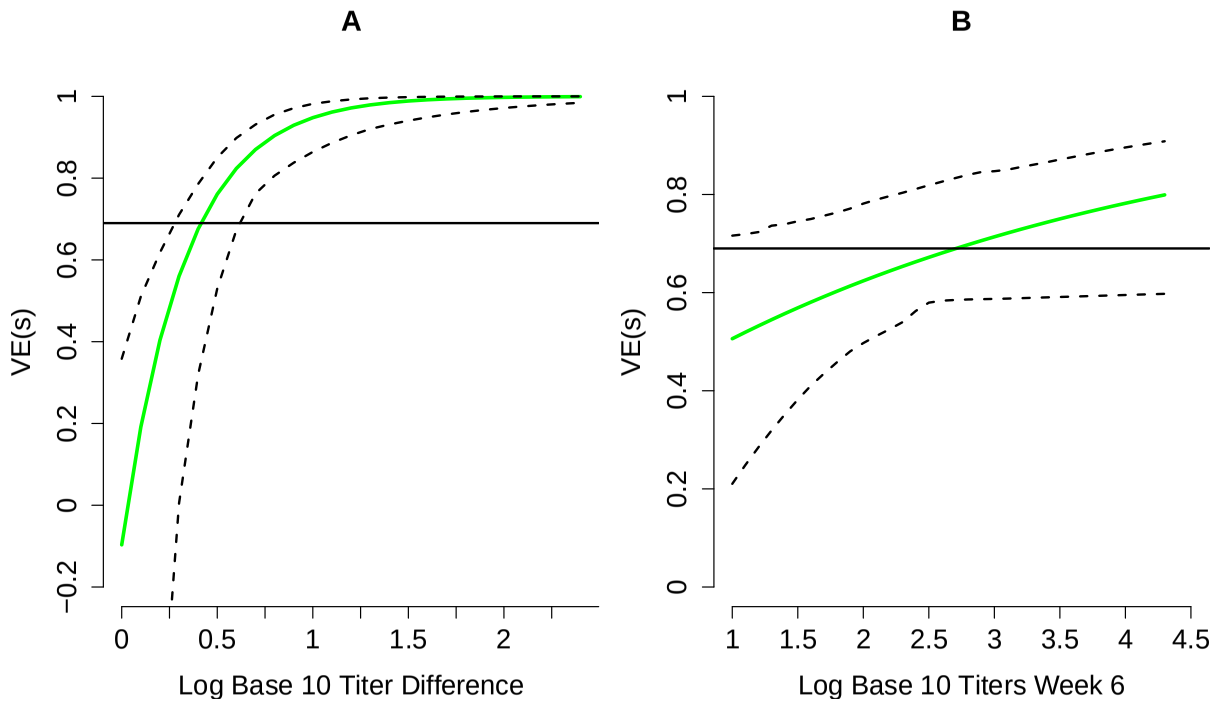

Figure 7.4: Figure A depicts the estimated $VE(s_1)$ over levels of $S(1)_{diff}$ with point-wise bootstrap confidence intervals. Figure B depicts the estimated $VE(s_1)$ over levels of $S(1)_6$ with point-wise bootstrap confidence intervals.

Chapter 8

DISCUSSION AND CONCLUSION

8.1 Discussion

In the recent work Pearl (2011), the usefulness of principal stratification as a tool for furthering science and investigation of the truth is called into question. We believe however, that the estimands introduced in this dissertation fall into the category of true scientific interest as was pointed out by Gilbert et al. (2011c). Principal stratification is basically a subgroup analysis or a study of VE modifies. It is a tool that allows us to estimate scientifically interesting estimands that we might not otherwise be able to investigate. Particularly in the vaccine setting, where CB is common, principal stratification allows for causal comparisons to be made that would otherwise not be possible with other techniques.

Specific SoPs are important targets for Phase I and IIa trials. Without good endpoints for evaluation and comparison of candidate vaccines, advancement is slowed, increasing the risk of poor vaccines being advanced to Phase III and potentially abandoning effective formulations prior to Phase III. There are very few SoP evaluation methods that allow for a time-to-event clinical endpoint and to our knowledge none that allow for or characterize the time-dependent effects of the vaccine. There is evidence of waning in assay levels and in VE in many recent vaccine trials. Methods of SoP evaluation that do not allow for or characterize time-varying effects may classify potential SoP as high quality ignoring their lack of durability or dismiss high quality surrogates in trials that have rapid waning in VE. Our methods allow us to investigate VE as a function of time in subgroups based on a potential SoP.

All of our estimation methods make assumptions about the data, some of which can be tested using the data and some that are not testable. Assumption A3, equal individual risk until time τ, is an untestable assumption that was relaxed in the Wolfson and Gilbert (2010) and should be evaluated logically on a case-by-case basis. Preforming sensitivity analysis for the violation of A3 when using the methods from this dissertation is always encouraged. Other testable assumptions

like the distribution of $S(1)|W$ for the parametric EML can be evaluated in the observed data of a BIP augmented trial, if fill sampling is not preformed. As pointed out above, A4-P can be evaluated if CPV is performed and A5 and A6 hold. No close out failures, A6, can be tested and observed in trials that conduct CPV. Minor violations to A6 should be considered via sensitivity analysis, rather than abandoning the CPV preformed. As the assumptions made may vary with the data structure, all assumption should be clearly outlined, and either validation or sensitivity analysis preformed, in any analysis using the methods developed in this dissertation.

8.2 Conclusions

The saturated Weibull model 6.1 allows for the characterization of time-varying effects in the time-to-event setting. All of the main estimation methods presented in this dissertation for this model seem to be adequate under many plausible scenarios for SoP evaluation. The pseudoscore method of CoR analysis also allows for the identification of CoRs that can then be considered as candidate SoPs. Even in the CoR analysis setting, not allowing for time variation can change the understanding of the association between the CoR and outcome; this was seen in the RV144 example.

The findings from the Step example, although very similar to the findings of Huang and Gilbert (2011), do suggest that although the clinical endpoint was described as infection prior to three years that the risk defined by the CDF or the hazard is greater at three years than that based on the observed number of events. This suggests if we had included the the follow-up information after unblinding and there had been no drop-out, the event rate would be expected to be quite a bit higher. Although no strong evidence was found, our analysis of the Step data suggested potential partial surrogate value of magnitude of response to Nef specific epitopes. This suggests, at the least, that measurement of response to Nef specific epitopes in future HIV vaccine trials of similar vaccines is warranted, although more study is need before this biomarker should be considered as a Phase I or IIa endpoint.

The findings from the Weibull analysis of the ZEST trial suggest that log base 10 Zoster titers 6 weeks after vaccination less log base 10 Zoster titers at day 1 is a strong SoP for VE, not just a partial SoP. There is also evidence to support this titer difference is a significantly better SoP than log base 10 Zoster titers 6 weeks alone. This suggests that it is the boost to Zoster titers from

vaccine rather than the post vaccine titer level that is a good predictor of VE, as VE varies greatly over the Zoster titer difference. There is no evidence of time-variation in the treatment effects for either candidate SoP, which may suggest reasonable durability of VE and titer difference as an SoP. Longer follow-up is needed to assess this fully. There is very little evidence to support log base 10 Zoster titers 6 weeks after vaccination as a partial surrogate.

8.3 Future Work

8.3.1 Work in Progress

We are currently working to implement estimation of the variance forms found for the CoR and SoP evaluation methods using pseudoscore. These estimated variances should allow for greater computational efficiency, by removing the need to bootstrap. It is hoped that pseudoscore methods can also be extended to allow for continuous baseline variables. Investigation into the application of pseudo-likelihood ratio tests similar to those developed in Chen and Fan (2005) is also underway and it is hoped that these tests will have better finite sample properties than the Wald tests used in the large sample simulations sets. We also hope to investigate a comparison of this method to that of Breslow and Wellner (2007) when there is no evidence of time-dependence.

Further investigation of the proposed extension of the Gilbert and Hudgens (2008) method of BIP sub-sampling leveraging other baseline variables and the proposed semi-parametric EML method allowing for BIP sub-sampling is also underway. It is hoped that these extensions will allow for the core methods developed in this dissertation to be applied to a wider range of data.

8.3.2 Concepts for Future Work

It is our hope that the methods in this dissertation can be extended for use in evaluation of general surrogates of protection. It is also our hope to frame the meta-analysis methods of GSoP evaluation in a causal manner and reconcile the meta-analysis framework, the Pearl framework and the causal framework for GSoP evaluation; showing their similarities and differences and determining which framework is best suited for the analysis of vaccine trial data.

We have an interest in pursuing the use of extensions to the methods developed here to estimate treatment effects, possibly time-varying treatment effects, in the presence of non-compliance. The

principal stratification framework has often been used in the compliance literature, where the counterfactual values are defined over the compliance strata, rather than the values of a post-treatment biomarker. This should be an interesting application of the time-dependent estimands, as allowing for variation over time and compliance may suggest very different conclusions than the time constant estimands.

BIBLIOGRAPHY

RP A'Hern, SR Ebbs, and MB Baum. Does chemotherapy improve survival in advanced breast cancer? a statistical overview. *Br J Cancer*, 57(6):615–618, 06 1988.

LQ Araujo, CR Macintyre, and C Vujacich. Epidemiology and burden of herpes zoster and postherpetic neuralgia in australia, asia and south america. *the Journal of the IHMF*, 14:40–44, 2007.

William E. Barlow. Robust variance estimation for the case-cohort design. *Biometrics*, 50(4):pp. 1064–1072, 1994.

William E. Barlow, Laura Ichikawa, Dan Rosner, and Shizue Izumi. Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, 52(12):pp. 1165–1172, 1999.

Ornulf Borgan, Bryan Langholz, SvenOve Samuelsen, Larry Goldstein, and Janice Pogoda. Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6:39–58, 2000.

N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75 (1):pp. 11–20, 1988.

Norman E. Breslow and Richard Holubkov. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):447–461, 1997.

Norman E. Breslow and Jon A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102, 2007.

Norman E. Breslow, Thomas Lumley, Christie M. Ballantyne, Lloyd E. Chambless, and Michal Kulich. Improved horvitz-thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*, 1, 2009.

SP Buchbinder, DV Mehrotra, A Duerr, DW Fitzgerald, R Mogg, D Li, PB Gilbert, JR Lama, M Marmor, C del Rio, MJ Mcelrath, DR Casimiro, KM Gottesdiener, JA Chodakewitz, L Corey, and MN Robertson. Efficacy assessment of a cell-mediated immunity hiv-1 vaccine (the step study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *The Lancet*, 372 (9653):1881–1893, 2008a.

Susan P Buchbinder, Devan V Mehrotra, Ann Duerr, Daniel W Fitzgerald, Robin Mogg, David Li, Peter B Gilbert, Javier R Lama, Michael Marmor, Carlos del Rio, M Juliana McElrath, Danilo R Casimiro, Keith M Gottesdiener, Jeffrey A Chodakewitz, Lawrence Corey, and Michael N Robertson. Efficacy assessment of a cell-mediated immunity hiv-1 vaccine (the step study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *The Lancet*, 372(9653):1881 – 1893, 2008b.

Efstathia Bura and Joseph L. Gastwirth. The binary regression quantile plot: Assessing the importance of predictors in binary regression visually. *Biometrical Journal*, 43(1):5–21, 2001.

M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67, 2000.

R. J. Carroll, Suojin Wang, and C. Y. Wang. Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 90(429):pp. 157–169, 1995.

Nilanjan Chatterjee. *A Pseudoscore Estimator for Regression Problems With Two-Phase Sampling*. Doctor of philosophy dissertation, University of Washington; Department of Statistics, Chair: Wellner, Jon and Norman Breslow, 1999.

Nilanjan Chatterjee and Yi-Hau Chen. A semiparametric pseudo-score method for analysis of two-phase studies with continuous phase-i covariates. *Lifetime Data Analysis*, 13:607–22, 2007.

Nilanjan Chatterjee, Yi-Hau Chen, and Norman E. Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461): 158–68, 2003.

Hua Yun Chen and Roderick J. A. Little. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94(447):896–908, 1999.

Xiaohong Chen and Yanqin Fan. Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *Canadian Journal of Statistics*, 33(3):389–414, 2005.

ML Clements-Mann. Lessons for aids vaccine development from non-aids vaccines. *AIDS Research Human Retroviruses*, Supplement 3:197–203, 1998.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Michael J. Daniels and Michael D. Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16(17):1965–1982, 1997.

Abhijit Dasgupta, Steven G. Self, and Somesh Das Gupta. Non-identifiable parametric probability models and reparametrization. *Journal of Statistical Planning and Inference*, 137(11):3380–93, 2007.

AJ Dunning. Comment on evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine*, 27(29):6268–6270, 2008.

Thomas R. Fleming. Surrogate markers in aids and cancer trials. *Statistics in Medicine*, 13(13-14): 1423–1435, 1994.

TR Fleming and DL Demets. Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine*, 125(7):605–613, 1996.

MN Flynn, DN Forthal, CD Harro, FN Judson, KH Mayer, MF Para, PB Gilbert, MG Hudgens, BJ Metch, SG Self, PW Berman, DP Francis, M Gurwith, WL Heyward, DV Jobes, ML Peterson, V Popovic, FM Sinangil, M Gurwith, DV Jobes, ML Peterson, FM Sinangil, PW Berman, DP Francis, WL Heyward, PB Gilbert, MG Hudgens, BJ Metch, SG Self, A Adamczyk, RL Baker, D Brand, SJ Brown, S Buchbinder, BP Buggy, J Cade, MC Caldwell, C Celum, C Creticos, RA Coutinho, K Lindenburg, P Daly, E DeJesus, R Di-Carlo, M Fenstersheib, N Flynn, D Forthal, B Gripshover, GJ Gorse, R Belshe, H Grossman, CD Harro, K Henry, RG Hewitt, R Hogg, JM Jacobson, J Jemsek, F Judson, JO Kahn, MC Keefer, H Kessler, B Koblin, J Kostman, M Lally, K Logue, M Marmor, K Mayer, D McKinsey, BM Miskin,

JO Morales, MJ Mulligan, RA Myers, R Novak, M Para, P Piliero, R Poblete, F Rhame, S Riddler, RW Richter, JH Sampson, M Sands, S Santiago, C Shikuma, MS Somero, E Thomas, M Thompson, SK Tyring, J Vincelette, PS Vrooman, BG Yangco, and rgp120 HIV Vaccine Study Grp. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent hiv-1 infection. *Journal of Infectious Diseases*, 191(5):654–665, 2005.

D Follmann. Augmented designs to assess immune response in vaccine trials. *Biometrics*, 62(4): 1161–1169, 2006.

CE Frangakis and DB Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Laurence S. Freedman, Barry I. Graubard, and Arthur Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11(2):167–178, 1992.

MH Gail, R Pfeiffer, HC. van Houwelingen, and RJ Carroll. On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1(3):231–246, 2000.

PB Gilbert and MG Hudgens. Evaluating candidate principal surrogate endpoints. *Biometrics*, 64 (4):1146–1154, 2008.

PB Gilbert, ML Peterson, D Follmann, MG Hudgens, DP Francis, M Gurwith, WL Heyward, DV Jobes, V Popovic, SG Self, F Sinangil, D Burke, and PW Berman. Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of hiv-1 infection in a phase 3 hiv-1 preventive vaccine trial. *Journal of Infectious Diseases*, 191(5):666–677, 2005.

PB Gilbert, L Qin, and SG Self. Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine*, 27(23):4758–4778, 2008.

PB. Gilbert, JO. Berger, D Stablein, S Becker, M Essex, SM. Hammer, JH. Kim, and VG. DeGruttola. Statistical interpretation of the rv144 hiv vaccine efficacy trial in thailand: A case study for statistical issues in efficacy trials. *Journal of Infectious Diseases*, 203(7):969–975, 2011a.

PB Gilbert, D Grove, E Gabriel, Y Huang, G Gray, SM Hammer, SP Buchbinder, J Kublin, L Corey, and SG Self. A sequential phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple hiv vaccine regimens. *Statistical Communications in Infectious Diseases*, 3(1), 2011b.

Peter B. Gilbert, Michael G. Hudgens, and Julian Wolfson. Commentary on "principal stratification a goal or a tool?" by judea pearl. *The International Journal of Biostatistics*, 7(1), 2011c.

Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):pp. 515–526, 1994.

Wen Gu and Margaret Pepe. Measures to summarize and compare the predictive capacity of markers. *The International Journal of Biostatistics*, 5(1):1–30, 2011.

Barton F. Haynes, Peter B. Gilbert, M. Juliana McElrath, and et al. Immune-correlates analysis of an hiv-1 vaccine efficacy trial. *New England Journal of Medicine*, 366(14):1275–1286, 2012.

Patrick J. Heagerty, Thomas Lumley, and Margaret S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–44, 2000.

PJ Heagerty and MS Pepe. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551, 1999.

M. Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1), 2010.

R Edgar Hope-Simpson. The nature of herpes zoster: A long-term study and a new hypothesis. *Journal of the Royal Society of Medicine*, 58(1):9–20, 1965.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):pp. 663–685, 1952.

Y Huang and PB Gilbert. Comparing biomarkers as principal surrogate endpoints. *Biometrics*, 2011.

Y. Huang, P. B. Gilbert, and J. Wolfson. Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics*, Under Review(Submitted), 2012a.

Ying Huang, Margaret Sullivan Pepe, and Ziding Feng. Evaluating the predictiveness of a continuous marker. *Biometrics*, 63(4):1181–188, 2007.

Ying Huang, Peter B. Gilbert, and Holly Janes. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*, 2012b. ISSN 1541-0420.

MD Hughes. Evaluating surrogate endpoints. *Controlled Clinical Trials*, 23(6):703–707, 2002.

HVTN. Phase 2, randomized, placebo-controlled trial to evaluate the safety and effect on post-hiv acquisition viremia of a multiclade hiv-1 dna plasmid vaccine followed by a multiclade hiv-1 recombinant adenoviral vector vaccine in hiv- uninfected, adenovirus type 5 neutralizing antibody negative, circumcised men. Study Protocol DAIDS DOCUMENT ID 10753, DAIDS, NIAID, NIH, DHHS, 2009.

MM Joffe and T Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2): 530–538, 2009.

Robert W. Johnson and Andrew S.C. Rice. Pain following herpes zoster: The influence of changing population characteristics and medical developments. *Pain*, 128(12):3–5, 2007.

Yun Li, Jeremy M.G. Taylor, and Michael R. Elliott. A bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*, 66(2):523–531, 2010.

Zhiguo Li, Peter B. Gilbert, and Bin Nan. Weighted likelihood method for grouped survival data in case-cohort studies with application to hiv vaccine trials. *The University of Michigan Department of Biostatistics Working Paper Series*, 70, 2007.

R.J.A. Little and D. B. Rubin. *Statistical analysis with Missing Data*. Wiley, 2nd edition, 2002.

M Juliana McElrath, Stephen C De Rosa, Zoe Moodie, Sheri Dubey, Lisa Kierstead, Holly Janes, Olivier D Defawe, Donald K Carter, John Hural, Rama Akondy, Susan P Buchbinder, Michael N Robertson, Devan V Mehrotra, Steven G Self, Lawrence Corey, John W Shiver, and Danilo R Casimiro. Hiv-1 vaccine-induced immunity in the test-of-concept step study: a casecohort analysis. *The Lancet*, 372(9653):1894 – 1905, 2008.

Chaya S. Moskowitz and Margaret S. Pepe. Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Statistics in Medicine*, 23(10):1555–1570, 2004a.

Chaya S. Moskowitz and Margaret S. Pepe. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*, 5:113–127, 2004b.

MJ Mulligan. Advances in human clinical trials of vaccines to prevent hiv/aids and other hiv prevention interventions. *Current Infectious Disease Reports*, 11(5):399–406, 2009.

K Murphy, P Travers, and M Walport. *Janeway's Immuno Biology: Seventh Edition*. Garland Science, Taylor and Francis Group, 2008.

J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):pp. 101–116, 1938.

Michael N. Oxman. Zoster vaccine: Current status and future prospects. *Clinical Infectious Diseases*, 51(2):197–213, 2010.

M.N. Oxman, M.J. Levin, G.R. Johnson, K.E. Schmader, S.E. Straus, L.D. Gelb, R.D. Arbeit, M.S. Simberkoff, A.A. Gershon, L.E. Davis, A. Weinberg, K.D. Boardman, H.M. Williams, J. Hongyuan Zhang, P.N. Peduzzi, C.E. Beisel, V.A. Morrison, J.C. Guatelli, P.A. Brooks, C.A. Kauffman, C.T. Pachucki, K.M. Neuzil, R.F. Betts, P.F. Wright, M.R. Griffin, P. Brunell, N.E. Soto, A.R. Marques, S.K. Keay, R.P. Goodman, D.J. Cotton, J.W. Gnann, J. Loutit, M. Holodniy, W.A. Keitel, G.E. Crawford, S.-S. Yeh, Z. Lobo, J.F. Toney, R.N. Greenberg, P.M. Keller, R. Harbecke, A.R. Hayward, M.R. Irwin, T.C. Kyriakides, C.Y. Chan, I.S.F. Chan, W.W.B. Wang, P.W. Annunziato, and J.L. Silber. A vaccine to prevent herpes zoster and postherpetic neuralgia in older adults. *New England Journal of Medicine*, 352(22):2271–2284, 2005.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

J Pearl. Principal stratification — a goal or a tool? *International Journal of Biostatistics*, 7(1), 2011. Article 20.

J Pearl and E Bareinboim. Transportability across studies: A formal approach. Technical Report R-372, University of California, Los Angeles, March 2011.

MS Pepe and TR Fleming. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86(413):108–113, 1991.

P Pitisuttithum, PB Gilbert, M Gurwith, W Heyward, M Martin, F van Griensven, D Hu, JW Tappero, K Choopanya, and Bangkok Vaccine Evaluation. Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 hiv-1 vaccine among injection drug users in bangkok, thailand. *Journal of Infectious Diseases*, 194(12):1661–1671, 2005.

Stanley A. Plotkin. Correlates of vaccine-induced immunity. *Clinical Infectious Diseases*, 47(3): 401–409, 2008.

Stanley A Plotkin. Vaccines: the fourth century. *Clin Vaccine Immunol*, 16(12):1709–19, Dec 2009.

R. L. Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982.

RL Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989.

Ross L. Prentice. On the design of synthetic case-control studies. *Biometrics*, 42(2):301–310, 1986.

B Pulendran, S Li, and HI. Nakaya. Systems vaccinology. *Immunity*, 33(4):516–529, 10 2010.

L Qin, PB Gilbert, L Corey, MJ McElrath, and SG Self. A framework for assessing immunological correlates of protection in vaccine trials. *The Joutnal of Infectious Diseases*, 196(9):1304–1312, 2007.

L Qin, PB Gilbert, D Follmann, and L Dongfeng. Assesing surrogate endpoints in vaccine trails with case-cohort sampling and the cox model. *Annals of Applied Statistics*, 2(1):386–407, 2008.

S Rerks-Ngarm, P Pitisuttithum, S Nitayaphan, J Kaewkungwal, J Chiu, R Paris, N Premsri, C Namwat, M de Souza, E Adams, M Benenson, S Gurunathan, J Tartaglia, J McNeil, D Francis, D Stablein, D Birx, S Chunsuttiwat, C Khamboonruang, P Thongcharoen, M Robb, N Michael,

P Kunasol, J Kim, and the MOPH-TAVEG Investigators. Vaccination with alvac and aidsvax to prevent hiv-1 infection in thailand. *New England Journal of Medicine*, 361(23):2209–2220, 2009.

J M Robins and S Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–55, Mar 1992a.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):pp. 846–866, 1994.

JM Robins and S Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992b.

Bernard Roizman and Philip Pellett. *The Family Herpesviridae: A Brief Introduction*, volume 5, pages 2479–2500. Lippincott Williams & Wilkins, 2007.

DB Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170, 2004.

Johnson RW, G Wasner, P Saddier, and R Baron. Herpes zoster and postherpetic neuralgia: optimizing management in the elderly patient. *Drugs and Aging*, 25(12):991–1006, 2008.

Michael Sachs. *Statistical Methods to Assess the Prospective Predictive Accuracy of a Medical Test*. Doctor of philosophy dissertation, University of Washington; Department of Biostatistics, Chair: Zhou, XH Andrew, 2011.

Thomas H. Scheike and Torben Martinussen. Maximum likelihood estimation for cox's regression model under casecohort sampling. *Scandinavian Journal of Statistics*, 31(2):283–293, 2004.

Kenneth Schmader. Herpes zoster in older adults. *Clinical Infectious Diseases*, 32(10):pp. 1481–1486, 2001.

Kenneth E. Schmader, Gary R. Johnson, Patricia Saddier, Maria Ciarleglio, William W.B. Wang, Jane H. Zhang, Ivan S.F. Chan, Shing-Shing Yeh, Myron J. Levin, Ruth M. Harbecke, Michael N.

Oxman, and for the Shingles Prevention Study Group. Effect of a zoster vaccine on herpes zoster-related interference with functional status and health-related quality-of-life measures in older adults. *Journal of the American Geriatrics Society*, 58(9):1634–1641, 2010.

Kenneth E. Schmader, Myron J. Levin, John W. Gnann, Shelly A. McNeil, Timo Vesikari, Robert F. Betts, Susan Keay, Jon E. Stek, Nickoya D. Bundick, Shu-Chih Su, Yanli Zhao, Xiaoming Li, Ivan S.F. Chan, Paula W. Annunziato, and Janie Parrino. Efficacy, safety, and tolerability of herpes zoster vaccine in persons aged 5059 years. *Clinical Infectious Diseases*, 2012.

Santosh C Sutradhar, William W B Wang, Katia Schlienger, Jon E Stek, Jin Xu, Ivan S F Chan, and Jeffrey L Silber. Comparison of the levels of immunogenicity and safety of zostavax in adults 50 to 59 years old and in adults 60 years old or older. *Clin Vaccine Immunol*, 16(5):646–652, May 2009.

Jeremy M. G. Taylor, Yue Wang, and Rodolphe Thibaut. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61(4):1102–1111, 2005a.

Jeremy M. G. Taylor, Yue Wang, and Rodolphe Thibaut. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61(4):pp. 1102–1111, 2005b.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

A.J. van Hoek, N. Gay, A. Melegaro, W. Opstelten, and W.J. Edmunds. Estimating the cost-effectiveness of vaccination against herpes zoster in england and wales. *Vaccine*, 27(9):1454–1467, 2009.

Tyler J. VanderWeele. Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters*, 78(17):2957 – 2962, 2008. ISSN 0167-7152. doi: 10.1016/j.spl.2008.05.029. URL http://www.sciencedirect.com/science/article/pii/S0167715208002496.

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

JE White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American journal of epidemiology*, 115(1):119, 1982.

J Wolfson. *Statistical Methods for Identifying Surrogate Endpoints in Vaccine Trails*. Doctor of philosophy dissertation, University of Washington; Department of Biostatistics, Chair: Gilbert, Peter, 2009.

J Wolfson and PB Gilbert. Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics*, 66(4):1153–1161, 2010.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

D. Zeng and D. Y. Lin. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 507–64, 2007.

Y Zheng, T Cai, MS Pepe, and WC Levy. Time-dependent predictive values of prognostic biomarkers with failure time outcome. *Journal of the American Statistical Association*, 103(481):362–368, 2008.

Y Zheng, T Cai, JL Stanford, and Z Feng. Semiparametric models of time-dependent predictive values of prognostic biomarkers. *Biometrics*, 66(1):50–60, 2010.

VITA

Erin E Gabriel received a Bachelor of Arts degree with honors in Mathematics in 2002 from Washington College and a Master of Arts degree in Economics from McGill University in 2004. Starting in September 2012, she will be a post-doctoral researcher at the Fred Hutchinson Cancer Research Center under the guidance of Dr. M Elizabeth Halloran.