USAGE OF SURROGATE ENDPOINTS IN THE DESIGN AND ANALYSIS OF CLINICAL TRIALS

by

Judah Abberbock

M.S. Biostatistics, Columbia University, 2013

B.A. Mathematics, Yeshiva University, 2011

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Judah Abberbock

It was defended on

July 17, 2017

and approved by

Gong Tang, Ph.D., Associate Professor

Department of Biostatistics

Graduate School of Public Health, University of Pittsburgh

Stewart Anderson, Ph.D., Professor

Department of Biostatistics

Graduate School of Public Health, University of Pittsburgh

Ying Ding, Ph.D., Assistant Professor

Department of Biostatistics

Graduate School of Public Health, University of Pittsburgh

Priya Rastogi, MD, Associate Professor

Department of Medicine

University of Pittsburgh

Dissertation Director: Gong Tang, Ph.D., Associate Professor

Department of Biostatistics

Graduate School of Public Health, University of Pittsburgh

USAGE OF SURROGATE ENDPOINTS IN THE DESIGN AND ANALYSIS OF CLINICAL TRIALS

Judah Abberbock, PhD

University of Pittsburgh, 2017

ABSTRACT

There has been a shift in the conduct of early-stage breast cancer trials in recent years from long adjuvant trials with overall or disease-free survival as the efficacy endpoint to shorter neoadjuvant trials with pathological complete response (pCR), a binary marker, at time of surgery as the endpoint. The Food and Drug Administration (FDA) currently embraces this transition and deems evidence in pCR improvement sufficient for drug approval on condition that long-term data are collected to eventually show efficacy in survival. Incorporating data on pCR in the design and analysis of such a trial is therefore of public health interest. Here, we propose one method to assess the power and sample size of such a trial with using observed neoadjuvant data and another method to estimate certain causal treatment effects on survival conditional on pCR. In the first part, we propose an exponential mixture model for survival time with parameters for the response rates and an estimated benefit in survival from achieving response. Under a fixed sample size, we obtain the empirical power through simulations from the proposed mixture model. We also propose a more efficient method than the empirical approach by applying an estimated average hazard ratio to the Schoenfeld formula. The performance of our methods is assessed via simulation studies. Data from two neoadjuvant cancer clinical trials are used to illustrate these methods. Second, we propose a method under the principal stratification framework to estimate the causal effect of treatment on a binary outcome, conditional on a post-treatment binary response marker in randomized controlled clinical trials. Specifically, we estimate the treatment effect among

those who would achieve response if given the treatment. We are able to identify this causal effect under two assumptions. First, we model the counterfactual probability of achieving response under treatment given baseline clinical markers and the outcome. Second, we assume a monotonicity condition: a patient who responds under control would respond under treatment as well. We compared the performance of proposed method with other standard approaches in simulation studies. Data from a neoadjuvant breast cancer clinical trial are used to demonstrate the proposed method.

Keywords: binary response marker, survival, power, sample size, neoadjuvant, cancer clinical trials, principal stratification, causal inference.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 OVERVIEW

Testing new drug therapies in general and for breast cancer in particular has become harder as standard treatments have improved over time. Because of this, new drug testing requires larger sample sizes and longer follow-up in order to demonstrate statistical significance for smaller treatment effects. An example of a typical phase III breast cancer trial is the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-40 study where 1206 women with HER2-negative breast cancer were randomized to one of three docetaxel-based neoadjuvant regimens and whether to receive bevacizumab or not. Long-term survival follow-up data were collected from Oct 31, 2007 to March 27, 2014. Median follow-up was 4.7 years. Because large and long trials like this are not always feasible, researchers often prefer to find an intermediate endpoint which can accurately predict the drug's treatment effect on the long-term outcomes before they occur to make new drug trials shorter and less costly. Such an intermediate endpoint is known as a surrogate endpoint.

While there are varying definitions of a surrogate endpoint, we adopt the definition of Wittes, Lakatos and Probstfield as "an endpoint measured in lieu of some other so-called 'true' endpoint" [40]. This definition describes the purpose of a surrogate endpoint, but makes no assumption of the relationship between the surrogate and true endpoint. The type of questions that can be answered about the true endpoint with measurement on the surrogate depends on the relation between treatment, surrogate, and true endpoint.

For this reason, it is important to distinguish between two types of surrogates: a prognostic surrogate and a treatment efficacy surrogate. A prognostic surrogate is statistically associated with the true endpoint of interest. For example, tumor response is a prognos-

tic surrogate for survival if tumor responders have improved survival over non-responders. Prognostic surrogates, while not the focus of this dissertation, are important nonetheless. Prognostic surrogates are useful to inform patients of their likely outcomes well before these outcomes are predicted to occur. This knowledge can help determine the proper course of action for future treatment and/or patient care.

The focus of this dissertation is on the second type of surrogate, the treatment efficacy surrogate. An efficacy surrogate is an intermediate variable whose association with treatment can predict a treatment effect on the true endpoint. Such a surrogate is important in drug development, where investigators would like to predict a treatment effect on a patient important "true" outcome by showing the drug has an effect on the surrogate measure. Running a randomized clinical trial with the surrogate as the primary endpoint has the potential of saving time and resources, as well as accelerating a drug's FDA approval.

However, it is possible and quite common for a surrogate to have little or no efficacy value, even though a proposed mechanism of action seems plausible to the investigator. One famous historical example is ventricular arythmia. Ventricular arythmia is associated with a four-fold increase in mortality due to cardiovascular complication, making it plausible as a surrogate for the "true" endpoint of mortality. However, ventricular arythmia proved to be a poor efficacy surrogate for overall mortality. Of the three drugs approved by the FDA for reducing ventricular arythmias, all 3 eventually showed an increased mortality rate compared to placebo in the Cardiac Arrhythmia Suppression Trial (CAST) [24, 37]. Fleming and Demets [17] have documented these "failures" of surrogate endpoints across a wide range of conditions including cardiologic conditions, cancer, and other diseases. Therefore, unless a surrogate is well-established as an efficacy surrogate, the FDA will not grant approval to the drug with a surrogate endpoint as its primary efficacy outcome.

Furthermore, even a prognostic surrogate is not necessarily a treatment efficacy surrogate for the same "true" endpoint. As an example, consider a drug that shows efficacy on a surrogate response measure. Additionally, patients achieving this surrogate response have a better prognosis than those non-responders, regardless of treatment assignment. Such a drug may still not have an effect on the true endpoint. Such a scenario is depicted in Table 1.

Table 1: Example data of a prognostic surrogate with no overall treatment effect

	Surrogate		True Endpoint	
Treatment (N=40)	Responders	20	Responders	10
			Non-responders	10
	Non-responders	20	Responders	5
			Non-responders	15
Control (N=40)	Responders	10	Responders	5
			Non-responders	5
	Non-responders	30	Responders	10
			Non-responders	20

In this scenario, 50% of subjects on treatment achieve surrogate response compared to 25% of control subjects. Additionally, 50% (15/30) of surrogate responders achieve the true endpoint compared to 30% (15/50) of surrogate non-responders. However, there is no treatment effect on the true endpoint as both groups have 15/40 responders to true endpoint.

For cancer studies with tumor response as a possible surrogate, Anderson et al. [1] caution investigators about the problem of confusing a prognostic surrogate with a treatment efficacy surrogate. They report that "common practice" at the time was for an investigator to present comparisons between tumor responders and non-responders, known as a "responder analysis." If responders survive significantly longer than non-responders investigators would conclude that increasing tumor response would be a way to increase survival. The investigators' error is to equate the prognostic surrogate of tumor response as necessarily implying tumor response is a valid efficacy surrogate. However, as the authors state, "It is generally impossible to refute the possibility that response is just a marker which selects the good prognosis patients: those who would have survived longer even if the therapy has no effect at all."

1.2 METHODS FOR SURROGATE ENDPOINT VALIDATION

Because potential surrogate endpoints may not be treatment efficacy surrogates, there is a need to develop criteria to validate potential surrogates as efficacy surrogates. Only after a surrogate has been validated can the surrogate replace the true endpoint in clinical studies. As reviewed by Molenberghs et al. [29] there have been a number of methods proposed to assess the validity of a potential surrogate.

1.2.1 Prentice's Criteria

Prentice [31] was the first to give a formal definition of a valid surrogate endpoint. He defined such a surrogate as "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." This definition captures the purpose of a surrogate endpoint in clinical trials: to use the surrogate instead of the true endpoint in order to concluded the treatment effective on the true endpoint measure.

Prentice proposed four criteria, that, when true, would validate a surrogate endpoint. For treatment indicator Z, surrogate S, and true endpoint T, the criteria are the following:

$$f(S|Z) \neq f(S) \tag{1.1}$$

$$f(T|Z) \neq f(T) \tag{1.2}$$

$$f(T|S) \neq f(T) \tag{1.3}$$

$$f(T|S, Z) = f(T|S) \tag{1.4}$$

These criteria, in words, say that treatment has a significant effect on both the surrogate (1.1) and true endpoint (1.2), the surrogate has a significant effect on the true endpoint (1.3), and the full effect of treatment is captured by the effect of the surrogate on treatment (1.4)[9].

As an example of the validation procedure, Molenberghs et al. [29] consider a continuous outcome T_j with binary treatment Z_j and binary surrogate S_j for subject j. The following four models can be used to evaluate Prentice's criteria:

$$S_j = \mu_S + \alpha Z_j + \epsilon_{Sj}; \tag{1.5}$$

$$T_j = \mu_T + \beta Z_j + \epsilon_{Tj} \tag{1.6}$$

$$T_j = \mu + \phi S_j + \epsilon_j \tag{1.7}$$

$$T_j = \mu_0 + \beta_S Z_j + \phi S_j + \epsilon, \tag{1.8}$$

where ϵ_{Sj} and ϵ_{Tj} are multivariate normal with mean 0 and variance-covariance:

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix} \tag{1.9}$$

The first three criteria are inequalities and can be verified by a rejection of a null hypothesis that α, β, and ϕ of (1.5), (1.6), and (1.7) are equal to 0. The fourth criteria requires β_S from equation (1.8) to equal 0. Because this requirement is an equality, this fourth critirea cannot be verified by rejection of a null hypothesis. Additionally, these criteria would fail to validate a surrogate endpoint in which only part of the treatment effect is explained by the surrogate endpoint, a situation where β_S from equation (1.8) would be nonzero. Because of this Freedman et al. [19] introduced the concept of "proportion explained" by the surrogate to estimate the extent of surrogacy of an intermediate endpoint.

1.2.2 Freedman's Proportion Explained

Freedman et al. [19] proposed to estimate the proportion of the treatment effect that is captured by the surrogate (PE). Using the above models and notation, PE(T,S,Z) is defined as:

$$PE(T, S, Z) = \frac{\beta - \beta_S}{\beta}. \tag{1.10}$$

As β_S becomes closer to zero, the closer this proportion is to one, indicating the extent to which the surrogate captures the full treatment effect on the true endpoint. However, PE is not restricted to values between 0 and 1 as there is no restriction that β_S be smaller than β. Thus the intuitive explanation of PE as a proportion is not accurate. Furthermore,

Freedman et al. [19] pointed out that if the model in (1.8) is incorrect and an interaction term between Z and S exists, PE could not be used. Molenberghs et al. [9] suggested using the relative effect (RE), defined as $\frac{\alpha}{\beta}$ in the above models (1.5) (1.6), instead of PE. It has been shown that a one-to-to correspondence between PE and RE exists [29]. Thus there are no statistical advantages to RE over PE, as RE is simply a reparameterization of PE.

1.2.3 Meta-analytic Framework

An additional measure of surrogacy has been proposed [10] for a meta-analytic validation of a surrogate endpoint. With data from multiple trials, one can regress the observed treatment effects on the observed treatment effects on the surrogate. The model is then used to predict the relation between the observed trial effects on the surrogates and the observed trial effects on the true endpoint. A valid surrogate should have a high correlation between these two measures. The full details of this Meta-analytic validation approach are described by Molenberghs et al. [30].

1.3 APPLICATIONS TO NEOADJUVANT BREAST CANCER TRIALS

The surrogate endpoint which is the topic of this dissertation is pathological complete response, pCR, at the time of surgery following neoadjuvant therapy. The three most commonly used definitions of pCR, according to the Food and Drug Administration (FDA) [38] are as follows:

1. ypT0/Tis: absence of invasive cancer in the breast.
2. ypT0/Tis ypN0: absence of invasive cancer in the breast and axillary nodes.
3. ypT0 ypN0: absence of invasive and in situ cancer in the breast and axillary nodes.

In a meta-analysis of 11,955 patients [12], all three definitions of pCR were related to improved event-free survival (EFS) and overall survival (OS). This provides evidence of pCR as a prognostic surrogate for survival which can be useful for clinical decision making a patient care. However, the meta-analysis failed to prove treatment surrogacy using the

meta-analytic approach [30]. Possible explanations for this failure are lack of power due to small effect sizes on pCR, the heterogeneity of the patient population for the trials included, or that pCR is not in fact, a valid treatment surrogate for survival endpoints.

Chapter 2 will focus on how to design a Neoadjuvant Clinical Trial with a survival outcome when there is information about the drug's effect on pCR. Chapter 3 will focus on how to conduct causal inference when conditioning on the poct-randomization variable pCR.

1.4 ASSESSMENT OF EFFECT SIZE AND POWER FOR SURVIVAL ANALYSIS THROUGH A BINARY SURROGATE ENDPOINT IN CLINICAL TRIALS

The second chapter of this dissertation explains how to determine the power and effect size of a clinical trial with a survival endpoint, when there is information on a binary intermediate endpoint associated with both treatment assignment and survival. Specifically, we propose to model the survival function of each group as a mixture of exponential models with the mixing proportion determined by the intermediate endpoint data and hazard rates determined by historical data. We use simulations to estimate the power of the trial using our models and propose to use average hazard ratio over the study period as a measure of the treatment effect size. Additionally, we show how Schoenfeld's formula for survival endpoints assuming proportional hazards can be used to estimate power, using the calculated average hazard ratio instead of the constant hazard ratio in the formula. We extend our approach to a setting with continuous accrual and follow-up and compare it with a single point of accrual setting. To illustrate our approach we apply our method to two neoadjuvant breast cancer clinical trials where we use data on pCR to determine the power of the trial for a survival endpoint.

1.5 PRINCIPAL STRATIFICATION FOR CAUSAL EFFECTS CONDITIONING ON POST-TREATMENT VARIABLES

The third chapter this dissertation relates to causal inference in the presence of a post randomization variable. In particular, we would like to limit our analysis to those patients who would achieve pCR under treatment and would therefore be likely to benefit from the added treatment. We use the method of principal stratification developed by Frangakis and Rubin [18]. Principal stratification creates subgroups of patients based on the patient's potential value of a post-randomization variable if assigned to treatment or control. For a binary surrogate S such as pCR, there are a total of four basic principal strata: (1) those who would not achieve pCR regardless of having added treatment, (2) those who would achieve pCR only if they receive added treatment, (3) those who would achieve pCR regardless of added treatment, and (4) those who would achieve pCR only if they do not receive added treatment. Conditioning on these strata maintains a causal interpretation because the conditioning is done on potential outcomes instead of observed outcomes. Using this approach we can therefore condition on those who would achieve pCR under a specific treatment but cannot condition on observed pCR response status. The former is independent of treatment assignment while the latter is dependent on treatment, and conditioning on it could introduce bias because we are no longer comparing completely randomized groups.

We will apply this method to data from the NSABP B-40 trial. In this trial women with operable human epidermal growth factor receptor 2 (HER2)-negative breast cancer were randomly assigned to receive (N=604) or not to receive (N=602) bevacizumab along with their neoadjuvant chemotherapy regimens [3]. Bevacizumab significantly increased the proportion of pCR (28.2% vs. 34.5%, P=.02). Median follow-up time was 4.7 years. In the long-term outcomes analysis, patients on bevacizumab did show improvement in DFS compared to the control patients, although the improvement was not statistically significant (HR=.80, P=.06) [4]. We will assess the treatment effect among those patients who would achieve pCR had they been assigned treatment.

The difficulty in the method is identifying those patients in the control group who would be pCR responders under treatment. To identify this we make two assumptions. First, we

assume that pCR responders in the control group would be pCR responders had they received treatment as well. This is known as the monotonicity assumption. Second, we assume a parametric model form for the probability of a control pCR non-responder achieving pCR under treatment. Once we estimate the parameters of the model we use the model to impute pCR status under treatment. After imputation we conduct the stratified analysis among the subgroup of patient that would be pCR responders under treatment drug. To account for uncertainty in the imputed values we use the technique of bootstrap resampling [27].

2.0 ASSESSMENT OF EFFECT SIZE AND POWER FOR SURVIVAL ANALYSIS THROUGH A BINARY SURROGATE ENDPOINT IN CLINICAL TRIALS

2.1 INTRODUCTION

We have seen a major shift in the conduct of breast cancer clinical trials in recent years. Historically, an experimental drug or treatment was administered with standard systemic therapy following surgery to one group of randomly assigned patients while patients randomly assigned to the control group received the standard therapy alone. Then patients from the two groups were followed over time for comparison of long-term outcomes such as disease-free survival (DFS), overall survival (OS) and progression-free survival (PFS). However, in recent years, there have been an increasing number of neoadjuvant trials where many of the systemic therapies are administered prior to the surgery of the breast [38].

The primary endpoint in recently initiated neoadjuvant breast cancer clinical trials is pathological complete response (pCR), a binary variable that categorizes tumors into responders and nonresponders. The US Food and Drug administration's (FDA) latest guidance accepts pCR definitions of either absence of invasive cancer in the breast and axillary nodes or absence of invasive and in situ cancer in the breast and axillary nodes (ypT0/Tis ypN0 and ypT0 ypN0 respectively in the current American Joint Committee on Cancer (AJCC) staging system) [38]. The rationale for using pCR as the trial endpoint is that efficacy of treatment can be determined at the time of surgery, usually six to seven months after therapy commences, instead of the typical 5-10 years of follow-up required to show improved efficacy on survival endpoints in the adjuvant setting. Furthermore, the strong prognostic

link between pCR and survival has been well documented [12, 39], making pCR an attractive candidate surrogate marker for survival.

It is important to note, however, that pCR has yet to be validated as a surrogate for treatment efficacy on survival endpoints. Furthermore, a recent meta-analysis [12] of 12 neoadjuvant trials failed to show evidence that pCR is a valid surrogate endpoint. That being said, while the FDA does not usually accept surrogate endpoints to demonstrate efficacy, it makes exceptions in cases of unmet need. The FDA has therefore cautiously accepted pCR as an endpoint for accelerated approval in high-risk subsets of breast cancer patients, provided certain additional requirements are met, including the recruitment of patients for a larger confirmatory trial with a survival endpoint [38].

This chapter addresses the power and sample size calculation for this confirmatory trial. Specifically, we will show how to use observed data from a neoadjuvant trail with pCR data to determine the sample size and power for the FDA-mandated confirmatory trial. This will provide a tool for investigators in the planning stage of the larger confirmatory trial.

Recently, two approaches of sample size determination for these confirmatory trials have been proposed. Berry and Hudis [7] advocate using the FDA meta-analysis [12] patient-level relations between pCR responders and non-responders to estimate the treatment effect size by considering each treatment arm as a mixture distribution of pCR responders and non-responders. Alternatively, Hatzis et al. [22] calculate the sample size using biased bootstrap resampling from the survival data of 127 triple negative breast cancer patients. We adopt the first approach in our method and rely on external estimates of the relative risk between pCR responders and non-responder.

To illustrate our approach, we consider two examples of neoadjuvant breast cancer trials with data on both pCR (defined in both as the absence of invasive cancer in the breast) and long-term follow-up. We use the available data on pCR to estimate the power and effect size of the two trials and compare it to the long term data available. The first trial is the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-27 study [2]. Women with operable breast cancer were randomly assigned to receive preoperative doxorubicin and cyclophosphamide (AC) (N=804) or docetaxel (T) added to AC (N=805). A third arm was randomly assigned to preoperative AC and post-operative T (n=802), but is excluded from

this analysis because the postoperative drug may confound the relation between pCR and survival. For the two arms with exclusively neoadjuvant therapy, there was a significant difference ($P < 0.001$) between the pCR proportion in T+AC (26.1%) and AC (12.9%). Median follow-up time was 6.5 years. In the long-term outcomes analysis there was no significant difference in DFS between T+AC and AC (HR=0.90, P=0.22).

The second trial is the NSABP B-40 study. Women with operable human epidermal growth factor receptor 2 (HER2)-negative breast cancer were randomly assigned to receive (N=604) or not to receive (N=602) bevacizumab along with their neoadjuvant chemotherapy regimens [3]. Bevacizumab significantly increased the proportion of pCR (34.5% vs. 28.2%, P=0.02). Median follow-up time was 4.7 years. In the long-term outcomes analysis, patients on bevacizumab did show improvement in DFS compared to the control patients, although the improvement was not statistically significant (HR=0.80, P=0.06) [4].

The rest of the chapter is organized as follows. In section 2.2 we briefly review the basic model and calculation used in clinical trials to determine an appropriate sample size for testing a significant treatment effect on survival. In section 2.3 we propose two alternative models which are more intuitive for neoadjuvant data with information on a binary surrogate endpoint. The characteristics of the model are then laid out. In section 2.4 we provide details on how to use the models for sample size and power calculations. In section 2.5 we conduct a simulation study to determine the power of the trial under a range of parameters. In section 2.6 we illustrate our method through data collected from the NSABP B-27 and NSABP B-40. We conclude by discussing how our results provide a better understanding of neoadjuvant clinical trials and its effect size and sample size determination.

2.2 A STANDARD APPROACH FOR RANDOMIZED CLINICAL TRIALS WITH TIME-TO-EVENT ENDPOINT

For a placebo-controlled two arm clinical trial with OS or DFS as its primary endpoint, the logrank test is used to test the null hypothesis of equal distribution of event times for the two groups [13, 28]. This test is robust as it makes no assumptions about the

underlying distribution for each group's survival. To determine an appropriate sample size to achieve sufficient power, however, additional assumptions are required about the underlying distributions. The usual distributional assumption is that each group's survival time follows an exponential distribution with a given hazard rate. This assumption results in a constant hazard ratio over time between the two groups, also known as the proportional hazards assumption. Under this assumption the required sample size, N, for a two-sided logrank test of size α to achieve a power of β, assuming equal allocation and a hazard ratio of λ between the two groups is the following:

$$N = \frac{4(z_\beta + z_{1-\alpha/2})^2}{\rho(log^2\lambda)} \tag{2.1}$$

where $z_{1-\alpha/2}$ and z_β are quantiles of the normal distribution and ρ is the proportion of events among study participants at the time of analysis [34]. This sample size calculation is popular due to its simplicity and few required inputs.

In the neoadjuvant setting we assume that pCR acts as a mediating variable for survival as suggested by findings that patients who achieved pCR had much better prognosis than those who did not [12, 39]. We would therefore like to allow for different survival distributions based upon pCR status. This is the underlying motivation for our model. Our model can then be used to provide power estimation for the comparison of long-term outcomes with the data available at the conclusion of the neoadjuvant component of the clinical trial.

2.3 METHODS

2.3.1 Mixture Models for Survival Data

There is a rich literature on using mixture models to model survival data. Boag [8] and Berkson and Gage [5] first used them to describe survival data where some of the study population is cured from the disease and will not have an observed event during the study period. These "cure models" have since been studied for their asymptotic properties [20] and for estimation under specific conditions, such as the proportional hazards assumption

[36]. Here we stratify patients by their treatment arm and pCR status and the distribution of time to event within each patient stratum is modeled parametrically. We assume two treatment arms (treatment vs. control). In both NSABP B-27 and B-40 very few patients ($< 1\%$) died before the primary surgery was performed. Therefore, for simplicity we assume in our models that all patients survive through the conclusion of the neoadjuvant component and have information regarding pCR status.

2.3.2 Model 1

Denote the pCR proportion among patients randomized to the control arm as δ. At the end of the neoadjuvant component, we assume that the improvement on pCR due to the intervention is Δ. We model a patient's pCR status from group i (0=control, 1=treatment) as a Bernoulli random variable with parameter δ and $\delta + \Delta$ for the control and treatment group, respectively.

We then assume that conditional on pCR status $j \in \{0, 1\}$, a patient's survival time T following surgery follows an exponential distribution with parameter λ_j. Here we assume that all patients survive up to pCR assessment and consider survival time $T = 0$ to be time of surgery. Assuming the hazard ratio between pCR and non-pCR patients as β_1 we have $\lambda_1 = \lambda_0\beta_1$. Estimates of β_1 are available in the literature for different subpopulations [12] and generally range from .2 to .6.

As a starting point we assume the survival functions depend only on pCR status irrespective of treatment group, that is, pCR is a perfect surrogate marker of survival. We refer to this as Model 1. Subsequently we will relax this assumption to reflect that differences in patient survival after pCR assessment between pCR and non-pCR groups may vary by treatment arms. We refer to this relaxed model as Model 2.

Model 1 can be thought of as the mixture of two exponential distributions with a mixing probability corresponding to the pCR proportion for each treatment group. Each group's probability density function for survival time, $f_i(t)$, is the following:

$$f_0(t) = (1 - \delta)\lambda_0 e^{-\lambda_0 t} + \delta\lambda_0\beta_1 e^{-\lambda_0\beta_1 t} \tag{2.2}$$

$$f_1(t) = (1 - \delta - \Delta)\lambda_0 e^{-\lambda_0 t} + (\delta + \Delta)\lambda_0\beta_1 e^{-\lambda_0\beta_1 t} \tag{2.3}$$

Additionally each group's survival and hazard function, $S_i(t)$ and $h_i(t)$ respectively, are easily derived from (2.2) and (2.3) :

$$
\begin{aligned}
S_0(t) &= (1-\delta)e^{-\lambda_0 t} + \delta e^{-\lambda_0\beta_1 t} \\
S_1(t) &= (1-\delta-\Delta)e^{-\lambda_0 t} + (\delta+\Delta)e^{-\lambda_0\beta_1 t} \\
h_0(t) &= \frac{(1-\delta)\lambda_0 e^{-\lambda_0 t} + \delta\lambda_0\beta_1 e^{-\lambda_0\beta_1 t}}{(1-\delta)e^{-\lambda_0 t} + \delta e^{-\lambda_0\beta_1 t}} \qquad (2.4) \\
h_1(t) &= \frac{(1-\delta-\Delta)\lambda_0 e^{-\lambda_0 t} + (\delta+\Delta)\lambda_0\beta_1 e^{-\lambda_0\beta_1 t}}{(1-\delta-\Delta)e^{-\lambda_0 t} + (\delta+\Delta)e^{-\lambda_0\beta_1 t}} \qquad (2.5)
\end{aligned}
$$

Dividing equation (2.5) by (2.4) yields the hazard ratio between the 2 groups:

$$
\beta(t) = \frac{\left[(1-\delta-\Delta)\lambda_0 e^{-\lambda_0 t} + (\delta+\Delta)\lambda_0\beta_1 e^{-\lambda_0\beta_1 t}\right]\left[(1-\delta)e^{-\lambda_0 t} + \delta e^{-\lambda_0\beta_1 t}\right]}{\left[(1-\delta-\Delta)e^{-\lambda_0 t} + (\delta+\Delta)e^{-\lambda_0\beta_1 t}\right]\left[(1-\delta)\lambda_0 e^{-\lambda_0 t} + \delta\lambda_0\beta_1 e^{-\lambda_0\beta_1 t}\right]} \qquad (2.6)
$$

2.3.3 Model 2

Since PCR may not be a perfect surrogate of patient survival, we also consider the case when the treatment may be associated with lower event rate, compared to the controls, even within the pCR and non-pCR stratum. We model this by adding a parameter β_2 to capture an additional benefit from treatment. For patients on the control arm, their hazard rate is still λ_0 $\lambda_0\beta_1$ and pCR non-responders and responders, respectively. For patients on the treatment arm their hazard rate would become $\lambda_0\beta_2$ and $\lambda_0\beta_1\beta_2$ for pCR non-responders and responders, respectively. The parameter β_2 is interpreted as the hazard ratio between treatment and control among those with the same pCR status. Accordingly, we have the following formula for the hazard ratio between the two groups:

$$
\beta(t) = \frac{\left[(1-\delta-\Delta)\lambda_0\beta_2 e^{-\lambda_0\beta_2 t} + (\delta+\Delta)\lambda_0\beta_1\beta_2 e^{-\lambda_0\beta_1\beta_2 t}\right]\left[(1-\delta)e^{-\lambda_0 t} + \delta e^{-\lambda_0\beta_1 t}\right]}{\left[(1-\delta-\Delta)e^{-\lambda_0\beta_2 t} + (\delta+\Delta)e^{-\lambda_0\beta_1\beta_2 t}\right]\left[(1-\delta)\lambda_0 e^{-\lambda_0 t} + \delta\lambda_0\beta_1 e^{-\lambda_0\beta_1 t}\right]} \qquad (2.7)
$$

We refer to this model as Model 2. Because of the complex nature of our models we rely on simulation instead of asymptotic theory to determine the power of a given trial based on our models.

2.4 POWER CALCULATION

2.4.1 Empirical Power

Using the above survival models we propose the following two methods for calculating the power of a trial. The first method is to use the observed neoadjuvant data to calculate the parameters of the survival models and then calculate empirical power. For the above models δ is calculated as the proportion of pCR responders in the control arm and Δ as the increased proportion of pCR responders observed in the treatment arm. λ_0 and β_1 can be estimated based on historical data describing the survival of pCR responders and non-responders for the subgroup of patients in the trial such as that found in the FDA meta-analysis [12]. We suggest treating β_2 as a sensitivity parameter and initially setting it to 1.0 as in Model 1. Other values of β_2 can then be chosen to calculate the power under the assumption of an additional survival difference between treatment and control within each pCR arm.

Once the parameter values are chosen, we propose to calculate the empirical power using simulations. First, for the given sample size and study end time, T_{end}, simulate the patient's pCR status by a Bernoulli draw with probabilities δ and $\delta + \Delta$ for the control and treatment group, respectively. Then simulate the event time following surgery of each patient from an exponential distribution with hazard rates of $\lambda_0\beta_2$ and $\lambda_0\beta_1\beta_2$ for pCR responders and non-responders respectively, with Model 1 simulations corresponding to $\beta_2 = 1$. An event time greater than T_{end} is considered censored for purposes of analysis to reflect administrative censoring. Then we compare the survival of the two study arms following the conclusion of the neoadjuvant component with a 2-sided logrank test with a predetermined α level. Then repeat the simulation many (1,000 or 10,000) times. The proportion of logrank tests that achieve statistical significance will be the empirical power of the randomized trial. This procedure can be repeated with different sample sizes until the desired power is achieved.

2.4.2 Approximate Power using Schoenfeld's Formula

The second method to determine the power of the study is an approximation using Schoenfeld's formula (2.1). In the equation ρ can be calculated based on the survival functions at

time T_{end}. Because our model incorporates a mixture of distributions, the hazard ratio between treatment and control changes over time and must be calculated based on the assumed parameters of our model. To simplify calculating the hazard ratio between the two groups at a given time point in the study, we initially assume all patients are accrued at a single time point and followed until time T_{end}, the time from surgery to study close. We propose to approximate the treatment effect size by numerically approximating the average hazard ratio given in equations (2.6) and (2.7) over the length of follow-up time. We numerically approximate by averaging the hazard ratio calculated at 1000 equidistant points over the entire post-surgery period. The average hazard ratio during the follow-up period will be used to replace the constant hazard ratio in Schoenfeld's formula (2.1) to approximate the sample size required to achieve a chosen power or vice versa.

To validate this approximation for the effect size we first plot the hazard ratio over time to see its trajectory for a given set of parameter values. Using average hazard ratio would be most informative as an effect size if the hazard ratio trajectory is mostly flat. Additionally, we check the validity of replacing λ with average hazard ratio in Schoenfeld's equation (2.1) by comparing our approximation method with our empirical power calculation. First we calculate the required sample sizes to achieve 80 and 90 percent power using Schoenfeld's equation (2.1), replacing λ with the calculated average hazard ratio and calculating ρ based on the modeled survival function. We calculate these required sample sizes for a total of 288 combinations of parameter values ($\delta \in \{0.1, 0.3, 0.5\}$, $\Delta \in \{0.1, 0.2, 0.3, 0.4\}$, $\lambda_0 \in \{0.08, 0.12, 0.16\}$, $\beta_1 \in \{0.2, 0.3, 0.4, 0.5\}$, $\beta_2 \in \{1, 0.9\}$ for $T_{end} = 5$). We then calculate the empirical power using our first method (with 1,000 runs) for the sample sizes obtained from the approximation under our model assumptions along with its 95% confidence interval using a normal approximation for proportions. If the empirical power matches its target of 80 and 90 used to calculate the approximate sample sizes, this would validate the using the average hazard ratio in Schoenfeld's formula for approximate power and sample size calculations and provide for a more efficient tool for sample size determination over simulations.

2.4.3 Accounting for Continuous Accrual and Follow-up

In practice, patients are accrued continuously over a period of time. We also extended our models to incorporate this feature. In the extension we assume a uniform distribution of entry time over the accrual period. Additionally, we assume that no patient dies before the primary surgery and pCR assessment, which we assume is half a year after accrual for every subject. First, we derive the survival and hazard functions for each treatment group. Through simulations we then calculate the empirical power of the logrank test assuming continuous accrual of 2 years, with 4.5 years of follow-up after the last patient accrual, with other parameter values equivalent to the single point accrual setting. We choose these accrual and follow up times so that the average follow-up time equals the follow up time assumed in our original setting of a single accrual time point and results of the two can be compared. Finally, we approximate the required sample sizes using Schoenfeld's formula, by applying the average hazard ratio in place of constant hazard ratio, to acheive 80 and 90 percent power and calculate the empirical power under these estimated sample sizes to ascertain the accuracy of our estimation technique at achieving their target. Because the results of this extension are similar to the results assuming a single time point accrual, we choose to present the derivation and simulation details in the appendix.

2.5 SIMULATION STUDIES

2.5.1 Setup

We conducted simulations to assess the empirical power for detecting a treatment difference in patient survival for a randomized neoadjuvant trial with a total of N patients allocated equally to two groups. In each run, we first simulated the patient's pCR status by a Bernoulli draw with probabilities δ and $\delta + \Delta$ for the control and treatment group, respectively. We then simulated the event time of each patient from an exponential distribution with hazard rates of $\lambda_0\beta_2$ and $\lambda_1\beta_2$ for pCR responders and non-responders respectively, with Model 1 simulations corresponding to $\beta_2 = 1$. Administrative censoring was considered to reflect

the closure of long-term outcome data in multi-center clinical trials. Therefore, an event time after this closure, T_{end}, was considered censored for purposes of analysis. We compared the survival following the conclusion of the neoadjuvant component with a 2-sided logrank test with $\alpha = 0.05$ chosen for significance. We repeated each simulation 10,000 times. The proportion of logrank tests that achieved significance equals the power of the randomized trial. All simulation were performed in R Version 3.1.2 [32].

The above simulation was carried out for $T_{end} = 5, \delta = 0.3$ along with every combination of the following parameters for Model 1: $N = 1000, 2000$; $\lambda_0 =, 0.08, 0.12$; $\Delta = 0.1, 0.2, 0.3$; $\beta_1 = 0.2, 0.3, 0.4$. For Model 2 we ran a simulation with $N = 1000$ and $\beta_2 = 1, 0.95, 0.9$ assuming $\lambda_0 = .08, \Delta = 0.2$, and $\beta_1 = 0.3$ to evaluate the sensitivity of the model with respect to β_2, the treatment effect within each pCR strata.

2.5.2 Simulation Results

Results of our simulation studies on the empirical power under Model 1 are presented in Table 2. The average hazard ratio between treatment and control ranged from 0.92 ($\beta_1 = 0.4, \delta = 0.1$) to 0.67 ($\beta_1 = 0.2, \delta = 0.3$). Increasing λ_0, the hazard for those who did not achieve a pCR, from 0.08 to 0.12 did not substantially change the average HR between groups. However, increasing λ_0 did increase power, with a greater gain in power for $\Delta = 0.2$ and 0.3. Decreasing β_1 caused a marginal decrease in the average HR. For example, holding Δ at 0.2 and λ_0 at .08, a decrease in β_1 from 0.4 to 0.3 corresponded to a decrease in average HR from 0.848 to 0.814 and an increased power from 0.457 to 0.616 under N=2000.

An increase in Δ led to a large decrease in average HR and increased power. For example, holding β_1 at 0.3 and λ_0 at 0.08, an increase in Δ from 0.1 to 0.2 decreased the average HR from 0.91 to .81 and led to an absolute increase in power of 0.24 and 0.42 for $N = 1000, 2000$ respectively.

Eighty percent power, a common threshold in clinical trial design, was not achieved for all scenarios where $\Delta = 0.1$. For $\Delta = 0.2$, eighty percent power was achieved only for $N = 2000$, $\beta_1 = 0.2$ and $\lambda_0 = 0.12$. The change in sample size from $N = 1000$ to $N = 2000$ caused an marked increase in power, particularly for Δ=0.2 or 0.3. For example, or $\lambda 0 = .08$, $\beta 1 = 0.3$, and $\Delta = 0.2$ the power almost doubled from 0.36 to 0.61.

Table 2: Average HR and Empirical Power for the log-rank test on survival difference under Model 1 with various parameter values setting δ=.3, T_{end}=5

		$\lambda_0 = .08$			$\lambda_0 = .12$		
			Empirical Power			Empirical Power	
β_1	Δ	Average HR	N=1000	N=2000	Average HR	N=1000	N=2000
0.4	0.1	0.922	0.102	0.152	0.921	0.122	0.193
0.4	0.2	0.848	0.264	0.457	0.845	0.352	0.607
0.4	0.3	0.774	0.530	0.805	0.772	0.670	0.923
0.3	0.1	0.906	0.126	0.199	0.902	0.161	0.264
0.3	0.2	0.814	0.364	0.616	0.810	0.489	0.774
0.3	0.3	0.725	0.694	0.932	0.720	0.836	0.985
0.2	0.1	0.886	0.155	0.259	0.882	0.210	0.357
0.2	0.2	0.776	0.484	0.762	0.770	0.642	0.897
0.2	0.3	0.669	0.838	0.985	0.663	0.942	0.999

Table 3 shows the results of the simulation for Model 2 with the relevant comparisons from Model 1, corresponding to $\beta_2 = 1$. As β_2, the hazard ratio of treatment to control irrespective

Table 3: Average HR and Empirical Power with additional treatment effect within pCR strata using Model 2

Δ	β_2	Average HR	Empirical Power
0.2	1	0.814	0.364
0.2	0.95	0.775	0.505
0.2	0.9	0.735	0.654
0.3	1	0.725	0.694
0.3	0.95	0.69	0.803
0.3	0.9	0.655	0.885

parameters: $N = 1000$, δ=0.3, λ_0=0.08, T_{end}=5

of pCR status, decreases, the corresponding average hazard ratio of the complete treatment and control groups decreases as well and power increases. For the examples in Table 2 the magnitude of the change in average HR is slightly less than the magnitude in change of β_2. The gain in power from Model 1 to Model 2 is quite substantial. With $N = 1000$ and $\Delta = 0.2$ in Table 2 the power increased from 0.364 to 0.654 when changing β_2 from 1 to 0.9.

Figures 1-4 show the plots of the hazard ratio given in equation (2.6) over time for a variety of parameter settings.

For all of these the hazard ratio appears stable over time, making the average hazard ratio a useful measure of effect size. Applying the average hazard ratio to Schoenfeld's equation (2.1) to determine sample size yielded consistent results compared to the empirical power. Overall, empirical power achieved its target, with the average empirical power of 79.4% and 89.5% for 80% and 90% power targets respectively over the 288 scenarios tested. Among the 288 scenarios 93.4 and 94.4 percent of the 95% confidence intervals contained their targets of 80 and 90 percent power, respectively. This validates the use of the average hazard ratio as a measure of effect size for the parameter range of our simulations and gives a more efficient method of calculating a sample size for a specific power target.

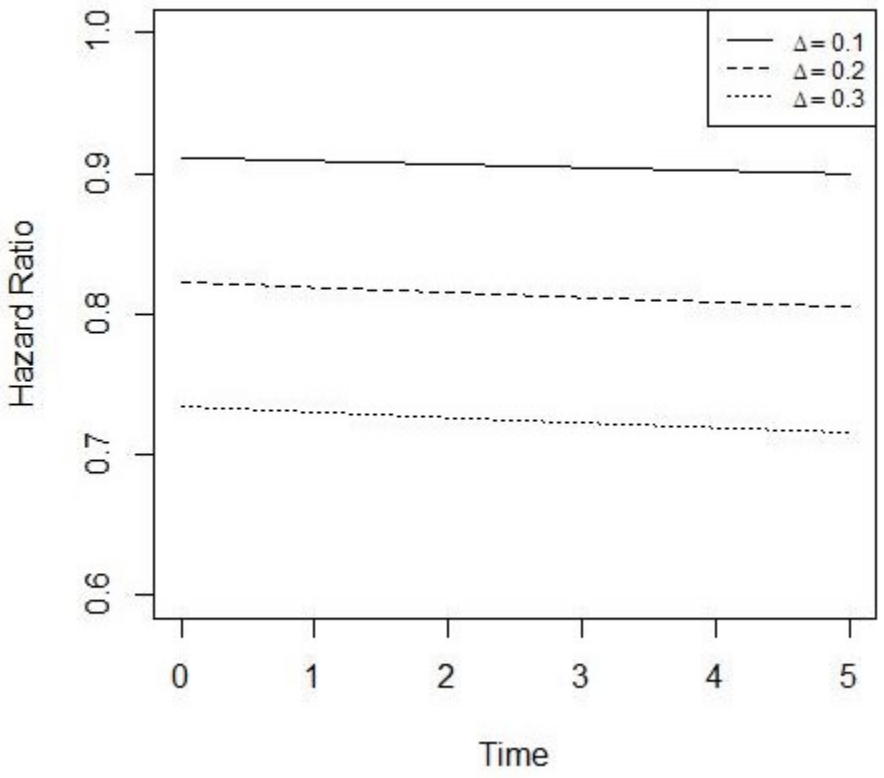

Figure 1: Plots of hazard ratio over time.$\delta = .3, \lambda_0 = .08, \beta_1 = .3, \beta_2 = 1$

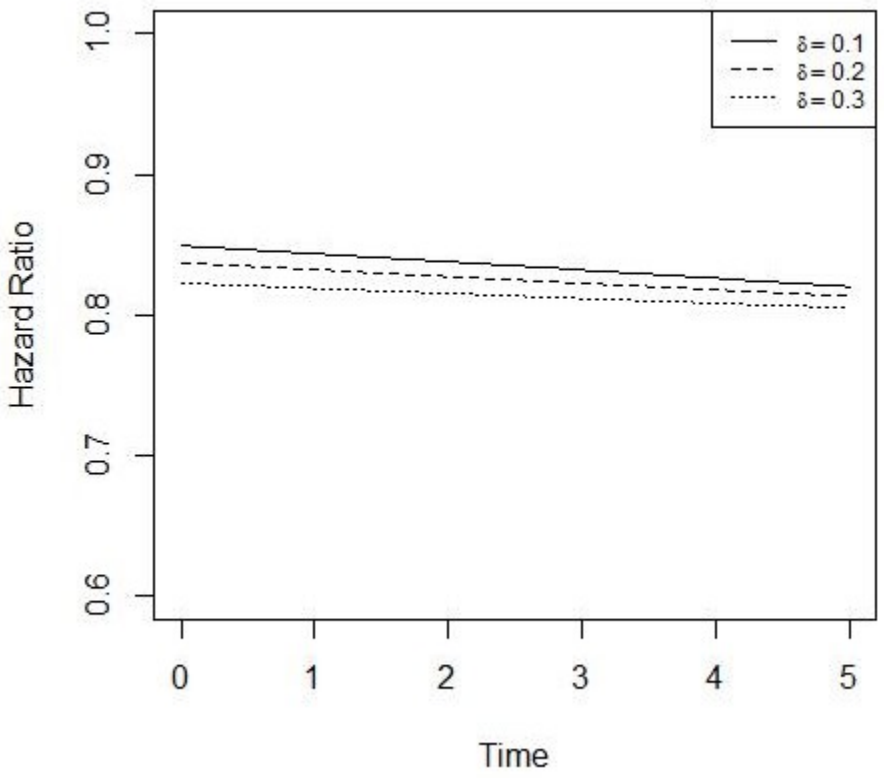

Figure 2: Plots of hazard ratio over time. $\Delta = .2, \lambda_0 = .08, \beta_1 = .3, \beta_2 = 1$

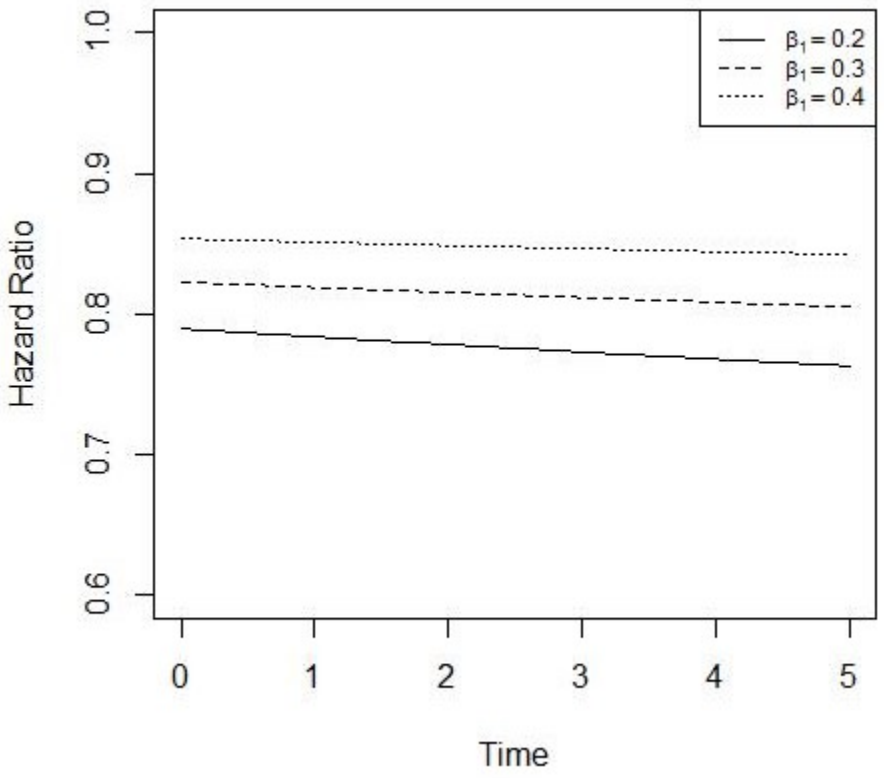

Figure 3: Plots of hazard ratio over time.$\delta = .3, \Delta = .2\lambda_0 = .08, \beta_2 = 1$

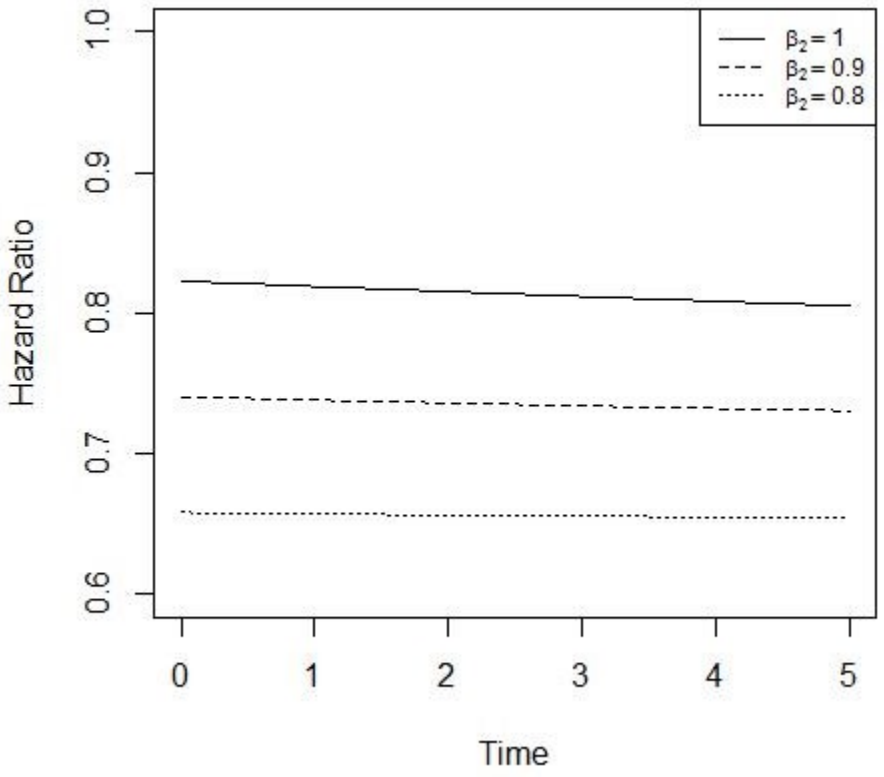

Figure 4: Plots of hazard ratio over time. $\delta = .3, \Delta = .2\lambda_0 = .08, \beta_1 = .3$

Figure 5 displays the relationship between pCR treatment effect (δ) the total required sample size using Schoenfeld's equation (2.1) to achieve 80% power. The greatest decrease in

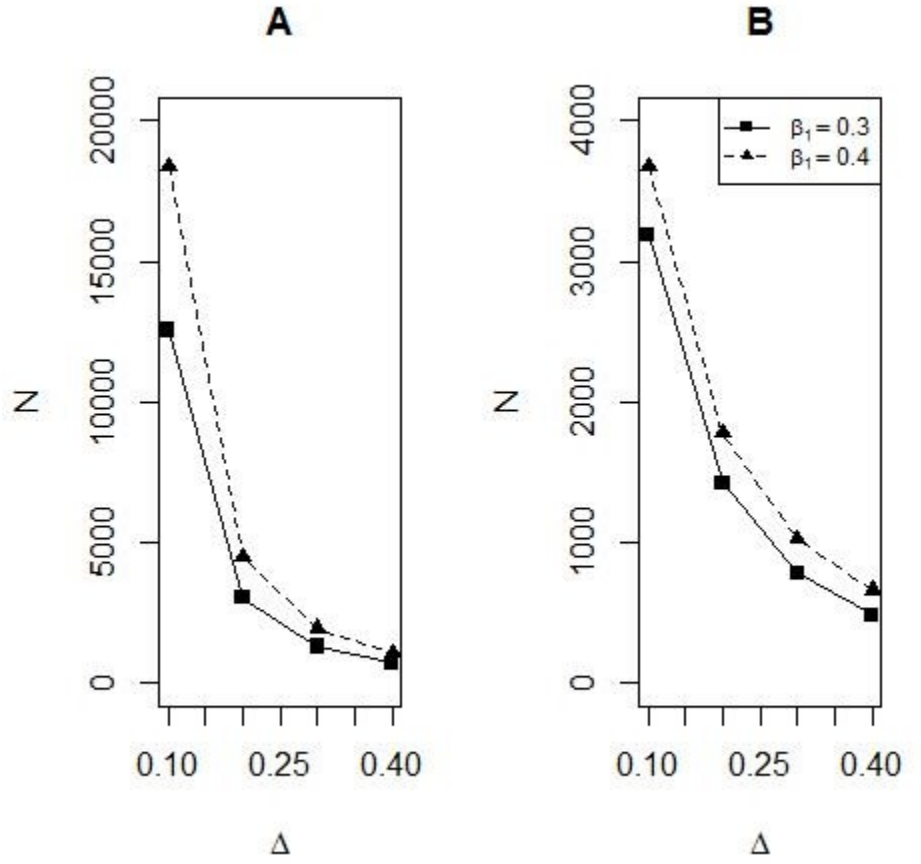

Figure 5: Plots of total sample size by Δ.
Both plots assume baseline pCR rate (δ) of .3, hazard rate for control non-pCR group (λ_0) of .08 and follow-up time of 5 years. A: Model 1 with $\beta_1 = .3, .4$. B: Model 2 with $\beta_1 = .3, .4$ and $\beta_2 = .9$

required sample size occurs from increasing Δ from 0.10 to 0.20 than any other equal increase in Δ. Additionally, Model 2, with $\beta_2 = 0.9$ compared to Model 1 with other parameters equals requires a substantially lower sample size and is not as sensitive to Δ compared to Model 1.

Using a continuous accrual process produced nearly identical average hazard ratios and empirical power to a process modeled by a single accrual point at the mean of the total accrual time (Tables A1 and A2 in appendix). Based on these results, the power of a neoadjuvant

trial with a continuous uniform accrual process can be reasonably estimated by assuming all participants are accrued at a single point in time at the mean accrual period. The time of study end in the power calculation is then the mean accrual plus follow-up length after last accrual. This simplifies the calculations for the average hazard ratios.

2.6 ILLUSTRATION OF METHOD WITH NSABP B-27 AND B-40 DATA

We applied our simulation technique using the neoadjuvant data of NSABP B-27 and B-40 to show how our method can be used in practice and to compare our results with the observed long-term survival data available from these studies. In NSABP B-27 a total of 1535 subject were randomized to the two strictly neoadjuvant arms. There was an observed pCR increase of 0.132 in the T+AC arm (pCR proportion=26.1%) over the T arm (pCR proportion=12.9%). To incorporate these results in our power calculation we set the parameters of our NSABP B-27 simulation to N=1536 (rounded up to nearest even number to create equal groups), $\delta = 0.129$ (reflecting pCR proportion in the control group), and $\Delta = 0.132$ (reflecting observed pCR difference between the two groups).

In NSABP B-40 a total of 1186 subjects had data on pCR status. There was an observed pCR increase of 0.063 in the bevacizumab group (pCR proportion=34.5%) over the no bevacizumab group (pCR proportion=28.2%). To incorporate these results in our power calculation we set the parameters of our NSABP B-27 simulation to N=1186, $\delta = 0.282$ (reflecting pCR proportion in the control group), and $\Delta = 0.063$ (reflecting observed pCR difference between the two groups).

For our method we require an estimate of the HR between a pCR responder and non-responder. For our simulations we used $\beta_1 = 0.35, 0.45$ for both trials to reflect the variability of the observed pCR responder vs. non-responder hazard ratios for DFS of 0.45 and 0.42 of B-27 and B-40 respectively. In reality, information on β_1 will not be available after the neoadjuvant component but can be estimated from the FDA meta-analysis [12] for a given subpopulation under study. To assess sensitivity to treatment differences within pCR strata, we ran our simulation for $\beta_2 = 1, 0.9, 0.8$, with β_2 corresponding to the HR of treatment vs.

control for patients within the same pCR stratum. Each simulation was rerun 1,000 times to calculate the empirical power.

We present the results of the simulation for NSABP B-27 and NSABP B-40 in Table 4. Assuming $\beta_2 = 1$ in B-27, the average hazard ratio is about 0.898 which is similar to the

Table 4: Average HR, empirical power, and estimated power using Schoenfeld's formula for NSABP B-27 and B-40 data

		NSABP B-27			NSABP B-40		
β_1	β_2	Average HR	Empirical Power	Estimated Power	Average HR	Empirical Power	Estimated Power
0.35	1	0.898	0.199	0.205	0.946	0.075	0.071
0.35	0.9	0.810	0.579	0.588	0.854	0.279	0.279
0.35	0.8	0.721	0.906	0.912	0.761	0.630	0.643
0.45	1	0.916	0.150	0.153	0.957	0.065	0.059
0.45	0.9	0.826	0.508	0.516	0.863	0.257	0.258
0.45	0.8	0.735	0.881	0.883	0.768	0.623	0.630

Parameter values: B-27:N=1536, δ=0.129, Δ=0.132; B-40: N=1186, δ=0.282, Δ=0.063

hazard ratio of 0.90 actually reported after long term follow-up. Additionally, the follow-up data [2] did not show a significant treatment difference in DFS in either pCR or non-pCR patients which corresponds to $\beta_2 = 1$ in our models. In B-40, where bevacizumab was given post-surgery in addition to neoadjuvantly, it is reasonable to suspect there may be an additional treatment effect beyond pCR. Indeed, the reported hazard ratio of 0.80 after follow-up falls between $\beta_2 = 0.9$ and $\beta_2 = 0.8$ in our model. In both scenarios our simulations show that there was a severe lack of power to detect a significant result for the long-term survival analysis.

2.7 DISCUSSION

Through our parametric model, we link the effect size of pCR with the predicted survivial effect size. This enables us to assess power for treatment comparison in survival outcomes for neoadjuvant trials or similar studies. The results of our simulations show that a pCR proportion reduction can have a modest effect on long-term survival, disregarding any additional benefit the treatment may have. Achieving $\Delta = 0.2$, as per the latest FDA guidance [38] would correspond to an average hazard ratio of $0.85, 0.81$, and 0.78 for β_1 at 0.4, 0.3, and 0.2 respectively with $\lambda_0 = 0.08$ and $\delta = 0.3$. These treatment effects would require sample sizes of 4486, 3042, and 2136 respectively to achieve 80% power for a confirmatory trial with a survival endpoint and 5 years of follow-up following surgery. A similar approach of using subpopulation estimates for sample size determination, based on Berry and Hudis [7], is currently being used in the I-SPY 3 Trials and is described elsewhere [6, 14].

Our simulations offer a possible explanation why the recent meta-analysis failed to validate pCR as a surrogate endpoint. As the authors of that study note, most of the clinical trials included in the meta-analysis failed to increase pCR by more than 10 percent. Based on our simulations even a sample size of $2,000$ patients, which most of the trials did not recruit, would yield insufficient power to detect a statistically significant difference in survival within five years. This can explain why no trend was observed in the meta-analysis, even if pCR were a valid surrogate endpoint.

Furthermore, the results of our Model 2 show that effect size is highly sensitive to the value of β_2. Therefore, the magnitude of binary pCR increased by treatment may not be enough by itself to predict accurately the effect size of survival. This phenomenon is apparent comparing the results of B-27 to B-40. Although B-27 had double the effect size on pCR compared to B-40 (0.132 vs. 0.063) the estimated magnitude of the effect size for B-40 was much greater that than observed in B-27 (HR of 0.80 vs. 0.90).

One limitation of our paper is that it assumes the exponential distribution for survival. This is the most basic parametric form for time-to-event data but may not fit a particular dataset well. Further research is necessary to use more flexible models for the survival functions. Additionally, both of our models assume that pCR is a valid surrogate for survival,

either totally in Model 1 or partially in Model 2. This assumes that any treatment difference in pCR leads to a beneficial change in survival. This is an assumption that is currently not yet verified and is the subject of recent debate [11, 14–16, 25]. As such, this paper does not address the validity of pCR as a surrogate endpoint. Rather, by using our models which assumes pCR is a valid surrogate, well-powered trials can be conducted to definitively answer this critical question.

3.0 PRINCIPAL STRATIFICATION FOR CAUSAL EFFECTS CONDITIONING ON POST-TREATMENT VARIABLES

3.1 INTRODUCTION TO CAUSAL EFFECTS

New interventions are constantly being proposed to improve our health, safety, and society. These can range from new governmental policy proposals to drugs targeted at specific diseases. The evaluation of new interventions is critical in determining which ones are worthwhile to implement. The essential objective of such an evaluation is to show that, compared with standard approach, the new intervention causes or yields some benefit or improvement. The exact definition of this "causal effect" has been formalized by Rubin [33].

Rubin's definition of the "causal effect" makes use of counterfactuals. A counterfactual is a hypothetical quantity which would have been observed if, counter to the facts, some other course of events had taken place. For example, in a randomized clinical trial, a patient assigned to an intervention group has a counterfactual potential outcome had him or her been assigned to the placebo group. Using the notion of counterfactuals, the causal effect of an intervention on a study subject in the intervention group is the difference between the outcome under the intervention and the counterfactual outcome had the intervention not taken place. For example, when one says his headache was "caused" by lack of sleep, he means that had he had adequate sleep he would not have a headache. The essential problem of proving causation, is that we cannot observe both the outcome under an intervention and the outcome without the intervention for each individual. To show that a new drug improves a person's survival we would have to compare the person's survival under treatment to his or her survival without the treatment, and we do not have the benefit of observing both.

Before addressing how such a comparison can be carried out we first introduce the formal definition of a "causal effect" under the framework of Rubin's Causal Model (RCM).

3.2 RUBIN'S CAUSAL MODEL

Let $Z_i \in \{0, 1\}$ be the treatment assignment of subject $i = 1, 2, ..., n$. Let $Y_i(j)$ be the endpoint of interest for subject i under treatment j (possibly counterfactual). For individual i, $Y_i(Z_i)$ represents the observed endpoint value, since subject i is assigned to treatment Z_i.If $Z_i = 1$, then $Y_i(1)$ is observed and $Y_i(0)$ is counterfactual; if $Z_i = 0$, then $Y_i(0)$ is observed and $Y_i(1)$ is counterfactual.

The causal effect on individual i is $Y_i(1) - Y_i(0); i = 1, .., n$. For example, consider a study comparing survival under a new chemotherapy regimen ($Z = 1$) to that under the standard care ($Z = 0$). $Y_i(j)$ is a binary variable indicating survival status after a predetermined number of year following initiation of the study. The individual causal effect on individual i is the difference in survival status of that individual if given the new drug and his or her survival status if given the standard of care. For each subject in the population, we can define the individual causal effect based on the counterfactual outcomes. The average of all these effects, known as the average causal effect (ACE) is:

$$E[Y_i(1) - Y_i(0)]$$

Had $\{Y_i(0), Y_i(1)\}$ been observed for all subjects the ACE can be estimated by:

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$$

In any trial or experiment we cannot observe the causal effect of a given intervention on an individual since we never know his or her unobserved counterfactual outcome. However, with the following two assumptions we can estimate the average (or "typical" in Rubin's words) causal effect by taking the difference between the average response of the two groups:

1. Random assignment: $\Pr(Z_i = 0) = \Pr(Z_j = 0)$ & $Z_i \perp Z_j$ $i, j = 1, ..., n$.

2. Stable Unit Treatment Value Assumption (SUTVA): Treatment assignment of other individuals do not effect the potential outcomes of any individual i.

SUTVA ensures that an individual's potential outcomes is independent of other's assigned treatment. This assumption is usually valid for treatment of non-infectious diseases. For infectious diseases there may be an association between one individual's treatment assignment and another individual's outcome, since the disease can be transmitted between individuals on different treatment arms. Randomization ensures that the average outcome for subjects in each treatment group Z is an unbiased estimate of $\mathrm{E}[Y_i(Z)]$. This is because individuals from each group is representative of the study population. Taking the difference between the average observed outcomes between the two groups leads to an unbiased estimator of the ACE.

Without randomization, the average outcome for each group may be biased in estimating $Y_i(Z)$. The bias can be introduced through the treatment selection process if, for example, the intervention group is healthier on average than the control group. In such a case, an observed effect may be wrongly attributed to the new intervention even though the intervention has no effect on the outcome of interest. Randomization eliminates such bias by comparing two groups with the similar characteristics, measured or unmeasured, except the treatment and outcome of interest.

3.3 POST-TREATMENT VARIABLE ADJUSTMENT OF CAUSAL EFFECTS

Randomized trials allow for the estimation of causal effects by creating two comparable groups with the only difference being treatment assignment. Adopting the notation of the Rubin Causal Model (RCM), Frangakis and Rubin [18] define a causal effect to be a comparison between the potential outcomes on a common set of subjects:

$$\{Y_i(1) : i \in E\} \text{ and } \{Y_i(0) : i \in E\}$$

where E is a subset of the study population. In a randomized trial we estimate the average causal effect of the intent to treat population by taking the difference between the observed outcomes from the two groups. Randomization creates two subsets who are equally representative of the study population. There are situations however, where we would like to condition on a post-treatment variable to study the treatment effect within levels of the post-treatment variable. The following examples were discussed in Frangakis and Rubin [18]:

(1) Investigators would like to estimate the causal effect of a treatment among those individuals who complied with the treatment. Here treatment compliance is the post-treatment variable.

(2) Many individuals drop out from a study with long follow-up. Investigators would like to estimate the causal effect of treatment among those who did not drop out. Here dropout can be considered a post-treatment variable.

(3) The causal effect of a treatment with information on a surrogate marker, such as tumor response or disease progression is of interest. Investigators would like to estimate the causal effect of treatment among those patients with the same surrogate response, for example those whose tumors responded. The surrogate marker is the post-treatment variable in this scenario.

The last example will be discussed in this dissertation, where the relevant surrogate marker and post-treatment variable under consideration is pCR.

3.3.1 Naive Adjustment for Post-treatment Variables

Let $S_i(j) \in \{0, 1\}$ be a binary intermediate variable, for example pCR in neoadjuvant trials, of subject i under treatment j (possibly counterfactual). A naive method of adjusting for post-treatment variables is to compare the outcomes for given values of the post-treatment variable between the two groups:

$$\{i : S_i(1) = s\} \text{ and } \{i : S_i(0) = s\}, s = 0, 1.$$

If treatment has any effect on the post-treatment variable then the above comparison

is not causal, as it violates the condition that the potential outcomes should be compared among the same set of individuals. Specifically, $\{i : S_i(1) = s\}$ and $\{i : S_i(0) = s\}$ are not the same subpopulation since an individual i may have a different intermediate response under treatment and control, that is $S_i(1) \neq S_i(0)$. To illustrate this problem, consider the following data on four individuals' counterfactual data in Table 5. Let $Y_i(z) = 1$ denote a binary outcome for patient i if given treatment Z. Values in parenthesis denote unobservable counterfactals. Here consider the comparison of the observed outcome between treatment

Table 5: Illustration of Naive Adjustment: Sample counterfactual data of 4 patients

i	Z	S(0)	S(1)	Y(0)	Y(1)
1	0	0	(1)	0	(0)
2	0	1	(1)	1	(1)
3	1	(0)	1	(0)	0
4	1	(1)	1	(1)	1

and control for those subjects with S=1. This would be a comparison between patient 2 from $Z = 0$ with those of patients 3 and 4 from $Z = 1$. Such a comparison would yield a conclusion that among the subgroup S=1 treatment group Z=0 had a better outcome since patient 2 has Y(0)=1 and only one (patient 4) has a value of Y(1)=1. In reality, patient 2 should only be compared with patient 4, since both have identical counterfactual characteristics $S_i(0) = S_i(1) = 1$. Such a comparison would lead to a different conclusion that the 2 treatments were equivalent in outcome Y.

Because of this issue, Frangakis and Rubin [18] developed a new framework of principal stratification to allow the introduction of post-treatment variables while maintaining a causal interpretation.

3.4 PRINCIPAL STRATUM EFFECT

For a two-arm trial with a binary intermediate endpoint there are a total of four basic principal strata as defined by Frangakis and Rubin [18]. Each takes the form $(S_i(0), S_i(1))$. Let $E_{jk} = \{i : S_i(0) = j, S_i(1) = k\}$. In the case of pCR, a principal stratum is represented by an individual's potential pCR status under control and under treatment. There are a total of four basic principal strata: $\{ E_{00}, E_{01}, E_{10}, E_{11} \}$. Using our example of pCR as the binary intermediate endpoint, E_{00} consists of individuals who would not achieve pCR, regardless of receiving treatment or control drug. E_{01} consists of individuals who would achieve pCR if they receive treatment but not if they receive control drug. E_{10} consists of individuals who would achieve pCR if they receive control drug, but not they received treatment. E_{11} consists of individuals who would achieve pCR regardless of receiving treatment or control drug.

Each individual falls into only one of these four groups, depending on their counterfactual pCR responses under treatment and under control. Using the principal stratum notation, Frangakis and Rubin [18] define a *principal effect* as a comparison between $Y_i(1)$ and $Y_i(0)$ among individuals i who are members of a specific principal stratum. This effect meets the definition of a causal effect since it is a comparison of the two potential responses among the same subpopulation, namely those belonging to a particular principal stratum. Each principal stratum is by definition independent of treatment assignment since it contains information on counterfactual, or potential outcomes rather than the observed outcome for a specific treatment assignment. Additionally, any union of the four basic principal stratum would also be a valid principal stratum as it leads to comparisons among a common set of individuals. In this thesis, we are interested in comparing potential outcomes among the union of $E_{01} \cup E_{11}$, those that would respond to the treatment regardless of whether or not they would respond under control.

Without additional assumptions, we cannot identify which subjects belong to each of the four basic principal strata and therefore cannot estimate the principal strata causal effects. Frangakis and Rubin [18] suggest two general approaches for estimating principal strata causal effects. One is to incorporate plausible restrictions to identify an individual's principal

stratum membership. For example, they suggest using covariates to predict principal strata membership. Then one can use a maximum likelihood approach to estimate the causal effect. The second is to use the principal stratum framework to conduct a sensitivity analysis for the causal effects. This is done by exploring a range of values for the unobserved quantities and how they impact the estimated causal effect of interest.

3.4.1 Treatment effect among treatment-responders

For our analysis we will compare potential survival outcomes under treatment and control among those individuals who would be pCR responders had they received the treatment:

$$\{Y_i(1)|S_i(1) = 1\} \text{ and } \{Y_i(0)|S_i(1) = 1\} \tag{3.1}$$

Such a comparison has the advantage of excluding patients whose tumors are not responsive to treatment and are unlikely to gain in survival from taking the treatment. By excluding these patients it may be possible to observe a treatment effect that would not have been observed in the overall trial since these excluded subjects could dilute the treatment effect. This can be viewed as a subgroup analysis where this strategy is employed to identify subjects among whom there is a treatment effect. We refer this subset as treatment-responders. A comparison between potential outcomes among treatment-responders would have a causal interpretation.

Without further assumptions, principal stratum (PS) individual membership for any individual i cannot be identified from the observed data because only one of their potential responses can be observed. Further assumptions are needed to identify PS membership so that causal inference conditional on PS strata can be carried out.

3.5 CURRENT APPROACHES TO IDENTIFY PRINCIPAL STRATUM CAUSAL EFFECTS

In this section, we review two recent approaches in the literature that identify causal effects for a given principal stratum. One is a Bayesian approach and the other employs a sensitivity

analysis technique. We will briefly review these approaches in the context of their specific applications before introducing our method.

Li et. al. [26] use a Bayesian approach to analyze data from the Collaborative Initial Glaucoma Treatment Study (CIGTS). This is a randomized study comparing the effects of surgery (Z=1) and medicine (Z=0) on intraocular pressure (IOP), defined as a binary measure with 18mmHg as the cutoff. In their analysis IOP measured at 12 months (S) is considered as a surrogate for IOP measured at 96 months (T). Let $\{s(0), s(1)\}$ be the counterfactual measure of IOP at 12 months with s(0) denoting the IOP value that would be obtained under the control assignment and s(1) denoting the IOP value that would be obtained under the treatment assignment. Likewise, let $\{t(0), t(1)\}$ be a counterfactual measure of IOP at 96 months. Then each patient falls into one of 16 categories describing their counterfactual measure of IOP at 12 months and their counterfactual measure of IOP at 96 months. Table 6 illustrates the 16 possible categories with p_{ij} being the probability of an individual falling into each category.

Table 6: Probabilities of the Counterfactual Model with Binary Intermediate and Outcome Variables

	{t(0),t(1)}			
$\{s(0), s(1)\}$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	p_{11}	p_{12}	p_{13}	p_{14}
(0,1)	p_{21}	p_{22}	p_{23}	p_{24}
(1,0)	p_{31}	p_{32}	p_{33}	p_{34}
(1,1)	p_{41}	p_{42}	p_{43}	p_{44}

A goal is to assess the chance of improvement in IOP at 96 months due to surgery among those individuals who would experience improvement due to surgery in IOP at 12 months. If there is an observed improvement at 96 months among this subset, this would give clinicians assurance that IOP improvement at 12 months is a good surrogate measure

for IOP improvement at 96 months. Using the notation from Table 6, this quantity is $p_{22}/(p_{21} + p_{22} + p_{23} + p_{24})$ in the above table, referred to by the authors as the associated proportion (AP).

Li et al. [26] proposed to estimate AP using a log-linear model. A saturated model requires 15 parameters since there is one restriction that the table probabilities sum to one. However, the data can only provide 6 parameter restrictions, since we can estimate from the observed data $Pr(T = t, S = s|Z)$ for each treatment arm ($4 \cdot 2 = 8$ estimates). Additionally, there is a restriction that $Pr(T = t, S = s|Z)$ sum to 1 within each treatment arm ($8 - 2 = 6$ parameter restrictions). To estimate all parameters of the log-linear model, additional restrictions are required.

To estimate AP, Li et. al. [26] first impose a monotonicity assumption on the data for both IOP at 12 and 96 months. Under this assumption, a patient receiving surgery cannot have a worse IOP at either 12 or 96 months than had they received medication. Formally, this assumes $S_i(1) \geq S_i(0)$ and $T_i(1) \geq T_i(0)$ for all i. In the above table this restriction sets $p_{13}, p_{23}, p_{33}, p_{43}, p_{31}, p_{32}, p_{34}$ to 0. This reduces the free parameters from 15 to 8. Because only 6 parameters are supported by the data the authors assume a prior distribution for these probabilities and employ a Bayesian approach to make inference on the log-linear model. They assume prior distributions on the parameters of the log-linear model and use data augmentation to estimate the parameters.

This approach allows inference on all the model parameters. However, because these estimates are only possible through assuming prior distributions, they can vary based upon the chosen priors. The authors chose priors to incorporate their belief that T is likely to match the values of S [26]. This key assumption leads to the identifiability of the causal parameters. An additional assumption helping identification is monotonicity on the outcome, $T_i(1) \geq T_i(0)$, which assumes the outcome under treatment is no worse than the outcome under control. In many applications it is unreasonable to assume this, since treatment may in fact be inferior to control.

Gilbert et. al. [21] use a sensitivity analysis to make causal inference for a vaccine clinical trial in which HIV infection is a surrogate for viral load. HIV infection, a binary variable, is the post-treatment variable and viral load is the final outcome of interest. Among

those vaccinated some individuals still become infected with HIV. It is believed that in these individuals the vaccine still may be helpful in reducing viral load compared to the viral load had the subject not been vaccinated. To quantify such an effect, it is of interest to estimate the causal effect of a vaccine on viral load among those who became infected with HIV. We cannot simple condition on HIV status since this is a post-randomization variable and any inference of a treatment effect is not guaranteed to be causal. In order to maintain a causal interpretation while conditioning on HIV infection status, the authors condition on the principal stratum of individuals who would be infected with HIV regardless of whether they took the vaccine or not. Without further assumptions, this is not estimable as we cannot determine individual membership to this stratum.

To estimate this causal effect the authors imposed two assumptions. First, they assume monotonicity on HIV infection. This means that any subject that had an HIV infection after taking vaccine is assumed to have had HIV infection had they not taken vaccine as well. Additionally, the authors impose a logistic model for the probability of being infected under vaccine predicted by HIV status under placebo as a binary indicator and observed viral load as a continuous variable. The parameters of this model are not estimable as there are no subjects with data on both potential HIV status under placebo and vaccine. Therefore, the authors employ a sensitivity analysis to estimate the causal effects under a range of parameter values in the logistic model. Shepard et. al [35] extend this sensitivity approach where there is information on baseline covariates. This type of sensitivity analysis is useful to give researchers a range of values for the treatment effect under difference values of the sensitivity parameter. However, it is often unclear what values of the sensitivity parameter to assume, since we cannot obtain estimates of this parameter from the data. Investigator must rely on their intuition and understanding of the disease studied to set plausible values of this parameter.

3.6 A NEW METHOD FOR ESTIMATING A PRINCIPAL STRATUM EFFECT

In this section we develop a new method for estimating a principal stratum effect with the following data setup from a randomized controlled trial. Let $Z_i \in \{0, 1\}$ be the treatment assignment of subject $i = 1, 2, ..., n$. Let $X_i \in \Gamma = \{0, 1, .., l\}$ be an observed baseline categorical variable, consisting of $l + 1$ categories, for subject i. In cases were the baseline variable is continuous, we can create X_i by choosing cutoff values for each category. The cutoffs should be chosen based on scientific knowledge of the variable and so that no category has very few subjects, which would make estimates in that category unstable. Let $S_i(j) \in \{0, 1\}$ be a binary post-randomization intermediate variable (possibly counterfactual) for subject i. Let $Y_i(j) \in \{0, 1\}$ be a binary endpoint of interest for subject i under treatment j (possibly counterfactual). For individual i, $\{S_i(Z_i), Y_i(Z_i)\}$ represents the observed intermediate and final endpoint values. If $Z_i = 1$, then $Y_i(1)$ is observed and $Y_i(0)$ is counterfactual; if $Z_i = 0$, then $Y_i(0)$ is observed and $Y_i(1)$ is counterfactual.

Our goal is to estimate the casual treatment effect among those who would achieve response on the intermediate variable had they been assigned treatment, $\{i : S_i(1) = 1\}$. Formally, we want to estimate the expected treatment difference:

$$E[Y_i(1) - Y_i(0)|S_i(1) = 1] = E[Y_i(1)|S_i(1) = 1] - E[Y_i(0)|S_i(1) = 1] \tag{3.2}$$

This requires us to stratify the study population based upon their potential intermediate response to treatment, $S_i(1)$. Let E_{+1} be the principal stratum of treatment responders, $\{i : S_i(1) = 1\}$. The E_{+1} prinicipal stratum contain the following three subgroups:

$$\{i : Z_i = 1, S_i(1) = 1\}$$

$$\{i : Z_i = 0, S_i(0) = 1, S_i(1) = 1\}$$

$$\{i : Z_i = 0, S_i(0) = 0, S_i(1) = 1\}$$

$\{S_i(1), Y_i(1)\}$ is observable for individuals in the first subgroup of E_{+1}: $\{i : Z_i = 1, S(1) = 1\}$. Because of randomization this subgroup is representative of the entire E_{+1} principal stratum of interest. We can therefore estimate $E[Y_i(1)|S_i(1) = 1]$ by:

$$\hat{p}_1 = \frac{\sum_i I(Z_i = 1, S_i(1) = 1, Y_i(1) = 1)}{\sum_i I(Z_i = 1, S_i(1) = 1)}, \text{ where I() is the indicator function.}$$

This leaves us with the following two control subgroups that need to be identified in order to estimate $E[Y_i(0)|S_i(1) = 1]$:

$$\{i : Z_i = 0, S_i(0) = 1, S_i(1) = 1\}$$

$$\{i : Z_i = 0, S_i(0) = 0, S_i(1) = 1\}$$

. However, for individuals in the control group, we observe $S_i(0)$ and $Y_i(0)$ but cannot observe $S_i(1)$, their potential intermediate outcome had they received treatment. In order to estimate $E[Y_i(0)|S_i(1) = 1]$ we first need to identify which individuals in the control group (Z=0) are part of the E_{+1} principal stratum, or those who would respond to treatment. We solve this problem in different ways for each of the two subgroups. To identify the subgroup $\{i : Z_i = 0, S_i(0) = 1, S_i(1) = 1\}$ we impose the following monotonicity assumption on (Z, S):

$$S_i(0) \leq S_i(1) \tag{3.3}$$

Under the monotonicity assumption (3.3) it is impossible for an individual to achieve response under control but not under treatment: $S_i(0) = 1, S_i(1) = 0$. If a subject responded under control, monotonicity assumes that they would respond if given treatment as well. This makes the principal stratum E_{10} empty. Therefore, any individual in the placebo group with $S_i(0) = 1$ will have $S_i(1) = 1$. This identifies the second subgroup of the E_{+1} principal stratum: $\{i : Z_i = 0, S_i(0) = 1, S_i(1) = 1\}$. What remains is to identify those control subjects who were non-responders but would respond had they been given treatment. These individuals make up the last subgroup $E_{01} : \{i : Z_i = 0, S_i(0) = 0, S_i(1) = 1\}$.

In the following we propose an imputation method to estimate $E[Y_i(0)|i \in E_{+1}]$. First we model the probability of response under treatment for a control subject who did not achieve response. The model will incorporate the baseline covariate X_i and the final outcome $Y_i(0)$:

$$Pr[S_i(1) = 1|S_i(0) = 0, Y_i(0) = j, X_i = x],$$

where $Y_i(0)$ is a binary indicator of the final endpoint with a value of 1 for success and 0 for failure and X_i is the categorical baseline variable of patient i. After estimating the model parameters, we impute the response status under treatment for those control subjects who did not achieve response under control. After imputation we estimate $E[Y_i(0)|i \in E_{+1}]$ as the sample proportion of those who achieve final endpoint under placebo, among the placebo individuals who make up the E_{+1} principal stratum as identified by their imputed values. To account for variability introduced by estimating the model parameters and imputation, we use bootstrap to create confidence intervals.

The motivation of our model choice to identify those control nonresponders who would respond under treatment comes from the following equivalence relation:

$$\begin{aligned} &Pr(S_i(1) = 1|S_i(0) = 0, X_i = x) \\ &= \sum_{j=0}^{1} [Pr(S_i(1) = 1|S_i(0) = 0, Y_i(0) = j, X_i = x)] \\ &\cdot [Pr(Y(0) = j)|S_i(0) = 0, X_i = x];\ x \in \Gamma = \{0, 1, .., l\}. \end{aligned} \tag{3.4}$$

Let

$$G_L(x) = Pr(S_i(1) = 1|S_i(0) = 0, X_i = x)$$

$$G_R(x, j) = Pr(Y_i(0) = j|S_i(0) = 0, X_i = x)$$

$$G_M(x, j; \boldsymbol{\beta}) = Pr(S_i(1) = 1|S_i(0) = 0, Y_i(0) = j, X_i = x),$$

where $\boldsymbol{\beta}$ is a vector of model parameters. We have a different equation for each level of the baseline variable X. In total we have $l + 1$ equations of the form:

$$G_L(x) = \sum_j [G_M(x, j; \boldsymbol{\beta}) \cdot G_R(x, j)], \qquad x \in \Gamma \tag{3.5}$$

The component $G_M(x, j; \boldsymbol{\beta})$ in (3.5) is the model we use to predict $S_i(1)$ for control subjects. This model has $S_i(1)$ as the outcome which is unobserved for all control subjects. Since $S_i(1)$ is never observed for subjects on the control arm, we cannot use standard regression techniques to estimate model parameters. We propose to use the equation system (3.5) to estimate the model parameters, $\boldsymbol{\beta}$. In order to proceed we need to estimate $G_L(x)$ and $G_R(x, j)$ for all $x \in \Gamma, j = 0, 1$. First we consider $G_L(x) = Pr(S_i(1) = 1|S_i(0) = 0, X_i = x)$.

Recall the monotonicity assumption imposed on the data:

$$S_i(0) \leq S_i(1)$$

Because this assumption is on the individual level i, it is true regardless of the tumor size category X of patient i. We therefore have for (Z, S, X):

$$S_i(0) \leq S_i(1), \text{given } X_i = x. \tag{3.6}$$

We use this monotonicity assumption to estimate $G_L(x) = Pr(S_i(1) = 1|S_i(0) = 0, X_i = x)$ as follows. Let

$$E_{jkx} = \{i : S_i(0) = j, S_i(1) = k|X_i = x\}, j, k = 0, 1, x \in \Gamma$$

denote the principal stratum for baseline category x. Because of the monotonicity assumption the principal stratum E_{10x} is empty since there are no individuals with $S(0) = 1$ and $S(1) = 0$. Let

$$p_{jkx} = Pr[E_{jkx}] = Pr(S_i(0) = j, S_i(1) = k|X_i = x)$$

denote the conditional probability of each principal stratum given $x \in \Gamma$. For each x, $Pr[E_{jkx}]$ can be estimated from the observed data. These are derived as solving for the MLEs treating the data as arising from a multinomial distribution with each principal stratum having its own probability of membership. Let N_{zsx} be the total number of subjects with baseline category x with $Z = z, S = s$ with $\sum_{Z;S=0,1;X} N_{zsx} = n$. Then the likelihood function for $(p_{00x}, p_{01x}, p_{11x})$ is:

$$L(p_{00x}, p_{01x}, p_{11x}|N_{00x}, N_{01x}, N_{10x}, N_{11x}) \propto (p_{00x} + p_{01x})^{N_{00x}} p_{11x}^{N_{01x}} p_{00x}^{N_{10x}} (p_{01x} + p_{11x})^{N_{11x}} \tag{3.7}$$

The resulting MLEs are the following:

$$\hat{p}_{00x} = \hat{Pr}[S_i(0) = 0, S_i(1) = 0 | X_i = x)] = \frac{\sum_i [I(S_i(1) = 0, Z_i = 1, X_i = x)]}{\sum_i I(Z_i = 1, X_i = x)} = \frac{N_{10x}}{N_{10x} + N_{11x}}$$

$$\hat{p}_{11x} = \hat{Pr}[S_i(0) = 1, S_i(1) = 1 | X_i = x] = \frac{\sum_i [I(S_i(0) = 1, Z_i = 0, X_i = x)]}{\sum_i I(Z_i = 0, X_i = x)} = \frac{N_{01x}}{N_{00x} + N_{01x}}$$

$$\hat{p}_{01x} = \hat{Pr}[S_i(0) = 0, S_i(1) = 1 | X = x] = 1 - \hat{p}_{00x} - \hat{p}_{11x}.$$

$\hat{p}_{00x}$ is estimated by the proportion of subjects in the treatment arm who did not achieve response among subjects with baseline category x; $\hat{p}_{11x}$ is estimated by the proportion of subjects in the control arm who are responders with baseline category x.

Then for each level of x we have the following nonparametric estimate of $G_L(x)$:

$$\begin{aligned}
\hat{G}_L(x) &= \frac{\hat{Pr}(E_{01}|x))}{\hat{Pr}(S_i(0) = 0|x)} \\
&= \frac{1 - \frac{\sum_i [I(S_i(1)=0, Z_i=1, X_i=x)]}{\sum_i I(Z_i=1, X_i=x)} - \frac{\sum_i [I(S_i(0)=1, Z_i=0, X_i=x)]}{\sum_i I(Z_i=0, X_i=x)}}{\frac{\sum_i [I(S_i(0)=0, Z_i=0, X_i=x)]}{\sum_i I(Z_i=0, X_i=x)}} \\
&= \frac{\hat{p}_{01x}}{\hat{p}_{00x} + \hat{p}_{01x}} \\
&= \frac{\hat{p}_{01x}}{1 - \hat{p}_{11x}}
\end{aligned} \tag{3.8}$$

Next consider the estimation of $G_R(x, j) = Pr(Y_i(0) = j | S_i(0) = 0, X_i = x)$. $\{Y_i(0), S_i(0), X_i\}$ are observed for all control subjects. We therefore have the following nonparametric estimate of $G_R(x)$:

$$\hat{G}_R(x, j) = \frac{\sum_i [I(S_i(0) = 0, Z_i = 0, X_i = x, Y_i(0) = j)]}{\sum_i I(S_i(0) = 0, Z_i = 0, X_i = x)} \tag{3.9}$$

For each $X = x \in \Gamma = \{0, 1, .., l\}$, substituting with the estimates $\{\hat{G}_L(x), \hat{G}_R(x)\}$ in equation (3.4) yields $l + 1$ equations of the form:

$$\hat{G}_L(x) = \sum_j [G_M(x, j; \boldsymbol{\beta}) \cdot \hat{G}_R(x, j)], \qquad x \in \Gamma \tag{3.10}$$

A popular model for the predicted probability of a binary outcome which we assume for our method is the logistic regression model with independent additive effects such as:

$$Pr(S_i(1) = 1 | S_i(0) = 0, Y_i(0) = j, X_i = x) = \frac{exp(\beta_0 + \beta_1 j + \beta_2 x)}{1 + exp(\beta_0 + \beta_1 j + \beta_2 x)} \tag{3.11}$$

Assumption (3.11) is the predicted probability resulting from a logistic regression model with $Y_i(0)$ and X_i as independent predictors of $S_i(1) = 1$. This is the probability of achieving a response ($S_i(1) = 1$) for individuals who did not respond to placebo ($S_i(0) = 0, Z_i = 0$) given their baseline X category and $Y_i(0)$ values. The choice of model is similar to assumption 2 of Gilbert et. al [21]. Here, we assume the independent effect of x is proportional to the category level. Because of this, the unknown parameters are fixed at 3 $(\beta_0, \beta_1, \beta_2)$, but the number of equations in the system depend on the number of categories of X. If there are fewer than three categories, we cannot uniquely solve for all the unknown parameters since there are more unknown parameters than equations in the system. If there are exactly three categories, we can solve for an exact solution for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ using the system of three equations of the form (3.5). However, because of the variability involved in estimating $\hat{G}_L(x)$ and $\hat{G}_R(x, j)$, the solution will be "overfitted" to these estimated quantities. We therefore suggest having more than three categories.

In such a situation we have an overparameterized system of equations. We solve for $\hat{\boldsymbol{\beta}}$ as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{X=0}^{l} \left[\hat{G}_L(x) - \sum_{j=0}^{1} G_M(x, j; \boldsymbol{\beta}) \cdot \hat{G}_R(x, j) \right]^2 \tag{3.12}$$

This minimizes the sum of the squared differences between the two sides of equation (3.10).

3.6.1 Imputation and Estimation

Because

$$S_i(1) \sim \text{Bernoulli}\left(\frac{exp(\beta_0 + \beta_1 j + \beta_2 x)}{1 + exp(\beta_0 + \beta_1 j + \beta_2 x)}\right) \text{ given } \{S_i(0) = 0, Y_i(0) = j, x\} \tag{3.13}$$

we would like to impute $S_i(1)$ among placebo non-responders by drawing from (3.13), with $(\beta_0, \beta_1, \beta_2)$ replaced by the estimated model coefficients $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$. Denote the imputed value $\tilde{S}_i(1)$. After each imputation, we can identify which of these patients fall in the third subgroup of the E_{+1} principal stratum: $\{i : Z_i = 0, S_i(0) = 0, \tilde{S}_i(1) = 1\}$. Those control individuals with $S_i(0) = 1$ are by definition a treatment-responder under the monotonicity

assumption. Therefore, these individuals do not require any imputation for our analysis. After the stochastic imputation, $E[Y_i(0)|S_i(1)=1]$ is estimated by:

$$\hat{p}_0 = \frac{\sum_i I(Z_i = 0, \tilde{S}_i(1) = 1, Y_i(0) = 1)}{\sum_i I(Z_i = 0, \tilde{S}_i(1) = 1)}$$

.

To create confidence intervals with size α around the estimate we use bootstrap with 200 replicates. The lower bound is the $200 \cdot \alpha/2$ quantile and the upper bound is the $200 \cdot (1-\alpha)/2$ quantile. We use bootstrap instead of multiple imputation for the following reason. Multiple imputation is a way to account for the variability of the imputed values. Here, we have a second source of variability: the parameters estimators for the imputation model, $\hat{\boldsymbol{\beta}}$. When using bootstrap, we estimate $\boldsymbol{\beta}$ for each bootstrap sample and then impute based on those values. Therefore, bootstrap accounts for the variability associated with both the model parameters and the imputed values.

3.6.2 Consistency of $\hat{\boldsymbol{\beta}}$

In this section, we provide conditions for our estimator $\hat{\boldsymbol{\beta}}$ to be consistent for $\boldsymbol{\beta}$. We first show that $\hat{\boldsymbol{\beta}}$ can be considered as an extremum estimator as defined by Hayashi [23]. Then we prove that the conditions set forth by Hayashi for consistency are satisfied by our estimator.

Definition of an Extremum Estimator: An estimator $\hat{\theta}$ is an extremum estimator if there a function $Q_n(\theta)$ such that:

$$\hat{\theta} = \arg\max_{\theta} Q_n(\theta);\ \theta \in \Theta \text{ [23]}.$$

One example of an extremum estimator is the maximum likelihood estimator where:

$$Q_n(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

In our method we minimize the objective function,

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &= \sum_{x=0}^{l} Q_n^{(x)}(\boldsymbol{\beta}) \\ &= \sum_{x=0}^{l} \left[\hat{G}_L(x) - \sum_{j=0}^{1} G_M(x, j; \boldsymbol{\beta}) \cdot \hat{G}_R(x, j) \right]^2. \end{aligned}$$

which is equivalent to maximizing $-Q_n(\boldsymbol{\beta})$. Therefore $\hat{\boldsymbol{\beta}}$ is an extremum estimator as defined above.

Next, define:

$$Q_0(\boldsymbol{\beta}) = \sum_{x=0}^{l} Q_0^{(x)}(\boldsymbol{\beta}); x \in \Gamma = \{0, 1, .., l\},$$

where $Q_0^{(x)}(\boldsymbol{\beta}) = \left[G_L(x) - \sum_{j=0}^{1} G_M(x, j; \boldsymbol{\beta}) \cdot G_R(x, j)\right]^2$. We present sufficient conditions for the existence of a unique local minimizer of $Q_0(\boldsymbol{\beta})$.

Lemma: There exists a unique local minimizer $\boldsymbol{\beta_0}$ for $Q_0(\boldsymbol{\beta})$ if:

(1) $Q_0^{(x)}(\boldsymbol{\beta_0}) = 0$, for all x.

(2)

$$rank\left|\frac{\partial Q^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right| \geq dim(\boldsymbol{\beta}),$$

where $Q^*(\boldsymbol{\beta}) = (Q_0^{(0)}(\boldsymbol{\beta}), Q_0^{(1)}(\boldsymbol{\beta}), .., Q_0^{(l)}(\boldsymbol{\beta}))^T$

Proof: From (1) we have that $\boldsymbol{\beta_0}$ minimizes $Q_0(\boldsymbol{\beta})$ since $Q_0(\boldsymbol{\beta}) \geq 0$ for all $\boldsymbol{\beta}$ and $Q_0(\boldsymbol{\beta_0}) = 0$.

Then from condition (2) and the Implicit Function Theorem there exists a unique function $g(\boldsymbol{G_L}, \boldsymbol{G_R})$ such that:

$g(\boldsymbol{G_L}, \boldsymbol{G_R}) = \boldsymbol{\beta_0}$, in the neighborhood of $(\boldsymbol{G_L}, \boldsymbol{G_R})$,

where $(\boldsymbol{G_L}, \boldsymbol{G_R}) = \{G_L(x), G_R(x, j); x \in \{0, 1, ..., l\}, j = 0, 1\}$.

Therefore, $\boldsymbol{\beta_0}$ is a locally unique minimizer of $Q_0(\boldsymbol{\beta})$.

We now show the conditions for consistency.

Theorem: If: (1) $\boldsymbol{\beta} = \boldsymbol{\beta_0}$ is a solution to $Q_0(\boldsymbol{\beta}) = 0$ and

(2)

$$rank\left|\frac{\partial Q^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right| \geq dim(\boldsymbol{\beta}),$$

Then $\hat{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}$.

Proof: From Proposition 7.1 in Hayashi [23]: $\hat{\theta}$ converges in probability to θ if there is a function $Q_0(\theta)$ satisfying the following two conditions:

I.(identification) $Q_0(\theta)$ is uniquely maximized on Θ at $\theta_0 \in \Theta$.

II.(uniform convergence) $Q_n(\cdot)$ converges uniformly in probability to $Q_0(\cdot)$.

From (1) and (2) we can apply the above Lemma to satisfy condition I. To show that condition II is satisfied here, let:

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &= \sum_{x=0}^{l} W_n^{(x)}(\boldsymbol{\beta})^2 \\ &= \sum_{X=0}^{l} \Big[\hat{G}_L(x) - \sum_{j=0}^{1} G_M(x,j;\boldsymbol{\beta}) \cdot \hat{G}_R(x,j)\Big]^2 \\ Q_0(\boldsymbol{\beta}) &= \sum_{x=0}^{l} W_0^{(x)}(\boldsymbol{\beta})^2 \\ &= \sum_{X=0}^{l} \Big[G_L(x) - \sum_{j=0}^{1} G_M(x,j;\boldsymbol{\beta}) \cdot G_R(x,j)\Big]^2 \end{aligned}$$

Then we have:

$$\begin{aligned} &|Q_n^{(x)}(\boldsymbol{\beta}) - Q_0^{(x)}(\boldsymbol{\beta})| \leq \sum_{x=0}^{l} |W_n^{(x)}(\boldsymbol{\beta})^2 - W_0^{(x)}(\boldsymbol{\beta})^2| \\ &= \sum_{x=0}^{l} |W_n^{(x)}(\boldsymbol{\beta}) - W_0^{(x)}(\boldsymbol{\beta})| \cdot |W_n^{(x)}(\boldsymbol{\beta}) + W_0^{(x)}(\boldsymbol{\beta})| \\ &\leq \sum_{x=0}^{l} 2\Big\{|W_n^{(x)}(\boldsymbol{\beta}) - W_0^{(x)}(\boldsymbol{\beta})|\Big\} \end{aligned}$$

since $0 \leq |W_n^{(x)}| \leq 1, 0 \leq |W_0^{(x)}| \leq 1$ since each is a difference of two probability estimates.

$$\begin{aligned} &\leq \sum_{x=0}^{l} 2\Big\{|\hat{G}_L(x) - G_L(x)| + \sum_{j=0}^{1} G_M(x,j;\boldsymbol{\beta}) \cdot |\hat{G}_R(x,j) - G_R(x,j)|\Big\} \\ &\leq \sum_{x=0}^{l} 2\Big\{|\hat{G}_L(x) - G_L(x)| + \sum_{j=0}^{1} |\hat{G}_R(x,j) - G_R(x,j)|\Big\} \end{aligned} \tag{3.14}$$

since $G_M(x,j;\boldsymbol{\beta})$ is bounded by 1 because it is a probability.

Therefore, because

$$\hat{G}_L(x) \xrightarrow{p} G_L(x), \text{ as } n \to \infty$$

$$\hat{G}_R(x,j) \xrightarrow{p} G_R(x,j), \text{ as } n \to \infty$$

we have as a result:

$$Q_n^{(x)}(\boldsymbol{\beta}) \Longrightarrow Q_0^{(x)}, \text{as } n \to \infty,$$

where $\Longrightarrow$ denotes uniform convergence in probability. Uniform convergence is implied since $\boldsymbol{\beta}$ is not involved in the last inequality (3.14). This proves condition II and completes the proof. Therefore $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.

3.6.3 A Summary of the Proposed Method

Here we provide a streamlined summary of our proposed method to estimate:

$$E[Y_i(1) - Y_i(0)|S_i(1) = k], k = 0, 1 \tag{3.15}$$

the causal effect of treatment among patients stratified by their potential response had they been given treatment.

Step 1: Given a dataset $\mathcal{D}$ of a randomized trial with treatment assignment Z_i, baseline covariate X_i, intermediate endpoint S_i, and endpoint of interest Y_i for subjects i=1,2,...,n, set $S_i(1) = 1$ for those with $S_i = 1$ because of the monotonicity assumption.

Step 2: For b = 1, . . . , B:

(a) Draw with replacement observations of the original dataset to create a bootstrap sample $\mathcal{D}^{(b)}$.

(b) Estimate the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ of the logistic regression model (3.11) based on the bootstrap sample $\mathcal{D}^{(b)}$:

$$\hat{\boldsymbol{\beta}^{(b)}} = \arg\min_{\boldsymbol{\beta}} \sum_{X} \left[\hat{G}_L(X)^{(b)} - \sum_{j=0}^{1} G_M(x, j; \boldsymbol{\beta}) \cdot \hat{G}_R(x, j)^{(b)}\right]^2 \tag{3.16}$$

Step 3: Among subjects with $S_i(0) = 0$, impute $S_i(1)$ according to

$$\text{Bernoulli}\left(\frac{exp(\beta_0 + \beta_1 \cdot Y_i(0) + \beta_2 \cdot X_i)}{1 + exp(\beta_0 + \beta_1 \cdot Y_i(0) + \beta_2 \cdot X_i)}\right)$$

with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(b)}$ from *Step 2*. Denote the imputed data as $\{\tilde{S}_i(1)^{(b)}\}$.

Step 4: Estimate $E[Y_i(1) - Y_i(0)|S_i(1) = k]$ by $\hat{\theta}^{(b)} = \hat{p}_1 - \hat{p}_0$ where:

$$\hat{p}_1 = \frac{\sum_i I(Z_i = 1, S_i(1) = k, Y_i(1) = 1)}{\sum_i I(Z_i = 1, S_i(1) = k)}$$

$$\hat{p}_0 = \frac{\sum_i I(Z_i = 0, \tilde{S}_i(1)^{(b)} = k, Y_i(0) = 1)}{\sum_i I(Z_i = 0, \tilde{S}_i(1)^{(b)} = k)}$$

are the estimated proportion of (endpoint) responders under treatment or control for those who would achieve response if given treatment.

Step 5: Compute the empirical bootstrap confidence interval as $I_{emp}(\theta) = (\hat{\theta}^{(b,l)}, \hat{\theta}^{(b,u)})$ where $\hat{\theta}^{(b,l)}$ and $\hat{\theta}^{(b,l)}$ are the empirical $(\alpha/2)$ and $(1 - \alpha/2)$ quantiles from the B bootstrap estimates.

3.7 A SIMULATION STUDY

To evaluate the statistical properties of our proposed estimate we conducted a simulation study. Our setup was chosen to resemble a large phase III neoadjuvant study with data for each patient on baseline tumor category, binary pCR response status, and binary survival status. We simulated 2000 subjects with six variables each,

$$D_i = \{X_i, S_i(0), S_i(1), Y_i(0), Y_i(1), Z_i\},$$

where:

1. X_i denotes baseline categorical tumor size; $X_i \in 0, 1, 2, 3$.
2. $S_i(0)$ denotes pCR response indicator under control assignment. 1 indicates a pCR response and 0 no pCR response.
3. $S_i(1)$ denotes pCR response indicator under treatment assignment. 1 indicates a pCR response and 0 no pCR response.
4. $Y_i(0)$ denotes survival indicator under control assignment. 1 indicates survival and 0 no survival.

5. $Y_i(1)$ denotes survival indicator under treatment assignment. 1 indicates survival and 0 no survival.

6. Z_i denotes treatment assignment. 0=control, 1=treatment.

Of the six variable, two will be counterfactual, For a control subject, $S_i(1), Y_i(1)$ are counterfactual. For a treatment subject $S_i(0), Y_i(0)$ are counterfactual.

We simulate each of the 2000 subjects data as follows. First simulate categorical baseline tumor category from a multinomial distribution with probabilities (.25,.25,.25,.25) for the values (0,1,2,3). Next simulate $S_i(0)$ given the tumor size category from a Bernoulli distribution with $Pr(S_i(0) = 1|X_i = j) = p(j)$ with $p(0, 1, 2, 3) = (.45, .40, .40, .35)$. We then simulate the survival status under control $Y_i(0)$ with a Bernoulli draw with $Pr(Y_i(0)|S_i(0) = 0) = .60$ and $Pr(Y_i(0)|S_i(0) = 1) = .80$ to reflect a 20 percent increased survival probability for pCR control responders over pCR control non-responders. We then simulate $S_i(1)$ given baseline tumor category (X_i), pCR response under placebo ($S_i(0)$) and survival status under placebo ($Y_i(0)$) as follows. For subjects with $S_i(0) = 1$ we set $S_i(1)$ to be 1. This enforces the monotonicity assumption $S_i(0) \leq S_i(1)$. For subjects with $S_i(0) = 0$ we generate $S_i(1)$ from a Bernoulli draw with:

$$Pr(S_i(1)|S_i(0) = 0, X_i, Y_i(0)) = \frac{exp(\beta_0 + \beta_1 Y_i(0) + \beta_2 X_i)}{1 + exp(\beta_0 + \beta_1 Y_i(0) + \beta_2 X_i)},$$

with $(\beta_0, \beta_1, \beta_2) = (-1, 1, -.2)$. We then simulate $Y_i(1)$ given pCR counterfactual information $(S_i(0), S_i(1))$ and counterfactual survival status under control ($Y_i(0)$) as a Bernoulli random variable with the following probabilities of $Y_i(1) = 1$:

$$Pr(Y_i(1) = 1|S_i(0) = 0, S_i(1) = 0, Y_i(0) = 0) = 0.5$$

$$Pr(Y_i(1) = 1|S_i(0) = 0, S_i(1) = 0, Y_i(0) = 1) = 0.6$$

$$Pr(Y_i(1) = 1|S_i(0) = 0, S_i(1) = 1, Y_i(0) = 0) = 0.85$$

$$Pr(Y_i(1) = 1|S_i(0) = 0, S_i(1) = 1, Y_i(0) = 1) = 0.9$$

$$Pr(Y_i(1) = 1|S_i(0) = 1, S_i(1) = 1, Y_i(0) = 0) = 0.85$$

$$Pr(Y_i(1) = 1|S_i(0) = 1, S_i(1) = 1, Y_i(0) = 1) = 0.9$$

These probabilities were chosen to make the survival probability under treatment greater for those who would obtain pCR under treatment but not under placebo and also greater for those patients who would survive under control than those who would not survive under control. We set these probabilities to be independent of baseline tumor status given a patient's counterfactual pCR information and counterfactual survival status under control. Lastly, we simulate a patient's treatment assignment with equal probability for each arm as a Bernoulli draw with $Pr(Z_i = 0) = Pr(Z_i = 1) = 0.5$. In this way we ensure that each subject's counterfactual information is independent of treatment assignment. For the simulated data the true average causal effect for principal stratum $S_i(1) = k$ is:

$$E[Y_i(1) - Y_i(0)|S_i(1) = k] = E[Y_i(1)|S_i(1) = k] - E[Y_i(0)|S_i(1) = k]$$
$$= \frac{Pr[Y_i(1) = 1, S_i(1) = k] - Pr[Y_i(0) = 1, S_i(1) = k]}{Pr[S_i(1) = k]}$$

$$\text{where :} \quad Pr[S_i(1) = k] = \sum_X \sum_{Y_i(0)} \Big\{ Pr[X = x] \cdot Pr[S_i(0) = 1|X = x]$$
$$+ Pr[X = x] \cdot Pr[S_i(0) = 0|X = x]$$
$$\cdot Pr[Y_i(0)|S_i(0) = 0, X = x]$$
$$\cdot Pr[S_i(1) = k|S_i(0) = 0, Y_i(0), X = x] \Big\}$$

$$Pr[Y_i(0) = 1, S_i(1) = k] = \sum_X \Big\{ Pr[X = x] \cdot Pr[S(0) = 1|X = x]$$
$$\cdot Pr[Y_i(0) = 1|S_i(0) = 1, X = x]$$
$$+ Pr[X = x] \cdot Pr[S_i(0) = 0|X = x]$$
$$\cdot Pr[Y_i(0) = 1|S_i(0) = 0, X = x]$$
$$\cdot Pr[S_i(1) = k|S_i(0) = 0, Y_i(0) = 1, X = x] \Big\}$$

$$Pr[Y_i(1) = 1, S_i(1) = k] = \sum_X \sum_{Y_i(0)} \Big\{ Pr[X = x] * Pr[S_i(0) = 1|X = x]$$
$$\cdot Pr[Y_i(0)|S_i(0) = 1, X = x]$$
$$\cdot Pr[Y_i(1) = 1|Y_i(0), S_i(0) = 1, X = x]$$
$$+ Pr[X = x] \cdot Pr[S_i(0) = 0|X = x]$$
$$\cdot Pr[Y_i(0)|S_i(0) = 0, X = x]$$
$$\cdot Pr[S_i(1) = k|S_i(0) = 0, Y_i(0), X = x]$$
$$\cdot Pr[Y_i(1) = 1|S_i(0) = 0, S_i(1) = k, Y_i(0), X = x] \Big\}$$

Under the above simulation parameter settings the true average causal effect for those who would achieve pCR under treatment is .107. This means that if the treatment was administered to all subjects who would achieve pCR under treatment there would be a 10.7%

increasement in survival, within the time frame under consideration, than had all of them taken the control instead.

For the above scenario we conducted 500 replications. For each replicate we followed the method described in Section 3.6.3. We used B=200 bootstrap samples to obtain estimates for each replicate. The R package optim() with the Nelder-Mead method was used to estimate $\boldsymbol{\beta}$. Initial parameter values were chosen to match the true values of $\boldsymbol{\beta}$. In our simulations many of the bootstrap samples did not converge when estimating $\boldsymbol{\beta}$, the parameters used to model the probability of pCR response for a control pCR non-responder. Additionally, the r package optim() reported convergence for extreme values of $\boldsymbol{\beta}$ when in fact the objective function would obtain the same value under more extreme values of $\boldsymbol{\beta}$ as well. We think that this problem of convergence is due to the small sample size of our simulation study and the relatively flat surface of our objective function. We therefore imposed the following two additional restrictions to aid the identification the model parameters. First we rejected any bootstrap sample that did not converge or reported convergence to any parameter $\beta_k; k = 0, 1, 2$ where the value of $\beta_k > 3$. Second, we imposed the restriction $\beta_1 > 0$ by rejecting any bootstrap sample which converged to $\beta_1 \leq 0$. This restricts the probability of pCR under treatment for a control pCR non-responder who survived under control to be greater than the probability for a similar subject who did not survive under control for the same baseline tumor category.

For the purpose of comparison, we compare the performance of our method with two other methods. In the first we impute the missing data $\{S_i(1)\}$ for the control non-responders using the true model parameters from the data generating process. This allows us to evaluate the peformance in a scenario where one could know the true model for the missing data $S_i(1)$. Second, we carry out a sensitivity analysis using similar methods to Gilbert et. al. [21] and Shepard et. al. [35]. Recall that for each $X = x \in \Gamma = \{0, 1, 2, 3\}$, we have an equations of the form:

$$\hat{G}_L(x) = \sum_j [G_M(x, j; \boldsymbol{\beta}) \cdot \hat{G}_R(x, j)], \qquad x \in \Gamma = \{0, 1, 2, 3\} \tag{3.17}$$

where $G_M(x, j; \boldsymbol{\beta}) = \frac{exp(\beta_0+\beta_1 j+\beta_2 x)}{1+exp(\beta_0+\beta_1 j+\beta_2 x)}$. For the sensitivity analysis we set the value of β_1. Then for each category of x we define $\beta_x = \beta_0 + \beta_2 x$. With this reparameterization we have

for each equation only one unknown, β_x. We solve for β_x for each equation independently. We then follow the rest of our method by imputing the missing value $S_i(1)$ for the placebo non-responders, $\{i : S_i(0) = 0\}$ with a Bernoulli draw:

$$S_i(1) \sim \text{Bernoulli}\left(\frac{exp(\beta_x + \beta_1 j)}{1 + exp(\beta_x + \beta_1 j)}\right) \text{ given } S_i(0) = 0, Y_i(0) = j, X_i = x \tag{3.18}$$

. Bootstrap estimates are obtained along with the $100(1-\alpha)\%$ confidence interval.

Table 9 shows the results of the simulation for the study size of 2000 subjects over 500 replicated datasets using the true model for imputation, our method, and a sensitivity analysis setting β_1 to 0,1,or 2. We report the average over all 500 datasets the estimated average casual effect of treatment among those who would respond to treatment ($\hat{\theta}$), the empirical bias of $\hat{\theta}$, the mean squared error, or MSE of $\hat{\theta}$, the 90% confidence interval length, and the 90% confidence interval coverage probability. To account for the uncertainty of the imputation model parameters and the variability of the missing data, we use 200 bootstrap samples to conduct inference. Let $\hat{\theta}^{br}; b = 1, .., 200; r = 1, .., 500$ be the estimate of the bth bootstrap sample from the rth replicate. Then we have:

$$\hat{\theta} = \tfrac{1}{500}\sum_{r=1}^{500}\tfrac{1}{200}\sum_{b=1}^{200}\hat{\theta}^{(br)}$$

$$MSE(\hat{\theta}) = \tfrac{1}{500}\sum_{r=1}^{500}\left\{[\tfrac{1}{200}\sum_{b=1}^{200}\hat{\theta}^{(br)} - \theta_0]^2 + \tfrac{1}{199}\sum_{b=1}^{200}(\hat{\theta}^{(br)} - \hat{\theta}^{(r)})^2\right\}$$

$$\text{Average length of 90\% Confidence Intervals} = \tfrac{1}{500}\sum_{r=1}^{500}\{\hat{\theta}^{(r,.95)} - \hat{\theta}^{(r,.05)}\}$$

$$\text{Coverage probability of 90\% Confidence Intervals} = \tfrac{1}{500}\sum_{r=1}^{500}\left\{I\{\theta_0 \in (\hat{\theta}^{(r,.05)}, \hat{\theta}^{(r,.95)})\}\right\}$$

where $\hat{\theta}^{(r,.05)}$ and $\hat{\theta}^{(r,.95)}$ are the .05 and .95 quantiles of the bootstrap distribution of $\hat{\theta}$ from the rth replicated dataset and $\hat{\theta}^{(r)}$ is the average of the 200 bootstrap estimates,$\hat{\theta}^{(br)}$, from replicate r.

In Table 7 we see that using the true model parameters to impute the missing data produces an unbiased estimate with coverage close to the target. The average confidence interval length for this scenario is .069. For the sensitivity analysis, when β_1 is correctly assumed to be 1, the estimate was unbiased with 90% coverage probability of .884. However, when β_1 is incorrectly specified, estimates where severely biased, depending on the direction of the misspecification error. Additionally, in these scenarios, severe undercoverage resulted. In all sensitivity scenarios confidence interval length was similar to the confidence interval length assuming the true model.

Table 7: Simulation Results of our Method Compared to Assuming True Model and a Sensitivity Approach

	$\hat{\theta}$	empirical bias	MSE	Average length of 90% CIs	coverage probability of 90% CIs
Imputation under $\boldsymbol{\beta} = (-1, 1, -.2)$	.108	.001	6.43e-04	.069	.916
Sensitivity Analysis					
$\beta_1 = 0$	.157	.050	3.29e-03	.072	.300
$\beta_1 = 1$	.108	.001	7.43e-04	.069	.884
$\beta_1 = 2$	.070	-.036	1.92e-03	.066	.434
Our Method	.099	-.008	1.76e-03	.123	.996

Our method produced an estimate with small bias of $-.008$. The MSE of our method was larger than the MSE when imputation was carried out under the true model parameters or with the sensitivity analysis when β_1 was correctly specified. The MSE was superior to the sensitivity analysis when β_1 was misspecified as above. The average confidence interval length under our methos was .123, which is much wider than the other methods. Additionally, our method resulted in overcoverage, while the sensitivity analysis resulted in severe undercoverage when β_1 was misspecified ($\beta_1 = 0, 2$). Li et. al. [26] similarly report overcoverage using their Bayesian technique and attribute it to the lack of full identifiability of their models.

3.8 DATA EXAMPLE: NSABP B-40 NOADJUVANT CLINICAL TRIAL

In this section we apply our method to the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-40 study. In this study, 1206 women with HER2-negative breast cancer were randomized to one of three docetaxel-based neoadjuvant regimens and whether to

receive bevacizumab or not. The purpose of this analysis is to compare the survival probability at a specific time point between the arm with bevacizumab added to the treatment regimen and the arm without bevacizumab among those patient who would obtain a pCR had bevacizumab been added to their treatment regimen.

To apply our method we use baseline tumor size as the additional categorical covariate. We categorize tumor size in 4 categories: less than 3cm, more than 3cm and less than 4cm, more than 4cm and less than 5cm, and more than 5cm. We collapse survival data into a binary endpoint, those who survived past 3 years and those who did not. Using this cutoff 1156 (96%) and 1118 (93%) subjects had complete data for 3-year overall survival and disease-free survival, respectively. Subjects whose observations were censored prior to 3-years of follow up were excluded for the analysis.

We run our method, as well as a sensitivity analysis with β_1 assumed to be 0,1, and 2. We generate 90% confidence intervals using bootstrap resampling with 200 replicates. We use the estimation procedure for our method and sensitivity analysis discussed in 3.6.3.

Table 8: Application in the NSABP B-40 study: treatment effect in 3 year overall survival probability among those who would obtain pCR under treatment

	$E(Y(1)\|\hat{S}(1)=1)$	$E(Y(0)\|\hat{S}(1)=1)$	$\hat{\theta}$	90% CI
Our Method	.974	.941	.032	(-.007, .071)
Sensitivity analysis:				
$\beta_1 = 0$	.974	.931	.042	(.006, .079)
$\beta_1 = 1$	.975	.948	.027	(-.005, .058)
$\beta_1 = 2$	.973	.950	.023	(-.004, .050)

Table 8 shows the results for 3 year OS. Using our method the estimated 3-year OS probability for treatment responders is 0.974 for the bevacizumab group and 0.941 for the non-bevacizumab group, a difference of .032 (90% CI= (-.007, .071)). Because 0 is within the 90% CI, we cannot conclude that adding bevacizumab is beneficial for this subgroup. The sensitivity analysis yielded treatment deferences of .042, .027, and .032 when β_1 was

assumed to be 0, 1, and 2, respectively. For the sensitivity analysis, the 90% CI for both β_1 of 1 and 2 contained 0. For $\beta_1 = 0$ the the 90% CI did not contain 0. Therefore, if an investigator believed that $\beta_1 = 0$, a conclusion of significance at 90% confidence level is reached. $\beta_1 = 0$ signifies no greater chance of pCR under bevacizumab for a subject who did not obtain a pCR without bevacizumab and survived 3 years compared to a similar subject who did not survive 3 years.

Table 9: Application in the NSABP B-40 study: treatment effect in 3 year disease-free survival probability among those who would obtain pCR under treatment

	$E(Y(1)\|\hat{S}(1)=1)$	$E(Y(0)\|\hat{S}(1)=1)$	$\hat{\theta}$	90% CI
Our Method	.920	.859	.061	(-.009, .129)
Sensitivity analysis:				
$\beta_1 = 0$	.920	.846	.073	(.024,.133)
$\beta_1 = 1$	.920	.870	.050	(.001, .106)
$\beta_1 = 2$	.920	.881	.039	(-.018, .088)

Table 9 shows the results for DFS. Using our method the estimated 3-year DFS probability for treatment responders is 0.92 for the bevacizumab group and 0.859 for the non-bevacizumab group, a difference of .061 (90% CI= (-.009, .129)). Because 0 is within the 90% CI, we cannot conclude that adding bevacizumab is beneficial for this subgroup. However, as our method has shown to result in overcoverage, it is very likely that bevacizumab is beneficial, but our method failed to detect this benefit.

The sensitivity analysis yielded treatment deferences of .073, .050, and .039 when β_1 was assumed to be 0, 1, and 2, respectively. Interestingly, the 90% CI under $\beta_1 = 0$ and 1 did not contain 0. If an investigator believed that β_1 was 1 or less, a conclusion of significance at 90% confidence level is reached. The β_1 parameter is the log odds ratio of obtaining a pCR if given bevacizumab for a non-bevacizumab subject who survived past 3 years and did not obtain a pCR compared to one who did not survive past 3-years, within a specific baseline tumor size category. $\beta_1 \leq 1$ indicated the odds of pCR is 2.7 times or less for a

3-year survivor compared to one who did not survive 3 years, given the same baseline tumor size for a subject in the non-bevacizumab group who did not obtain a pCR.

3.9 DISCUSSION AND FUTURE WORKS

In this chapter we presented a method under the principal stratification framework to estimate the causal effect of treatment on a binary endpoint, conditioning on a post-treatment binary response marker in randomized controlled clinical trials. Identification of the causal effect is achieved through two assumptions. First, a subject who responds under control would respond if given treatment. Second, we assume a parametric model for the probability of response for a control non-responder if given treatment. Baseline clinical markers and the binary outcome are predictors in the same model. After estimating the model parameters we impute the missing counterfactual data and use bootstrap samples to conduct statistical inference.

Our method was shown in simulations to have negligible bias, in contrast to a sensitivity analysis-based approach where the sensitivity parameter is misspecified. Our method, however, produced wider confidence intervals than the sensitivity analysis-based approach, or when the true value of the model parameters for the assumed model is known. This can lead to Type II errors where an investigator may wrongly conclude a drug ineffective when in fact there is a treatment effect. However, under our method, if a significant result is achieved, one can conclude the treatment effective, as opposed to a sensitivity approach where one must be wary of a Type I error resulting from misspecifying the sensitivity parameter.

In addition to analyzing breast cancer clinical trial data, our method can be applied to clinical trials testing HIV vaccines. In this context, the causal effect of the vaccine on viral load among those who would be infected under treatment (and under control if monotonicity is assumed) is of interest. To use our method, an investigator would need to categorize viral load into a binary variable with some threshold.

The main weakness of our method comes from the problem of identification of the model parameters used to impute the missing counterfactual information. We believe this to be

primarily due to the flat surface of our objective function. Possible improvements to our method may be obtained by refining the objective function.

APPENDIX

ACCOUNTING FOR CONTINUOUS ACCRUAL AND FOLLOW-UP

A.1 SURVIVAL AND HAZARD FUNCTIONS WITH AN ACCRUAL PERIOD

To model the survival distributions of each treatment group we assume an accrual period of length T_a and a follow-up period after last patient accrual of length T_f. The total time of the study, T is $T_a + T_f$. In practice, most patients survive until time of surgery. We therefore assume that each patients will survive at least 0.5 years after enrollment. This assumption simplifies our calculations as it allows us to assume that all patients have observed pCR status.

Denote the pCR proportion among patients randomized to the control arm as δ. Let Δ be the increase in pCR proportion due to the intervention. Let T_a and T_f be the total accrual time and follow up time after accrual, respectively. We assume the entry time of each patient follows a uniform distribution over $(0, T_a)$. A patient's survival probability is assumed to be 1 for the first .5 years after entry and subsequently follow an exponential distribution with hazard rate λ_0 and λ_1 for pCR non-responders and responders, respectively. The following is the survival function $S_0(t)$ for the control group:

1. For $t \in (.5, T_a)$:

$$\begin{aligned} S_0(t) &= \frac{1}{2t} + (1 - \frac{1}{2t}) \cdot \frac{1}{t - 1/2} \int_0^{t-1/2} \delta e^{-\lambda_1(t-1/2-s)} + (1-\delta) e^{-\lambda_0(t-1/2-s)} ds \\ &= \frac{1}{2t} + \frac{1}{t} \int_0^{t-1/2} \delta e^{-\lambda_1 s} + (1-\delta) e^{-\lambda_0 s} ds \\ &= \frac{1}{2t} + \frac{1}{t} \big[\delta \frac{1}{\lambda_1} (1 - e^{-\lambda_1(t-1/2)}) + (1-\delta) \frac{1}{\lambda_0} (1 - e^{-\lambda_0(t-1/2)}) \big] \end{aligned}$$

2. For $t \in (T_a, T_a + \frac{1}{2})$:

$$\begin{aligned} S_0(t) &= \frac{T_a - (t-1/2)}{T_a} + (\frac{t-1/2}{T_a}) \cdot \frac{1}{t - 1/2} \int_0^{t-1/2} \delta e^{-\lambda_1(t-1/2-s)} + (1-\delta) e^{-\lambda_0(t-1/2-s)} ds \\ &= \frac{T_a - (t-1/2)}{T_a} + \frac{1}{T_a} \big[\delta \frac{1}{\lambda_1} (1 - e^{-\lambda_1(t-1/2)}) + (1-\delta) \frac{1}{\lambda_0} (1 - e^{-\lambda_0(t-1/2)}) \big] \end{aligned}$$

3. For $t > T_a + \frac{1}{2}$:

$$\begin{aligned} S_0(t) &= \frac{1}{T_a} \int_0^{T_a} \delta e^{-\lambda_1(t-1/2-s)} + (1-\delta) e^{-\lambda_0(t-1/2-s)} ds \\ &= \frac{1}{T_a} \big[\delta \frac{1}{\lambda_1} (e^{-\lambda_1(t-T_a-.5)} - e^{-\lambda_1(t-1/2)}) + (1-\delta) \frac{1}{\lambda_0} (e^{-\lambda_0(t-T_a-.5)} - e^{-\lambda_0(t-1/2)}) \big] \end{aligned}$$

The survival function for the treatment group $S_1(t)$ is obtained by replacing δ with $\delta + \Delta$, assuming pCR responders and non-responders in the treatment group have equivalent hazards to those responders and non-responders in the control group. If the treatment group is assumed to have a lower hazard than the control group within each pCR stratum, as in Model 2 of Section 4, λ_0 and λ_1 are replaced by $\lambda_0\beta_2$ and $\lambda_1\beta_2$ for the treatment group's survival. β_2 represents the hazard ratio between treatment and control within each pCR stratum. The hazard function $h_i(t)$ is derived from the survival function through the relation:

$$h_i(t) = -\frac{S_i'(t)}{S_i(t)}$$

For the control group $S_0'(t)$ is as follows:

1. For $t \in (.5, T_a)$:

$$S_0'(t) = -\frac{1}{2t^2} + \frac{1}{t}(\delta e^{-\lambda_1(t-1/2)} + (1-\delta)e^{-\lambda_0(t-1/2)}) - \frac{1}{t^2}\big[\delta\frac{1}{\lambda_1}(1 - e^{-\lambda_1(t-1/2)}) + (1-\delta)\frac{1}{\lambda_0}(1 - e^{-\lambda_0(t-1/2)})\big]$$

2. For $t \in (T_a, T_a + \frac{1}{2})$:

$$S_0'(t) = -\frac{1}{T_a} + \frac{1}{T_a}(\delta e^{-\lambda_1(t-1/2)} + (1-\delta)e^{-\lambda_0(t-1/2)})$$

3. For $t > T_a + \frac{1}{2}$:

$$S_0'(t) = \frac{1}{T_a}\big[\delta(e^{-\lambda_1(t-1/2)} - e^{-\lambda_1(t-2.5)}) + (1-\delta)(e^{-\lambda_0(t-1/2)} - e^{-\lambda_0(t-2.5)})\big]$$

$S_1'(t)$ for the treatment group is obtained by replacing δ with $\delta + \Delta$. The hazard ratio between the treatment and control group is the ratio of the treatments' hazard functions. Because of the complex nature of the hazard ratios we conduct simulations to determine the power of the trial.

A.2 SIMULATION

A.2.1 Setup

We conducted simulations to assess the empirical power for detecting a treatment difference in patient survival for a randomized neoadjuvant trial with a total of N patients allocated equally to two groups. In each run, we first simulated the patient's pCR status by Bernoulli distributions with success probabilities δ and $\delta + \Delta$ for the control and treatment group, respectively. We simulated the entry time,T_{entry} of each patient from a uniform distribution over $(0, T_a)$. We simulated the event time of each patient from an exponential distribution with hazard rates of λ_0, λ_1 for the control group and $\lambda_0\beta_2$, $\lambda_1\beta_2$ for the treatment group for pCR non-responders and responders respectively, with Model 1 simulations corresponding to $\beta_2 = 1$. We added .5 to each event time to account for our assumption that each patient lived until pCR assessment at .5 years. Administrative censoring was considered to reflect the closure of long-term outcome data in multi-center clinical trials. Therefore, an event time beyond an individual's followup time, $T_{follow} = T_{end} - T_{entry}$), was considered censored for purposes of analysis. We compared the survival of the two treatment groups with a 2-sided logrank test with $\alpha = 0.05$ chosen for significance. We repeated each simulation 10,000 times. The proportion of logrank tests that achieved significance will be the empirical power of the randomized trial. All simulation were performed in R Version 3.1.2 [32].

We chose parameter values for our simulation to be similar to the parameter setting of section 3.1; the time of study end of 6.5 was chosen to be 5 years following the mean time of pCR assessment (.5 years after the 1 year mean accrual time). The simulation was carried out for $T_{accrual} = 2, T_{end} = 6.5, \delta = .3$ along with every combination of the following parameters for Model 1: $N = 1000, 2000$; $\lambda_0 =, 0.08, 0.12$; $\Delta = 0.1, 0.2, 0.3$; $\beta_1 = 0.2, 0.3, 0.4$, where $\beta_1 = \lambda_1/\lambda_0$. For Model 2 we ran a simulation with $\Delta = 0.2$ and $\beta_2 = 1, 0.95, 0.9$ assuming $N = 1000, \lambda_0 = .08, \Delta = 0.2$, and $\beta_1 = 0.3$ to evaluate the sensitivity of the model with respect to β_2, the treatment effect within each pCR strata.

A.2.2 Assessment of Sample Size Justification

Similar to section 2.5, we numerically approximate the average hazard ratio. Because we assume each patient lives at least .5 years we approximate the hazard ratio over the interval $(.5, T_{end})$ by averaging the hazard ratio at 1000 time points over $(.5, T_{end})$. We then calculated required sample sizes to achieve 80 and 90 percent power using Schoenfeld's equation (2.1), replacing λ with the calculated average hazard ratio and calculating ρ based on the modeled survival function. We calculated the required sample sizes for a total of 288 combinations of parameter values ($\delta \in \{0.1, 0.3, 0.5\}, \Delta \in \{0.1, 0.2, 0.3, 0.4\}, \lambda_0 \in \{0.08, 0.12, 0.16\}, \beta_1 \in \{0.2, 0.3, 0.4, 0.5\}, \beta_2 \in \{1, 0.9\}$ for $T_{accrual} = 2$ and $T_{end} = 6.5$). Using these sample sizes we calculated the empirical power (with 1,000 runs) under our model assumptions along with its 95% confidence interval using a normal approximation for proportions.

A.2.3 Results

Results of our simulation studies on the empirical power under Model 1 and Model 2 incorporating a continuous accrual process are presented in Table A1 and Table A2.

Compared to the values obtained in Section 4 where a single accrual time was assumed, average hazard ratios and empirical power are nearly identical to those values from Section 4 with the same parameter values. Applying the average hazard ratio to Schoenfeld's equation (2.1) to determine sample size yielded consistent results compared to the empirical power. Overall, empirical power achieved its target, with the average empirical power of 80.5% and 90.2% for 80% and 90% power targets respectively over the 288 scenarios tested. Among the 288 scenarios 93.4 and 91.7 percent of the 95% confidence intervals contained their targets of 80 and 90 percent power, respectively.

Table A1: Average HR and empirical Power with 2 years Accrual and 6.5 years Endtime using Model 1

		$\lambda_0 = .08$			$\lambda_0 = .12$		
			Power			Power	
β_1	Δ	Average HR	N=1000	N=2000	Average HR	N=1000	N=2000
0.4	0.1	0.924	0.097	0.151	0.922	0.125	0.200
0.4	0.2	0.849	0.260	0.457	0.847	0.350	0.602
0.4	0.3	0.776	0.516	0.807	0.774	0.668	0.925
0.3	0.1	0.907	0.122	0.202	0.904	0.158	0.271
0.3	0.2	0.816	0.359	0.615	0.812	0.484	0.778
0.3	0.3	0.727	0.679	0.934	0.724	0.829	0.985
0.2	0.1	0.888	0.153	0.263	0.884	0.207	0.359
0.2	0.2	0.778	0.478	0.765	0.773	0.631	0.906
0.2	0.3	0.672	0.829	0.987	0.667	0.935	0.999

Table A2: Average HR and power with an additional treatment effect within each pCR group with 2 years Accrual and 6.5 years Endtime using Model 2

Δ	β_2	Average HR	Power
0.2	1	0.816	0.359
0.2	0.95	0.776	0.499
0.2	0.9	0.737	0.641
0.3	1	0.727	0.679
0.3	0.95	0.692	0.789
0.3	0.9	0.657	0.876

parameters: $N = 1000$, δ=0.3, λ_0=0.08, T_{end}=5

BIBLIOGRAPHY

[1] James R Anderson, Kevin C Cain, and Richard D Gelber. Analysis of survival by tumor response. *Journal of Clinical Oncology*, 1(11):710–719, 1983.

[2] Harry D Bear, Stewart Anderson, Roy E Smith, Charles E Geyer, Eleftherios P Mamounas, Bernard Fisher, Ann M Brown, Andre Robidoux, Richard Margolese, Morton S Kahlenberg, et al. Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer: National surgical adjuvant breast and bowel project protocol b-27. *Journal of Clinical Oncology*, 24(13):2019–2027, 2006.

[3] Harry D Bear, Gong Tang, Priya Rastogi, Charles E Geyer Jr, Andre Robidoux, James N Atkins, Luis Baez-Diaz, Adam M Brufsky, Rita S Mehta, Louis Fehrenbacher, et al. Bevacizumab added to neoadjuvant chemotherapy for breast cancer. *New England Journal of Medicine*, 366(4):310–320, 2012.

[4] Harry D Bear, Gong Tang, Priya Rastogi, Charles E Geyer, Qing Liu, André Robidoux, Luis Baez-Diaz, Adam M Brufsky, Rita S Mehta, Louis Fehrenbacher, et al. Neoadjuvant plus adjuvant bevacizumab in early breast cancer (nsabp b-40 [nrg oncology]): secondary outcomes of a phase 3, randomised controlled trial. *The Lancet Oncology*, 16(9):1037–1048, 2015.

[5] Joseph Berkson and Robert P Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

[6] Donald A Berry. Right-sizing adjuvant and neoadjuvant clinical trials in breast cancer. *Clinical Cancer Research*, 22(1):3–5, 2016.

[7] Donald A Berry and Clifford A Hudis. Neoadjuvant therapy in breast cancer as a basis for drug approval. *JAMA oncology*, 1(7):875–876, 2015.

[8] John W Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11 (1):15–53, 1949.

[9] Tomasz Burzykowski, Geert Molenberghs, and Marc Buyse. *The evaluation of surrogate endpoints*. Springer Science & Business Media, 2006.

[10] Marc Buyse, Geert Molenberghs, Tomasz Burzykowski, Didier Renard, and Helena Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67, 2000.

[11] Marc Buyse, Tomasz Burzykowski, and Everardo D Saad. Neoadjuvant as future for drug development in breast cancer-letter. *Clinical Cancer Research*, 22(1):268–268, 2016.

[12] Patricia Cortazar, Lijun Zhang, Michael Untch, Keyur Mehta, Joseph P Costantino, Norman Wolmark, Hervé Bonnefoi, David Cameron, Luca Gianni, Pinuccia Valagussa, et al. Pathological complete response and long-term clinical benefit in breast cancer: the ctneobc pooled analysis. *The Lancet*, 384(9938):164–172, 2014.

[13] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[14] Angela DeMichele, Douglas Yee, Donald A Berry, Kathy S Albain, Christopher C Benz, Judy Boughey, Meredith Buxton, Stephen K Chia, Amy J Chien, Stephen Y Chui, et al. The neoadjuvant model is still the future for drug development in breast cancer. *Clinical Cancer Research*, 21(13):2911–2915, 2015.

[15] Angela DeMichele, Douglas Yee, Melissa Paoloni, Don Berry, and Laura J Esserman. Neoadjuvant as future for drug development in breast cancer-response. *Clinical Cancer Research*, 22(1):269–269, 2016.

[16] Helena Earl, Elena Provenzano, Jean Abraham, Janet Dunn, Anne-Laure Vallier, Ioannis Gounaris, and Louise Hiller. Neoadjuvant trials in early breast cancer: pathological response at surgery and correlation to longer term outcomes–what does it all mean? *BMC Medicine*, 13, 2015.

[17] Thomas R Fleming and David L DeMets. Surrogate end points in clinical trials: are we being misled? *Annals of internal medicine*, 125(7):605–613, 1996.

[18] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[19] 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.

[20] ME Ghitany and Ross A Maller. Asymptotic results for exponential mixture models with long-term survivors. *Statistics: A Journal of Theoretical and Applied Statistics*, 23 (4):321–336, 1992.

[21] Peter B Gilbert, Ronald J Bosch, and Michael G Hudgens. Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59 (3):531–541, 2003.

[22] Christos Hatzis, W Fraser Symmans, Ya Zhang, Rebekah E Gould, Stacy L Moulder, Kelly K Hunt, Maysa Abu-Khalaf, Erin W Hofstatter, Donald Lannin, Anees B Chagpar, et al. Relationship between complete pathologic response to neoadjuvant chemotherapy and survival in triple-negative breast cancer. *Clinical Cancer Research*, 22(1):26–33, 2016.

[23] Fumio Hayashi. Econometrics. princeton. *New Jersey, USA: Princeton University*, 2000.

[24] Cardiac Arrhythmia Suppression Trial II Investigators et al. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med*, pages 227–33, 1992.

[25] EL Korn, MC Sachs, and LM McShane. Statistical controversies in clinical research: assessing pathologic complete response as a trial-level surrogate end point for early-stage breast cancer. *Annals of Oncology*, page 507, 2015.

[26] Yun Li, Jeremy MG Taylor, and Michael R Elliott. A bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*, 66(2):523–531, 2010.

[27] Roderick JA Little and Donald B Rubin. Statistical Analysis with Missing Data, 2002.

[28] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, 50(3):163–170, 1966.

[29] Geert Molenberghs, Marc Buyse, and Tomasz Burzykowski. The history of surrogate endpoint validation. In *The evaluation of surrogate endpoints*. Springer Science & Business Media, 2006.

[30] Geert Molenberghs, Marc Buyse, and Tomasz Burzykowski. A meta-analytic validation framework for continuous outcomes. In *The evaluation of surrogate endpoints*. Springer Science & Business Media, 2006.

[31] Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

[32] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

[33] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[34] David A Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrics*, pages 499–503, 1983.

[35] Bryan E Shepherd, Peter B Gilbert, Yannis Jemiai, and Andrea Rotnitzky. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization,

conditional on covariates, with application to hiv vaccine trials. *Biometrics*, 62(2): 332–342, 2006.

[36] Judy P Sy and Jeremy MG Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[37] Cardiac Arrhythmia Suppression Trial. Investigators. preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med*, 321(6):406–412, 1989.

[38] US Food and Drug Administration. Guidance for industry: Pathologic complete response in neoadjuvant treatment of high-risk early-stage breast cancer-use as an endpoint to support accelerated approval. , 2014.

[39] Gunter von Minckwitz, Michael Untch, Jens-Uwe Blohmer, Serban D Costa, Holger Eidtmann, Peter A Fasching, Bernd Gerber, Wolfgang Eiermann, Jörn Hilfrich, Jens Huober, et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *Journal of Clinical Oncology*, pages JCO–2011, 2012.

[40] Janet Wittes, Edward Lakatos, and Jeffrey Probstfield. Surrogate endpoints in clinical trials: cardiovascular diseases. *Statistics in medicine*, 8(4):415–425, 1989.