Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Megan Price Date

Issues in Causal Inference and Applications to Public Health

By

Megan Price

Doctor of Philosophy

Biostatistics and Bioinformatics

Vicki Hertzberg
Adviser

Michael Frankel
Committee Member

Qi Long
Committee Member

Robert Lyles
Committee Member

Lance Waller
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

Issues in Causal Inference and Applications to Public Health

By

Megan Price

M.S., Case Western Reserve University, 2003

B.S., Case Western Reserve University, 2002

Adviser: Vicki Hertzberg, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2009

Abstract

Issues in Causal Inference and Applications to Public Health

By Megan Price

We present three examples of public health research problems for which causal inference methods are better suited than commonly used traditional analytical methods. We expand and generalize our causal inference approaches in systematic ways to provide insight into their potential use beyond these specific motivating examples.

First is adjusting for confounding in observational studies. Although there is a growing trend to use propensity score analyses to confirm results from traditional adjustment methods, there has been little systematic comparison of propensity score and traditional regression adjustment methods, particularly when the majority of confounders are dichotomous variables. This leaves open the question of how to interpret potentially conflicting results from the two methods. We simulate comparison groups with higher and lower frequencies of confounders, and compare the performance of traditional and propensity score methods in terms of estimated treatment effect.

Next, we examine the performance of Frangakis and Rubin's (2002) principle stratification method for estimating treatment effects when outcome measures are 'truncated' by death. In our example from the ProTECT study [Wright et al., 2007] of traumatic brain injury patients, we have the added complication of missing mortality status due to loss to follow-up. We are not aware of any other research that examines the performance of principle stratification analyses when the post-randomization variable upon which stratification is based is missing among some observations. We examine the sensitivity of causal effect estimates to assumptions about the structure of the principle strata themselves versus possible patterns of missingness, and show that, for our example, the former are more influential.

Last, there have been recent efforts to define a prognostic score for stroke and traumatic brain injury patients, to enable tailoring of definitions of 'favorable' outcomes based on a patient's predicted outcome. We propose a new application of Hansen's (2006, 2008) prognostic scoring methods to this problem, and compare our prognostic score results to those generated by prognostic models from the existing literature. We also conduct a formal power analysis comparing analyses using outcomes based on a patient's prognosis versus traditional outcome measures.

Issues in Causal Inference and Applications to Public Health

By

Megan Price

M.S., Case Western Reserve University, 2003

B.S., Case Western Reserve University, 2002

Adviser: Vicki Hertzberg, PhD

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2009

Acknowledgments

I owe many thanks to many people. Memory is an imperfect thing, so if you think your name belongs here, but you find it missing, please know that my gratitude is not lessened, but that the brain cell assigned the task of remembering your name was reassigned sometime over the past six years.

Dr. Vicki Hertzberg - the best advisor a person could ask for. You have good ideas, you are thoughtful about the transition from student to practitioner of the craft, and, most importantly, you always have my back. I could ask for no better leader into the field.

Dr. Qi Long, Dr. Robert Lyles, and Dr. Lance Waller - many thanks to my committee members for all of your constructive feedback. The transition from a draft to a finished document was a challenge, but the final product is so much better thanks to your input.

Brian Schmotzer - my honorary committee member. Thank you for your patience, for your mocking, for your faith, and for your friendship. I never could have done this without you.

Kathy - thank you for picking up the phone all hours of day and night, for being so patient, and for talking me back from the proverbial ledge countless times.

Mark and Carrie - thank you for your presence.

My cohort - thank you for all those nights going over homework problems, all those phone calls, and all those lunches and dinners.

Biostatistics (and Bioinformatics) students, faculty, and staff - thank you for providing a caring and supportive environment for me to come to work in day after day after day. Even on the worst of days, I always knew that I had found the right place to call home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Causal inference is one of the most important, most subtle, and most neglected of all the problems of Statistics.

[Dawid, 1979]

> Problems involving causal inference have dogged at the heels of statistics since its earliest days. Correlation does not imply causation, and yet causal conclusions drawn from carefully designed experiments are often valid. What can a statistical model say about causation?

[Holland, 1986]

1.1 Overview

Causal inference methods have been present in the literature for many years, and yet they are both under-utilized and, all too frequently, improperly applied when they are implemented. This paper presents three specific examples of public health research problems that can (and should) be analyzed using causal inference methods, and then expands and generalizes those methods in systematic ways to provide insight into their potential use beyond these specific examples. Additionally, guidelines are provided for circumstances under which a causal inference method may be preferred to a traditional method or vice versa, and the

specific ways in which results from the different methods may differ, enabling a researcher to make informed analytical choices and correctly interpret potentially conflicting results.

Many research questions of interest in public health present unique methodological and analytical problems to statisticians. In an ideal world, all research would be based on strongly controlled and carefully designed experiments, enabling direct comparisons of groups and estimation of causal effects. However, it is often impossible and/or unethical to structure public health research in this way. For example, with many new treatments it is not clinically defensible to randomize patients into treatment and control groups. Instead, patient's doctors must be able to make individualized treatment decisions. Or, a research question may involve an exposure of interest, rather than a treatment effect, and a researcher may have no control over which individuals are exposed to a high level of air pollution or a potentially contaminated food. Therefore, much public health research is based on observational studies, where treatment and control (or exposed and unexposed) groups may not be directly comparable. In fact, they may differ in systematic ways that could potentially influence an estimation of treatment effect.

In other circumstances, randomization into treatment and control groups may be possible, but patients may be so severely injured, or their illness so significant, that they die before the primary outcome of interest can be measured. In this instance, the way in which an outcome is missing (the observation has been truncated by death) is informative, and analyses need to take this into account.

In yet another example, a researcher may be interested in designing a clinical trial with end points targeted or individualized to a specific group of patients. By linking the definition of a 'good' outcome to a patient's prognosis prior to treatment, it may be possible to design more efficient clinical trials, with the power to identify a significant treatment effect with fewer patients.

All of these examples can be handled using the causal inference framework of counterfactuals. What would that patient's outcome have been if he or she had received treatment instead of control? Would that patient have survived under both treatment and control,

or just treatment? Is it possible for this patient to attain a certain functional status, or based on his or her medical history is there a 'ceiling' past which this patient is unlikely to improve? How does his or her observed outcome compare, not to the general patient population, but to his or her initial prognosis of outcome?

The remainder of this chapter will establish the general causal inference framework (section 1.2) and the specific causal inference methods implemented throughout the rest of this paper (sections 1.3 - 1.6). Chapter two summarizes the current literature on all three methods. Chapter three presents two simulation studies comparing propensity score and traditional regression methods for adjusting for confounding in observational studies. In chapter four we present the results of an application of Zhang and Rubin's [2003] principle stratification technique for data that are truncated due to death. We look at the sensitivity of causal estimates to various assumptions about both the principle strata structure as well as the pattern of data missing due to loss to follow-up. Lastly, in chapter five we present a power analysis comparing the newly developed sliding dichotomy method for designing clinical trials in stroke and traumatic brain injury research to more traditionally defined outcome measures. Data for this power analysis incorporate an application of Hansen's [2006] prognostic scoring methods to the development of predictive models.

1.2 Causal Inference - General

The counterfactual framework mentioned at the end of the previous section has been used as a basic theory of causation in philosophy since the 1700s [Hume, 1748], but the most well-known analysis is Lewis's in 1973, updated and revised in 1999 [Menzies, 2008]. Although the philosophical theory and analysis differs from how statisticians think of causal effect analysis, we do rely on the philosophical notion of counterfactuals as "unactualised possibilities" [Menzies, 2001]. More formally, Pearl describes the case with binary true/false variables x and y and defines a counterfactual as "a probability statement about the truth of y, had x been true, when it is known that y had been false when x was false [Lindley, 2002]." Although the specific mathematical model to estimate causal effects (Rubin's causal

model) was not established until the 1970s [Rubin, 1974], the counterfactual language was being used as early as 1918 by Fisher

> If we say, 'This boy has grown tall because he has been well fed,' we are not merely tracing out the cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter.

The formalization of counterfactuals in statistical analyses was introduced by Neyman [1990] in the context of agricultural experiments.

Causal inference is a challenging topic in statistics, where we have a much longer history of declaring associations between variables with confidence, and tend to translate inference into causal relationships with great trepidation. The first critical difference (of many) between a traditional analysis resulting in association and an analysis resulting in a causal link is the imposition of order on a set of variables. If x is a cause of y, then x should occur chronologically prior to y. Additionally, if x is a cause of y, it becomes conceptually problematic if y could also coherently be considered a cause of x [Holland, 1986].

Holland emphasizes the importance of "measuring the effects of causes" [1986] as the place within concepts of causation that statistics has the most to offer. In setting up a model for causal inference (referred to as 'Rubin's model' in most literature), Holland begins with a population U of individual units u '...on which causes or treatments may act.' [1986] Holland further defines an indicator variable S, with the value $S(u)$ for a specific unit u within the population equal to t for treatment and c for control, in the simplest example. For consistency with later notation, we will instead use Z as the treatment indicator. The outcome variable Y is specified as a 'post-exposure variable' (the only way that it could be an effect of cause Z) and broken down into $Y_t(u)$ and $Y_c(u)$ for the result when unit u receives treatment and control respectively. The key here is that both $Y_t(u)$ and $Y_c(u)$ refer to the *exact same* population unit u. Therefore, "the causal effect of t (relative to c) on u (as measured by Y)" [Holland, 1986] is typically defined to be

$$Y_t(u) - Y_c(u) \tag{1.1}$$

(though we could just as easily estimate alternative comparisons such as ratios). Unfortunately, it is rarely possible to observe both treatment and control on the same unit (the 'Fundamental Problem of Causal Inference' according to Holland); or, as Rubin says [2005], "[e]ach potential outcome is observable, but we can never observe all of them."

Two possible solutions to this problem exist - the first, the 'scientific solution' is rarely applicable in public health research. The scientific solution refers to laboratory-type experiments where it is possible to control conditions such that one experiment conducted on a machine u under treatment and another experiment conducted some time later on machine u under control are considered directly comparable and the fundamental problem has been overcome. The second, the 'statistical conclusion' is what we will be implementing for the rest of this paper. The statistical solution refers to the average causal effect over all the us in U, expressed as

$$E(Y_t - Y_c) \tag{1.2}$$

which we cannot observe. Therefore, we assume equation 1.2 is equal to what we can observe

$$E(Y_t) - E(Y_c). \tag{1.3}$$

Unfortunately, this estimation, and its interpretation, still rely on untestable assumptions. In some instances, the average causal effect may answer the research question of interest. In other cases, we may be interested in the estimated causal effect on an individual unit u. Under the assumption of constant effect the effect of treatment is the same on every unit, thus making the average causal effect applicable to each individual unit and making it possible to draw causal inferences at the unit level.

1.2.1 Independence and the Stable-Unit-Treatment-Value Assumption (SUTVA)

Estimation of the average causal effect implies that the population U is large and that the decision to assign any one unit u to treatment or control is carried out in a random fashion

such that the indicator variable Z is independent of Y_t and Y_c and all other population variables. Then the following equations hold:

$$E(Y_t) = E(Y_t|Z = t) \tag{1.4}$$

$$E(Y_c) = E(Y_c|Z = c). \tag{1.5}$$

Only under SUTVA can we combine equations 1.4 and 1.5 to estimate the causal effect of interest as defined in equation 1.2.

Rubin uses the Stable-Unit-Treatment-Value Assumption (SUTVA) as a useful decision rule for determining "... which questions are formulated well enough to have causal answers [Rubin, 1986]." The assumption is defined for any number N of units u, any number T of treatments t, and some outcome $Y_t(u)$ for each treatment-unit combination. SUTVA is then

> ... the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive, and this holds for all $u = 1, \ldots, N$ and all $t = 1, \ldots, T$. SUTVA is violated when, for example, there exist unrepresented versions of treatments ($Y_t(u)$ depends on which version of treatment t was received) or interference between units ($Y_t(u)$ depends on whether unit u' received treatment t or t') [Rubin, 1986].

We rely on SUTVA as a useful decision rule to respond to a frequently cited Holland claim that "[f]or causal inference, it is critical that each unit be *potentially exposable* to any one of the causes. As an example, the schooling a student receives can be a cause, in our sense, of the student's performance on a test, whereas the student's race or gender cannot." As will be shown in later sections, we propose that when the cause under analysis is the *perception* of an attribute such as race or gender, a conceptual paradigm within which this attribute can be considered a cause has been established. To use Holland's example, while a student's race may not itself be considered a cause of his or her test performance, the causal effect of a teacher's perception of race or gender on the design and grading of a test may be validly estimated. Indeed, Rubin himself provides for this type of analysis

in his application of SUTVA to sex discrimination. Rubin's position is that as long as units, treatments, and outcomes are well-defined, and in such a way that SUTVA applies, causal effects of Holland's so-called attributes (such as race and gender) may be estimated. Rubin's problem with potentially causal statements such as "If the females at firm f had been male, their starting salaries would have averaged 20% higher," is not the attempt to draw causal conclusions regarding gender, but rather the lack of specificity in the definition of the treatment 'gender.' If, on the other hand, one clarified the above statement to indicate that the 'treatment' was "...replacing an 'F' with an 'M' on a job application form," Rubin sees no difficulty in estimating the causal effect of gender [1986].

SUTVA is generally assumed to be true of randomized trials, in that the treatment assignment mechanism is known, and is 'ignorable' (with slightly modified notation to maintain consistency) in the sense that

$$P(Z|X, Y_c, Y_t) = P(Z|X, Y_{obs})$$

where Y_{obs} is technically incomplete data, since the complete data Y consists of both observed and unobserved potential outcomes [Rubin, 2005]. A slightly weaker version of this assumption is applied to observational studies, where treatment assignment may be assumed to be 'strongly ignorable' based on conditional independence assumptions, and defined by Rosenbaum and Rubin [1983a] (with a slight modification to maintain consistent notation):

$$(Y_t, Y_c) \perp Z|X,\ 0 < P(Z = t|X) < 1 \tag{1.6}$$

i.e., after adjusting for covariates the potential outcomes are independent of the treatment assignment, and there do not exist values of x for which treatment (or control) is assigned with certainty. See Rosenbaum [1984] for a discussion of the potential ramifications when this assumption does not hold and ways to test the applicability of assuming strongly ignorable treatment assignment. This will be further addressed in chapter three since these analyses primarily focus on the estimated size of the causal effect, and Rubin warns, "...more careful consideration of the implications of SUTVA is required whenever

sizes of causal effects are of interest or the null hypothesis regarding typical causal effects are to be evaluated, because then actual values under more than one treatment must be contemplated [1986]."

1.3 Propensity Scores - Confounding in Observational Studies

As mentioned in the previous section, conditional independence assumptions (1.6) are required to draw causal conclusions from observational data. One popular causal inference method is the use of propensity scores to adjust for covariates X in 1.6. In essence, it summarizes each patient's covariates compared to the rest of the sample, and then evaluates how similar or dissimilar that observation is to others who received treatment or control. The goal behind propensity score analyses is that within subclassifications of the propensity score, the distributions of covariates X are as similar as possible between treatment and control groups. Another way to think about the propensity score is as teasing out the analogous case-control study hidden in each observational study [Hansen, 2006]. One of the features of propensity scores that also makes them similar to randomization (rather than traditional adjustment for confounding) is a focus on the relationship between covariates and treatment assignment regardless of the relationship between covariates and the outcome of interest [Rubin and Thomas, 2000]. Put another way, propensity score methods differ from traditional methods in that they "...adjust for confounding by modeling aspects of the marginal association of the exposures of interest with the confounders rather than by modeling the independent association of the confounders with the outcome [Robins et al., 1992a]."

The propensity score is the coarsest function of X that fulfills the definition of a balancing score, $b(x)$, namely that for some dichotomous treatment indicator z the "...distribution of x given $b(x)$ is the same for treated ($z = 1$) and control ($z = 0$) units; that is, in Dawid's (1979) notation, $x \perp z|b(x)$ [Rosenbaum and Rubin, 1983a]." The propensity score is then formally defined as

$$e(x_i) = P(Z_i = t|X_i = x_i) \tag{1.7}$$

See appendix for proof [Rosenbaum and Rubin, 1983a] that Z and X are conditionally independent given $e(x)$ and $0 < P(Z = t|e(x)) < 1$.

For a dichotomous treatment indicator, and assuming Z_is are independent given X_is this can be written as:

$$P(Z_1 = z_1, \ldots, Z_N = z_N | X_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{z_i} \{1 - e(x_i)\}^{1-z_i}.$$

This makes it possible to estimate the average causal effect of interest ($E(Y_t - Y_c)$) because under ignorable treatment assignment and propensity score $e(x)$ (a specific balancing score) "the expected difference in observed responses to the two treatments at $[e(x)]$ is equal to the average treatment effect at $[e(x)]$, that is,

$$E\{Y_t|e(x), Z = t\} - E\{Y_c|e(x), Z = c\} = E\{Y_t - Y_c|e(x)\}.''$$

([Rosenbaum and Rubin, 1983a]; with slightly modified notation for internal consistency)

1.3.1 Calculating a Propensity Score

In a randomized trial, the treatment assignment mechanism is known, and therefore the propensity score $e(x)$ has one known definition for that trial. In observational studies the assignment mechanism must be estimated from the observed data, and therefore numerous different definitions of $e(x)$ may be suggested. Translated into Bayesian language, the propensity score may be considered the "... posterior predictive probabilities of assignment to treatment 1 for a unit with vector x of covariates [Rosenbaum and Rubin, 1983a]."

For the case of a dichotomous treatment assignment (only treatment and control groups), the propensity score may be calculated based on a logistic regression model. All pre-treatment covariates should be considered for this model, and any preferred model-building procedure may be used (see chapter 3 of this paper, Rosenbaum and Rubin, 1984, D'Agostino Jr., 1998, among others for specific examples). Once a preliminary model has been chosen, the resulting balance of covariates between the treated and control groups should be assessed,

and the propensity score model updated appropriately until sufficient balance is achieved. Of course, propensity scores can only balance samples for *observed* covariates - only randomization can plausibly account for unobserved covariates as well.

There is some debate in the literature regarding whether propensity score models should include all potential confounding variables or only those that remain significantly associated with the treatment or exposure indicator. As Miettinen [1976] states with regard to a multivariate confounder score (a precursor to the propensity score)

> [w]ith the initial, full model fitted to the data, the statistical significance of the coefficients for many of the (potential) confounding factors is often found to be quite low. In these situations there may be a temptation to reduce the model in a stepwise fashion until all remaining terms have (nominally) 'significant' coefficients. Such reduction of the model would tend to defeat the purpose of multivariate control of confounding, since 'nonsignificance' does not mean lack of confounding, and the deletion of many 'nonsignificant' terms from the model may lead to substantial confounding by the aggregate of the deleted factors. Moreover, the deletion, even if not detrimental, would not serve a purpose of parsimony analogous to that in other contexts. For, the (potential) confounding factors are extraneous to the real issue - conditional association between the exposure and the disease - and no inferences need to be made about these controlled variates.

Miettinen's point highlights the need to base propensity score model decisions on a combination of model goodness of fit and the balance checking techniques outlined in the following section.

1.3.2 Checking Covariate Balance and Evaluating Quality of Propensity Score Model

Propensity scores also provide a level of transparency that is not available from traditional covariate adjustment methods. Propensity score methodology uses the language of 'balance' to refer to associations between treatment and other covariates. The stronger the associations, or the more associations that exist, the more out of balance the two groups are. If a propensity score has been successfully calculated, an intermediate step to assess the newly

achieved balance after adjusting for the propensity score informs a researcher as to whether or not he or she has attained two comparable groups, *prior to ever including the outcome measure in analyses*. If the dataset remains unbalanced, a new propensity score should be calculated, and these steps repeated until adequate balance is achieved. It is inappropriate to move on to the adjustment step in analyses prior to confirming that balance has been achieved. Unlike traditional regression methods, which require the involvement of the outcome measure from the beginning, repetitive calculations of propensity scores do not hinder the final analysis of treatment effect on outcome. As Miettinen states in his summary of the use of multivariate confounder scores [1976], "[t]he adequacy of the control of confounding and the validity of the assessment of the residual association are largely matters of faith, with little opportunity for direct verification, when the analysis is conducted completely under a multivariate model."

Unfortunately, very few studies explicitly address balance-checking and indeed may not conduct this critical step, which leaves the possibility that when propensity score methods are implemented, they may be applied incorrectly. Indeed, Ho et al [2007] argue that it is the balance checking step, as opposed to any theoretical properties, that make propensity scores a useful addition to a researcher's analytical toolbox.

> The propensity score tautology in our view is the main justification for using this technology in practice ... That is, it works when it works, and when it does not work, it does not work (and when it does not work, keep working at it) ... The tautology thus provides a way to make irrelevant the knowledge of whether we have satisfied the conditions necessary to use the theoretical results about the true or consistently estimated score ... At least given the current state of the literature, only the propensity score tautology is useful in practice. Other theoretical results have no bearing on practice.

The literature provides a number of balance checking techniques. In particular, Rosenbaum and Rubin [1984] suggest comparing t-test results for the difference of means of continuous covariates between treatment and control groups before and after subclassifying on the propensity score. They also suggest simple bar charts to demonstrate the frequency distributions of categorical covariates among the propensity score subclassifications. Love [2005]

has numerous suggestions for a well-implemented propensity score method (mostly summarized and updated from Rubin, 2001). These include comparing the standardized differences between treatment and control groups before and after propensity score adjustment,

$$d = \frac{100(\overline{x}_t - \overline{x}_c)}{\sqrt{\frac{s_t^2 + s_c^2}{2}}}$$

for continuous variables and

$$d = \frac{100(p_t - pc)}{\sqrt{\frac{p_t(1-p_t) + p_c(1-p_c)}{2}}}$$

for dichotomous, which Love recommends should be less than 10%. He also suggests checking the ratio of the variance of the logit of the propensity score for treatment versus control, which should be close to one. As Rosenbaum and Rubin [1983a] point out, differing variance structures between groups, both in terms of the propensity score itself and the distribution of covariates, can actually *increase* bias. However, it is also worth noting that the majority of Rubin's guidelines for balance assessment assume continuous covariates, preferably normally distributed, and he explicitly states [2001] that, "[w]ith markedly nonnormal covariates, analogous conditions for reliability of regression adjustment can be more complex. An obvious condition with nonnormally distributed propensity scores is the overlap of distributions of the propensity scores in the two groups."

Regardless of their distribution, propensity scores themselves should always be examined between treatment and control groups to ensure sufficient overlap for comparison, i.e., for each combination of covariates resulting in a given propensity score for a treated individual there should exist a control individual with a similar combination of covariates and therefore a similar propensity score. For extreme values of propensity score, where treatment and control groups differ greatly, additional consideration should be given to the interpretation of results. If a subsection of treated individuals lack comparable control individuals, estimated treatment effects may be skewed. Some suggest omitting these observations, however this obviously reduces sample size, and in some instances may simply be prohibited by the goals of the analysis [Rubin, 2001].

1.3.3 Propensity Score Adjustment

Traditional methods to adjust for confounding include regression adjustment (often referred to as epidemiologic models that include confounders in the final regression model) and matching treatment and control individuals based on identical (or at least similar) confounder values. However, both methods may be problematic in real world applications. Rubin has repeatedly voiced concerns about the possibility for traditional regression adjustment to actually add rather than reduce bias when response surfaces are nonlinear [1979], and stated that vastly differing groups were problematic for regression methods: "[t]he statistical literature has, for many years, warned that regression analysis cannot reliably adjust for differences in observed covariates when there are substantial differences in the distribution of these covariates in the two groups [2001]." As demonstrated by Cochran as early as 1957, in cases where the two comparison groups of interest differ substantially, regression adjustment merely pulls each group's covariates toward a common mean, which may not accurately represent either group's covariate distribution. As for matching techniques, Cochran noted [1972] that as the number of covariates on which it is desirable to match increases, the number of subclassifications within which a sample must contain both treatment and control observations increases exponentially and rapidly becomes untenable. Therefore one of the primary advantages of the propensity score method is that it provides a summary measure for a potentially large number of covariates, without the subclassification difficulties of direct matching on covariate values. Depending on which adjustment method is used, propensity score analyses may suffer from similar bias problems in the presence of non-linear response surfaces as traditional methods.

Rosenbaum and Rubin have shown [1983a] that propensity score analyses can be used to calculate an unbiased estimate of average treatment effect in observational studies. In their initial paper [1983a], only three methods of adjustment were outlined - pair matching, subclassification, and covariance (regression) adjustment. Since then, these three methods have been expanded and added to - observations may be weighted based on inverse probability of treatment assignment (i.e., $1/propensity\ score$) [Robins et al., 2000] or inverse odds ratios [Hirano et al., 2003], and matching may be $1:1$ or $1:k$ (i.e., matching one treated

individual to one control or matching one treated to any number k controls; Rosenbaum and Rubin, 1985, Dehejia and Wahba, 1998, Ho et al., 2007).

When dividing the propensity score into categories, either to include as a categorical covariate in the final regression model or to determine subclasses of the sample itself within which to conduct separate regression analyses, each researcher may make his or her own decisions regarding how many subclasses to choose. However, the use of five subclasses has become somewhat standard, as Cochran showed that when subclassifying on the covariates themselves five groups is often sufficient to reduce bias by 90% [1968] and Rosenbaum and Rubin have shown that five subclasses of the propensity score remove a similar amount of bias [1984].

Despite this wide variety of possible adjustment techniques, the majority of analyses based on propensity scores lack a discussion of how and why one technique was chosen over another. Additionally, most studies in fields other than statistics appear to select Rosenbaum and Rubin's third method, regression adjustment, arguably for its simplicity [Shah et al., 2005]. Propensity score regression adjustment involves simply including a single covariate (the propensity score itself, either as a continuous, raw score or collapsed into categories) in the final regression model instead of all the potential confounders individually used to calculate the propensity score. Unfortunately, the risk of additional bias shown by Rubin [1979] for traditional regression adjustment remains for propensity score regression adjustment. D'Agostino [1998] again reminds us that propensity score regression adjustment should always be conducted with great caution, since this method is particularly sensitive to different variance structures between the treatment and control groups and nonlinear response surfaces.

Fortunately, successful stratification into groups (subclassification) with fairly homogenous propensity scores should result in very little bias, as shown by Rosenbaum and Rubin [1983a]. Assuming that x is conditionally independent of z given some balancing score $e(x)$, then if treatment assignment is strongly ignorable and subclassifications based on $e(x)$ are perfectly homogeneous then the average treatment effect (weighted accordingly if the sample size is unevenly allocated to the strata) is unbiased, since within strata $E(Y_t) - E(Y_c)$ is

calculated based on identical treated and control units. In the more realistic setting, where strata are not exactly homogenous the remaining bias in x can be written as

$$B_s = \sum_{s=1}^{S} w_s \int E(x|b)[P(b|z=1, b \in I_s) - P(b|z=0, b \in I_s)]db \tag{1.8}$$

where b is some balancing score $b(x)$, subclasses I_s defined by the balancing score, and w_s is the direct adjustment weight for each subclass.

It is important to note that 1.8 is an estimate of the remaining imbalance in x between comparison groups, NOT an estimate of potential bias in the final estimate of treatment effect. Equation 1.8 could be used to determine if an estimated propensity score had successfully achieved balance between groups prior to calculating an estimate of treatment effect.

In chapter three we will look more closely at the bias problem in the context of comparing the performance of traditional and propensity score methods while systematically altering the balance of covariates to generate more and less similar comparison groups of interest.

1.4 Principle Stratification

There are many examples in public health, as well as in other fields, of outcome measures that are referred to as missing, but would be more accurately described as truncated. For example, a clinical trial where the outcome measure can only be assessed on those patients who survive to a given endpoint, or an economic study where salary can only be assessed for those who are employed [Zhang et al., 2006], or an education intervention study, where the effect of the intervention on test scores can only be assessed for those who remain in school [Zhang and Rubin, 2003]. All of these examples require a specific type of analysis that accurately takes the truncation of the outcome measure into account.

All of these are also examples of a broader category of problems involving post-treatment covariates. Section 1.2 introduced propensity scores as a way of adjusting for *pre*-treatment variables, and chapter three will present a systematic analysis comparing propensity scores to more traditional regression methods of adjustment. Unfortunately, such traditional meth-

ods, when applied to *post*-treatment variables, do not result in causal estimates. Frangakis and Rubin [2002b] introduced principle stratification, based on values of a post-treatment variable, as a way of organizing the data such that causal estimates may be achieved. The primary advantage of the principle stratification method is that the strata themselves are not affected by treatment, and therefore can be treated as a pre-treatment covariate [Frangakis and Rubin, 2002b]. Chapter four presents an application of principle stratification to what is referred to as the 'truncation due to death' problem, but this method is also frequently used to identify surrogate endpoints and to adjust for compliance problems.

Using a slight modification of the notation of Frangakis and Rubin [2002b], traditional post-treatment covariate adjustment compares

$$P(Y_i^{obs}|S_i^{obs} = s, Z_i = t) \text{ and } P(Y_i^{obs}|S_i^{obs} = s, Z_i = c) \tag{1.9}$$

where Y is the primary outcome of interest, Z is the treatment indicator, and S is the post-treatment covariate. Essentially, this compares the outcomes between the two treatment groups within the same value of some post-treatment variable. However, it fails to take into account that S, just like Y, can be defined in terms of unobserved potential values (counterfactuals) depending on which treatment group a unit is in. In other words, if the post-treatment variable S is affected by treatment, the above comparison will not be a causal effect because the two groups $S_i(t) = s$ and $S_i(c) = s$ may not represent the same group of units. Epidemiologists refer to this as post-treatment selection bias [Frangakis and Rubin, 2002b].

Principle strata are then defined by the pair of counterfactuals $(S_i(t), S_i(c))$. Although we cannot actually observe both, we form strata based on hypothetical groups of units that would possess identical pairs of values to each other, though not necessarily identical pairs of values under each treatment. So, for example, one principle strata may consist of patients who would all develop pneumonia under treatment, but none of whom would develop pneumonia under control. A separate, distinct strata would consist of patients who would contract pneumonia under both treatment and control. Within any one principle

strata, outcome comparisons between treatment groups result in causal estimates since these comparisons are now being made within comparable subsets of units.

1.4.1 Truncation Due to Death

In public health research, when an outcome of interest is missing because a patient died before it could be measured, some researchers treat this observation as simply missing, and omit it from analyses, while others insert an outcome value corresponding to the lowest or worst possible outcome. Both methods have advantages and disadvantages. In the case of omitting the missing, this is an accurate reflection of lack of data, but produces an obvious bias, not to mention reduces sample size. Assigning death the lowest outcome value *may* be an accurate measure of the 'value' of death on the outcome scale, but it may not. Ideally, an analysis would take into account the information provided by the fact that an observation was truncated by death - although this results in a missing outcome value, these data are not missing in the same sense as an observation that is lost to follow-up. We know what happened to that patient, that patient died. It is this train of thought that proposes that placing death on the lowest end of an outcome scale is misleading, since death is not actually located on the same measurement scale as the outcome measure. As Rubin states in reference to a quality of life (QOL) study, "[t]o assign a particular value to QOL when dead is to assume we know how to trade off a particular QOL and being dead (and out of misery). Not only do we not know how to do this, but the trade-off could vary by individual, so we prefer simply to represent the actual truth at this point, and not bring in such extraneous value judgements [2006]."

For example, in a study analyzing the effect of Progesterone on recovery from Traumatic Brain Injury (TBI), the outcome of interest was a measure of functional status, but this outcome was missing for several patients who died before functional status could be assessed [Wright et al., 2007]. Although one could argue that death corresponds to a functional status of zero, this does not necessarily represent the value judgements of individual patients. Indeed, if asked, individual patients may rank death above (as 'preferred') to certain lower levels of functional status, such as permanent vegetative state or severe disability. It is this

philosophical argument, that individual patients may assign different values to death, and this should be reflected in estimates of treatment effect, that motivates Zhang and Rubin's [2003] application of the principle stratification method to these types of problems.

When applied to the truncation due to death problem, principle strata described in the previous section are composed of those who would survive under both treatment and control (labeled LL), those who would die under both treatment and control (DD), those who would live under treatment but die under control (LD), and those who would die under treatment but live under control (DL). Zhang and Rubin [2003] argue that the only true causal effect that can be estimated is that of treatment on outcome within any one strata. Remember from section 1.2 that true causal effects can only be estimated when comparing outcomes on a common subset of units. If principle stratification is not taken into account in problems like these, we will erroneously draw comparisons between mixtures of multiple groups. For example, if we compare outcomes among survivors only in the treatment and control groups, we are actually comparing a mixture of the LL and LD groups among the treated individuals and the LL and DL groups among the control individuals. This will not result in a valid causal estimate of treatment on outcome since the LD and DL group member's mortality is clearly affected by treatment and therefore they differ fundamentally from the LL group. Since complete identification of any of the principle strata is impossible, Zhang and Rubin [2003] propose a method of bounding causal effect estimates, based on a weighted average of the possible distribution of patients and outcomes among the four strata. Calculation of causal effect bounds requires first formally defining the distribution of patients into the four principle strata.

Slightly modifying Zhang and Rubin's [2003] notation to maintain consistency with the current example (and with the analysis of this example in chapter four), the four principle strata are formally defined by their potential outcomes in Table 1.1.

Where $S_i(Z)$ is observed mortality status under treatment or control, $Y_i(Z)$ is observed outcome (DRS) under treatment and control (which is sometimes a valid outcome measure and sometimes truncated by death, indicated by *), and LL, LD, DL, and DD refer to the principle strata - those who would live under both treatments, those who would live

Table 1.1: Principle Strata

Prob. of Principle Stratum Membership	Principle Stratum	$S_i(t)$	$S_i(c)$	$Y_i(t)$	$Y_i(c)$
π_{LL}	LL	1	1	$\in \Re$	$\in \Re$
π_{LD}	LD	1	0	$\in \Re$	*
π_{DL}	DL	0	1	*	$\in \Re$
π_{DD}	DD	0	0	*	*

under control but die under treatment, those who would die under control but live under treatment, and those who would die regardless of treatment, respectively.

Instead, what we observe is Table 1.2

Table 1.2: Principle Strata - Observed

Observed Group	% of population	Principle Stratum	Z_i	$S_i(Z)$	$\overline{Y}_i(Z)$
OBS(tL)	P_{tL}	LL or LD	1	1	$\in \Re$
OBS (tD)	P_{tD}	DD or DL	1	0	*
OBS (cL)	P_{cL}	LL or DL	0	1	$\in \Re$
OBS (cD)	P_{cD}	DD or LD	0	0	*

Once principle strata have been formally defined, the probability of any one individual falling into any one of the principle strata must be bounded, based on the observed groups listed above.

Zhang and Rubin's [2003] original paper deriving these equations was based on an education example where the principle strata consisted of those who would graduate or drop-out of high school. We have modified their notation slightly to match our previously established principle strata based on mortality status. First they use a series of simultaneous equations to bound the proportion of individuals in the DL strata:

$$\pi_{LL} + \pi_{LD} = \frac{\sum I(Z_i = t) I(S_i^{obs} = L)}{\sum I(Z_i = t)} = \frac{\#\, in\, OBS(t, L)}{\#\, Assigned\, Z_i = t} \equiv P_{tL}$$

$$\pi_{DL} + \pi_{DD} = \frac{\sum I(Z_i = t) I(S_i^{obs} = D)}{\sum I(Z_i = t)} = \frac{\#\, in\, OBS(t, D)}{\#\, Assigned\, Z_i = t} \equiv 1 - P_{tL}$$

$$\pi_{LL} + \pi_{DL} = \frac{\sum I(Z_i = c) I(S_i^{obs} = L)}{\sum I(Z_i = c)} = \frac{\#\, in\, OBS(c, L)}{\#\, Assigned\, Z_i = c} \equiv P_{cL}$$

$$\pi_{LD} + \pi_{DD} = \frac{\sum I(Z_i = c) I(S_i^{obs} = D)}{\sum I(Z_i = c)} = \frac{\#\, in\, OBS(c, D)}{\#\, Assigned\, Z_i = c} \equiv 1 - P_{cL}$$

Additionally, since the πs need to be proper probabilities and sum to one, we can say:

$$\pi_{LD} = 1 - \pi_{LL} - \pi_{DD} - \pi_{DL}$$

and if π_{LD} were known, it would be possible to solve for π_{LL} (rearranging the third equation above)

$$\pi_{LL} = P_{cL} - \pi_{DL}$$

and π_{DD} (rearranging the second equation above)

$$\pi_{DD} = 1 - P_{tL} - \pi_{DL}.$$

Solving this system of equations requires restricting at least one of the probabilities. Combining the above three equations, and assuming π_{LD} some known quantity, π_{DL} can be limited:

$$\pi_{LD} = 1 - (P_{cL} - \pi_{DL}) - (1 - P_{tL} - \pi_{DL}) - \pi_{DL} \Rightarrow \pi_{DL} = \pi_{LD} + P_{cL} - P_{tL}$$

so

$$max(0, P_{cL} - P_{tL}) \leq \pi_{DL} \leq min(P_{cL}, 1 - P_{tL})$$

(though alternative probabilities could similarly be limited). Once the distribution of patients into the principle stata has been estimated, the causal effect within the principle strata of interest (in this case, LL) can be estimated. The large sample bounds on $Y_{z,s}$ are determined by estimating "...the maximum value of $[\overline{Y}_{LL}(t)]$ is the average value of Y in the $[\pi_{LL} \backslash (\pi_{LL} + \pi_{LD})]$ fraction of the [OBS(t,L)] group with the largest value of Y. ...the minimum value of $[\overline{Y}_{LL}(t)]$ is the average value of Y in the $[\pi_{LL} \backslash (\pi_{LL} + \pi_{LD})]$ fraction of the [OBS(t,L)] group with the smallest value of Y. [Zhang and Rubin, 2003]." Similar calculations can be done for the control group, and then the lower bound of the average causal effect is the difference between the minimum $\overline{Y}_{LL}(t)$ and maximum $\overline{Y}_{LL}(c)$ and the upper bound is the difference between the maximum $\overline{Y}_{LL}(t)$ and the minimum $\overline{Y}_{LL}(c)$(see table 4.7 for large sample bounds). In other words, the treatment effect is estimated within

the proportion of observed treatment patients who survive who would be expected to be members of the LL strata.

1.4.2 Simplifying Assumptions

The stable-unit-treatment-value assumption described in section 1.1.4 is the first standard assumption necessary to move forward with causal estimates in this case as well. Two additional assumptions are not necessary to conduct causal estimates, but when applicable may narrow the estimated bounds.

First is the monotonicity assumption (referred to as A1 in future chapters and sections), which states that there is no 'denier' or DL group - i.e., no patients would die under treatment but live if assigned to the control group. Under this assumption $\pi_{DL} = 0$ so the new estimations of the remaining πs are:

$$\pi_{LL} = P_{cL}$$

$$\pi_{LD} = P_{tL} - P_{cL}$$

$$\pi_{DD} = 1 - P_{tL}$$

This assumption may be conceptually appealing in clinical trails, since we obviously hope that a treatment does no harm, but it may or may not actually reflect the observed data.

Second is referred to by Zhang and Rubin as 'stochastic dominance' or 'ranked average score.' [2003] This is the case that the group of always survivors (LL) is on average healthier than the rest of the sample, and therefore this subset should be considered to have at least as good of an outcome as the DL group under control and the LD group under treatment. Again, this assumption may or may not be applicable for any specific set of observed data (and is referred to as A2 in future chapters and sections).

A third assumption is what Rubin refers to as the 'exclusion' assumption in the compliance setting. It states that if a treatment has no effect on the intermediate measurement

(mortality status, in our example) then it cannot have any effect on the primary outcome of interest. In other words, if receiving treatment rather than control would not change a patient's likelihood of death, then we also have to assume that it would not change that patient's functional status or quality of life or similar outcome measure. Unfortunately, we are primarily interested in estimating the causal effect of treatment within the group that would always survive, so making this assumption negates the very hypothesis we are trying to test! Rubin argues that although this assumption coupled with monotonicity results in the 'classical instrumental variables estimate,' in the vast majority of cases where truncation due to death occurs, this assumption is simply not applicable [Rubin, 2006]. We will not return to this assumption for the remainder of the paper.

In chapter four we apply Zhang and Rubin's [2003] methodology to the progesterone study mentioned earlier and assess the sensitivity of causal estimates to the principle strata structure assumptions suggested by Zhang and Rubin. Additionally, we extend their method to include covariate and Bayesian analyses.

1.5 Sliding Dichotomy

In the late 1990s researchers began questioning if the reason phase III clinical trials were failing to find successful treatments for ailments such as stroke and traumatic brain injury (TBI) was not due to the absence of successful treatments but rather to poorly defined or specified outcome measures resulting in under-powered trials [Barer, 1998, Maas et al., 1999]. The reasoning was the patient populations were so heterogenous, that not only were analyses often inadequately adjusting for these baseline differences but that outcome measures themselves needed to be adjusted to patient characteristics. In the TBI literature this resulted in the suggestion of a 'sliding dichotomy' by Murray et al [2005], which is an elaboration of Barer's 1998 suggestion. The idea is to develop a prognostic score for individual patients, typically based on covariates such as age, medical history, and illness severity, and use this prognostic score to define new endpoints. Rather than going as far as patient-specific endpoints, Murray et al suggest dividing patients into 'bands,' each of which

will have its own definition of a 'good' outcome (see figure 1.1). Traditionally, three 'bands' or groups are chosen based on tertiles of predicted probability of a favorable outcome.

This approach is clinically appealing, as it reflects the genuine assessment of a patient that occurs in a hospital and is factored into treatment decisions. It also leads to the possibility of better powered (i.e., more efficient) clinical trials. Currently, the standard in TBI or stroke research is to design a phase III clinical trial such that it has 80% to 90% power to detect a treatment effect resulting in a 10 percentage point increase in good outcomes. However, this assumes that every patient enrolled in the trial has an equal probability of improving past some threshold to a 'good' outcome and this is simply clinically untrue [Murray et al., 2005].

There are two ways that the sliding dichotomy can be used to increase the power of a study to detect a significant treatment effect, and therefore reduce the number of patients that must be recruited. One option, suggested by Machado et al [1999] is simply to only recruit patients with an 'intermediate' prognosis of good outcome, eliminating those at either extreme end of the spectrum. This certainly more closely resembles the equal probability of a good outcome implied by the standard design described above, but it also reduces the generalizability of the study results.

Alternatively, one could define a sliding dichotomy, with different definitions of 'good' outcome for groups of patients based on a prognostic score, and then design the study to recruit a certain proportion of the patient population from each of a few groups, say those with poor, intermediate, and good prognoses. This maintains the generalizability of the study while also providing more personalized assignment of a 'good' outcome measure to improve efficiency of the study [Young et al., 2003].

This does, however, present a new problem, which is how to define the prognostic score in a statistically consistent, clinically relevant, way? Currently numerous prognostic models exist in the stroke and TBI literature, but very few have been validated on multiple samples [Perel et al., 2006].

Another gap in the current literature on this topic is a formal treatment of just how much of

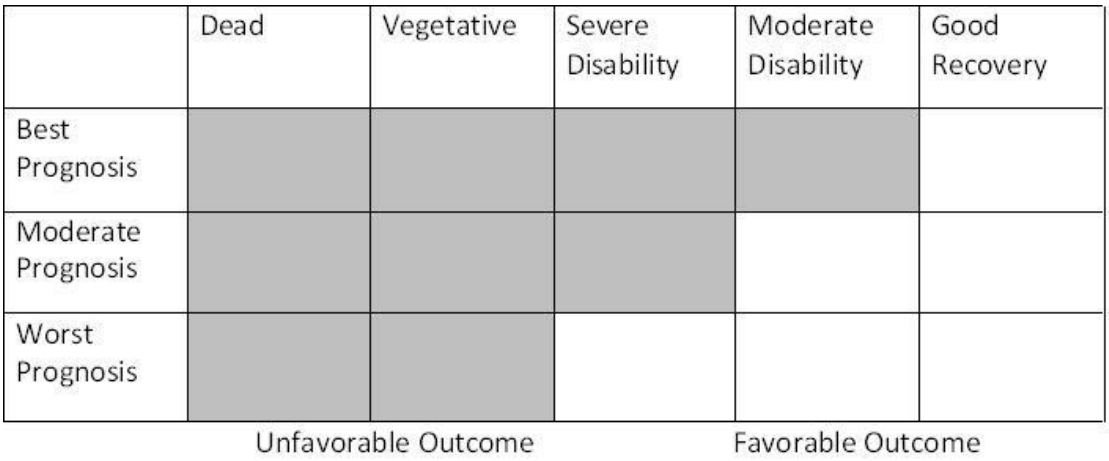

Figure 1.1: Graphical Representation of Sliding Dichotomy - GOS categories on top row, shaded regions indicate outcomes that would be considered 'favorable' or 'unfavorable' for given prognostic category

a gain in power/sample size can be achieved by using a sliding dichotomy approach to define favorable outcomes as compared to a traditionally defined favorable outcome. Currently Machado et al [1999] and Young et al [2003] provide the most complete analyses of this problem, which we explore in more depth in chapter five.

Within one specific research area, it may be possible to achieve a standard predictive model. For example, Glasgow Outcome Scale (GOS) results at three months could always be used to categorize patients into prognostic groups within which to define 'good' outcomes based on change in GOS results at six months. Instead, one general method for estimating a specific predictive model for each study may be preferable. Hansen has been working on prognostic score theory, as an extension of his research in propensity scores [2006, 2008]. Although his motivation is to further balance potential confounders in observational studies, as an assistant measure to the propensity score, we believe that his methods could readily be applied to the sliding dichotomy problem. Additionally, the general theory underlying Hansen's development of prognostic scores (using only the control subjects within a sample to build a model) has a long history of theoretical development [Peters, 1941, Belson, 1956, Cochran, 1969, Rubin, 1984, Gastwirth and Greenhouse, 1995]. The unique application of Hansen's [2008] prognostic score method to develop predictive models to categorize patients

into prognostic groups is developed in chapter five.

1.6 Prognostic Scores

The propensity score described in section 1.3 can be thought of as a summary of the association between treatment assignment and a collection of covariates. Similarly, the prognostic score can be thought of as a summary of the association between a potential outcome and a collection of covariates [Hansen, 2008]. By conditioning on the regression of Y_c on x (where Y_c is the outcome under control and x is a vector of covariates), essentially identifying those covariates associated with outcome among the control group, one can achieve a type of balance like that obtained by conditioning on the propensity score. This idea has its roots in work by Peters [1941], Belson [1956], Cochran [1969], Rubin [1984], and Gastwirth and Greenhouse [1995] who all suggested "...estimating the treatment effect as the treatment group mean of $y_i - \widehat{E}(Y_{ci}|X = x)$ [Hansen, 2008]." In other words, the observed outcome minus the predicted outcome based on parameters estimated from a regression model using only the control patients. Hansen [2008] defines these predictived outcome values as prognostic scores $\Psi(X)$.

Just as Hansen compares achieving propensity balance to identifying the randomized experiment hidden in an observational study [2008], he likewise compares the application of prognostic scores to scientific experimentation:

> In a second experimental ideal, it is the process by which outcomes are generated that is repeatable, understood, and carefully controlled, not the process of assigning units to treatment. Studies approaching this idea use experimental control in the interest of removing associations between covariates and potential outcomes, not treatment assignment. If in advance of studying a new experimental manipulation, an investigator conducts tests absent the new manipulation in order to better understand accompanying conditions and their influence on the outcome, then it is this second ideal that her procedure seeks to attain.

Hansen states that "[p]rinciples of sufficiency and of conditional independence support a theory of prognostic balance that parallels Rosenbaum and Rubin's [1983a] account of

propensity balance, with a few important differences [2008]." In general, if

$$Y_c \perp X | \Psi(X), X \epsilon A \tag{1.10}$$

for any measurable set A, then $\Psi(X)$ is a prognostic score, and achieves 'prognostic balance' along the outcome measure for given X in the same sense that a propensity score achieves balance along the treatment assignment for given X. Unfortunately, this prognostic balance cannot be checked across the entire sample (only the control observations), so unlike the complete-sample balance checking that is possible with the propensity score, it may be difficult to evaluate the quality of a prognostic score $\Psi(X)$. However, it does still share the favorable property that multiple models may be assessed without revealing the potential treatment effect.

Hansen then defines the average causal effect of interest (1.2), absent effect modification, as

$$E(Y_t - Y_c) = E[E(Y|Z = t, \Psi(X)) - E(Y|Z = c, \Psi(X))] \tag{1.11}$$

if $Y_c \perp Z|X$ and $P(Z = 1|\Psi(X)) < 1$ with probability one (similar to the SUTVA conditions for a propensity score $e(x)$).

Hansen suggests using the prognostic score as an additional tool for balancing observational data, in much the same way as a propensity score is used - as a covariate in regression adjustment, as a matching metric, and/or for calculating weights. It is possible that adjusting for a prognostic score could improve the analyses of TBI and stroke clinical trials, but we are primarily interested in applying Hansen's method to calculate predicted probabilities for defining prognostic categories within which to specify different outcome cutpoints. This work, in addition to a formal power analysis of the sliding dichotomy approach, is described in chapter five.

Chapter 2

Literature Review

2.1 Propensity Scores

The theoretical groundwork for propensity scores was laid by Rosenbaum and Rubin in a series of papers [1983a, 1983b, 1984, 1984]. The method has been used in a variety of fields over the past two decades [Imai and van Dyk, 2004, Grunkemeier et al., 2002], with a growing body of literature expanding on these initial applications and analyzing the performance of propensity score analyses under a variety of circumstances. However, there currently still lacks a consensus regarding whether and how the estimated treatment effect size differs between propensity score and traditional adjustment methods, particularly when the confounders of interest are dichotomous.

Robins et al [1992b] generalized propensity scores from the case of two groups (treatment and control or exposed and unexposed) to continuous, ordinal, or discrete treatments or exposures. Drake [1993] conducted simulations to compare different model specifications for the propensity model to traditional linear regression adjustment (with two normally distributed covariates), Dehejia and Wahba [1999] conducted a sensitivity analysis of propensity score performance under varying model specifications and variable selections, Cepeda et al [2003] and Austin et al [2007] compared propensity score analyses to traditional logistic regression, and Kang and Schafer compared the performance of traditional and propensity

score adjustment methods to doubly robust methods [2007].

Recent papers outside the statistics literature have compared traditional and propensity score methods in specific case studies [Austin and Mamdani, 2006, Posner et al., 2001], examined potentially biased results when estimating hazard and odds ratios using propensity score methods [Austin et al., 2007], compared different propensity score methods to each other [Kurth et al., 2006], and literature reviews have summarized recent usage of traditional versus propensity score methods [Shah et al., 2005, Sturmer et al., 2006] and summarized the use of propensity score methods in specific fields, providing some basic guidelines for their implementation [Glynn et al., 2006].

The literature reviews by Shah and Sturmer focused on publications including both propensity score and traditional methods, and compared whether or not a significant effect was detected with each method. Between these two reviews, more than 200 publications were summarized. Both reviews found few differences between traditional and propensity score methods and claim that propensity score analyses have a tendency to be more conservative, i.e., propensity score analyses detect a significant effect less often than traditional methods. Additionally, Sturmer [2006] looked at the size of the estimated treatment effects and found only nine examples (13% of their literature review) in which estimated effect sizes from propensity score methods differed from regression model estimates by more than 20%. However, given the lack of detailed information provided by most articles regarding model selection and propensity score estimation, it is difficult to determine how comparably these two methods were carried out in individual cases.

Cook and Goldman [1989] also compared traditional and propensity score methods, with respect to logistic regression, looking to test the hypothesis that propensity score analyses had the potential to exaggerate significance levels when confounders were highly correlated with the exposure of interest (which is the opposite of Drake's [1993] findings regarding linear regression). However, through both a case study and a series of simulations, they were able to show that significance levels (as determined by p-values associated with the exposure variable of interest) were distributed as expected, except in cases where the simulated correlation between confounders and the exposure variable was so extreme as to be highly

unlikely to occur in actual circumstances (squared multiple correlation coefficients of greater than 0.9).

Since Rubin and Thomas suggested that more statistical work needed to be done to evaluate the performance of propensity score methods [1997], many additional analyses have been performed, but these appear to be overwhelmingly in the subset of propensity score methods involving matching. For example, in the same article where Rubin and Thomas make their suggestion [1997] they conduct a simulation study on the performance of propensity score matching in "... practical settings when the conditional variance is quadratic rather than constant and the propensity score matching is not exact." An earlier paper by Gu and Rosenbaum [1993] compared different matching procedures and offered practical data analysis advice, and later Zhao [2004] also compared the performance and data requirements of different matching metrics. Likewise, Hansen compares several different matching methods for estimating the effect of coaching on SAT scores [2004]. Similar guidelines are needed for regression adjustment analyses, since this method appears to be even more commonly implemented outside of the statistics literature.

Also missing from the existing propensity score literature is any modeling that strays from assumptions of normally distributed confounders. A notable exception is Rosenbaum and Rubin's [1983b] examination of the robustness of causal conclusions to the omission of a binary covariate. This analysis is somewhat similar to the simulations in chapter three, except that Rosenbaum and Rubin compare their results under varying assumptions regarding the relationship between the omitted covariate and the outcome of interest and treatment assignment, all under propensity score analyses. No comparison is made to traditional adjustment methods under comparable circumstances.

2.2 Truncation Due to Death/Principle Stratification

The principle stratification approach is commonly applied to compliance issues in randomized trials [Jin and Rubin, 2008, Frangakis et al., 2002] but much less often applied to truncation due to death problems. Indeed, although the principle stratification framework

was not formalized until 2002 (Frangakis and Rubin), Imbens and Rubin [1997] used principle stratification-type language in their labeling of complier, never-taker, always taker, and defier groups in their compliance problem. Zhang and Rubin first applied the principle stratification framework to the truncation due to death problem [2003], Imai later showed that the bounds derived in the Zhang and Rubin paper are sharp bounds [2008], and Zhang, Rubin, and Mealli extended this approach to include Bayesian analyses [2006]. In a journal article summarizing two public lectures, Rubin [2006] outlined the principle stratification approach and the kinds of truncation due to death problems to which it is suited. Mattei and Mealli [2007] applied the principle stratification framework to a randomized trial of breast self-examination, with the added complication of noncompliance. Matsuyama and Morita [2006] use principle stratification to estimate the local average treatment effect of chemotherapy in a clinical trial studying non-small-cell lung cancer. MacKenzie et al [2007, 2008] used an implied principle stratification approach to their truncation due to death problem in assessing the impact of trauma-center care on functional outcome, but unfortunately insufficient details were provided regarding how the principle strata (and treatment effect) were estimated. The principle stratification approach has also been used to identify surrogate endpoints ([Frangakis and Rubin, 2002a, Mealli and Rubin, 2003, Weir and Walley, 2006] among others).

Prior to developing their principle stratification approach, Frangakis and Rubin [1999] established problems with standard intention to treat analyses in the presence of noncompliance and resulting missing outcome data. Similarly, Chen, Liu, and Zhang [2005] outline alternative adjustment methods for post-randomization covariates and quantify the potential bias of an estimate using traditional regression adjustment. Kurland, Johnson, and Diehr [2007] summarize the different types of research questions that are best suited to unconditional, fully conditional, partly conditional, and joint models of outcome data truncated by death. Kurland et al mention principle stratification as an alternative to the models they describe. Similarly, McConnell et al [2008] include an overview of principle stratification methods in their review of possible approaches to the truncation due to death problem.

More commonly, the truncation due to death problem arises in longitudinal studies, where

researchers approach the problem by modeling drop-out patterns both due to loss to follow-up and death [Diehr et al., 1995, Pauler et al., 2003, Dufouil et al., 2004, Kurland and Heagerty, 2005, Harel et al., 2007] . Methods include conditioning on mortality status and time to death. Interestingly, in longitudinal studies it is fairly common to (attempt to) take both loss to follow-up and truncation due to death into account, but in single time point datasets (such as the ProTECT study analyzed in chapter four) with the application of principle stratification, we are not aware of any analyses that attempt to take these two types of 'missing' data into account.

Although not directly applicable to the problem presented in chapter four, Cheng and Small [2006] develop confidence intervals for their causal bounds based on principle stratification in a three-arm randomized trial with noncompliance. Developing confidence intervals for our own point estimate bounds is an important next step, discussed in chapter four.

Lastly, Frangakis et al [2007] have begun to broaden the category of missing covariates in this type of problem, and propose a principle stratification framework for when *input* data (rather than outcome measures) are missing due to truncation due to death.

The literature also includes several alternative methods for analyzing this type of problem. Although initially based on the principle stratification framework, Egleston et al [2007] and O'Malley and Normand [2005] develop an alternative method (to Zhang and Rubin, 2003) for estimating treatment effects under truncation due to death/noncompliance based on maximum likelihood estimation. Robins [1998] suggests alternative methods to handling noncompliance using structural nested models, non-nested marginal structural models, and continuous-time structural nested models. Hayden, Pauler, and Schoenfeld [2005] develop an alternative estimate of the survivor average causal effect, not explicitly involving principle stratification. Other researchers have begun to include a death category in their outcome measure, and Diehr et al [2005] review the appropriateness of this choice for a few commonly used quality of life and functional status scales.

2.3 Prognostic Scores/Sliding Dichotomy

Much work in this area has been conducted in both the stroke and traumatic brain injury (TBI) literature, because these two fields utilize many of the same functional outcome scales and similar definitions of favorable outcomes. Additionally, both fields suffer from similar challenges in detecting significant treatment effects in clinical trials.

As mentioned in section 1.5, Barer [1998] initially suggested that stroke mega-trials might be missing treatment effects due to poorly operationalized outcome measures. Barer argued that "... a standard method of classifying stroke patients into severity groups" within which separate 'good' outcome measures could be defined, could contribute to better designed, and thus smaller, more efficient, clinical trials. In 1999, Maas et al provided "... an overview of the results of recent trials of neuroprotective agents in head injury" and echoed Barer's conclusions regarding stroke research: "... the failure to find statistically significant benefit from various neuroprotective agents in recent trials in head injury does not necessarily mean that such agents are ineffective but may in part be caused by problems in the design and analysis of clinical trials." Maas et al [1999] recommend more research into predictive modeling to "... discover the benefits for design of future trials." Machado, Murray, and Teasdale [1999] likewise evaluated the design of clinical trials in head injury research and concluded that trials targeted at patients with an intermediate prognosis could greatly reduce the required sample size to detect a significant treatment effect. Young, Lees, and Weir [2003, 2005] conducted simulations to show that patient-specific cut points resulted in more power to detect treatment effect, as compared to standard cutpoints uniformly applied across the patient population.

Although the Stroke Treatment with Ancrod Trial (STAT) study did not formally employ a sliding dichotomy method of assessing outcome, it did include patients with prior disabilities, and for that subset of patients, recovery to at least their pre-(most recent) stroke functional status (as measured by the Barthel Index) was considered a favorable outcome [Sherman et al., 2000]. However, since this was not the primary aim of the study, the results and discussion fail to address whether including a flexible definition of 'favorable' for this subset

of patients contributed to the study's ability to detect a significant treatment effect and/or implications for future studies.

The previously mentioned research into more efficiently designed clinical trials has prompted new work on developing and applying predictive models. Mukherjee et al. [2000] developed a mathematical outcome prediction model for head injury consisting of Glasgow Coma Scale (GCS) motor score, brain stem reflexes, and reaction level scale, however the focus here was more on family counseling than designing future trials (i.e., predictive models designed to better inform families about what to expect in the coming days, weeks, and months). Andrews et al [2002] compare decision tree analysis versus logistic regression for identifying predictive variables. Mendelow et al [2003] proposed a formal sliding dichotomy approach for their International Surgical Trial in Intracerebral Haemorrhage (STICH) trial, which consisted of assigning patients to good or poor prognosis groups, with a separate definition of favorable outcome for the two groups. The prognostic score was defined as a linear combination of GCS, age, and volume of intracerebral hemorrhage. In 2005 they reported the results of this study, and unfortunately, even with the "... more sensitive prognosis-based outcome assignment," and a sample size that exceeded "... the total number of patients in all nine previous randomised controlled trials" the results were still inconclusive as to whether surgery improved outcome for patients suffering intracerebral hemorrhage. Weimar et al [2004] set out to externally validate a prognostic model of functional outcome (defined by the Barthel Index) for patients suffering acute cerebral ischemia. Their model included age and National Institutes of Health Stroke Scale and was validated on 1,307 patients from the stroke data bank of the German Stroke Foundation. Three years later, Weimar et al [2006] used the same model in a simulation of various inclusion thresholds to identify optimum criteria for acute stroke trials. Rather than propose one specific recommendation, Weimar et al present a summary of sample size versus enrollment time trade-offs for different combinations of prognostic thresholds. Saver and Yafeh used a sliding dichotomy approach (called 'baseline severity-adjusted end points') to confirm the affect of tPA on outcome for stroke patients in the National Institute of Neurological Disorders and Stroke tissue plasminogen activator trial [2007]. Hukkelhoven et al [2005] developed and validated a predictive model

for functional status (as defined by Glasgow Outcome Scale (GOS)) among patients suffering a TBI. Their model included age, motor score, pupil reaction, hypotension, hypoxia, computed tomography (CT) classification, and traumatic subarachnoid haemorrhage and was assessed in terms of discrimination and calibration. Internal validity was checked using bootstrapping and external validity was carried out using data from the European Brain Injury Consortium and the Traumatic Coma Data Bank. Hukkelhoven et al also compared the performance of their model to three other prognostic models from the literature [Choi et al., 1991, Signorini et al., 1999, Andrews et al., 2002] and their model appeared to have better discriminative ability. King, Carlier, and Marion [2005] used GOS at 3 months, hypotension, diffuse axonal injury, and pupil response to predict 12 month GOS among TBI patients. Hernandez et al [2004] developed a predictive model for GOS at six months for TBI patients including age, motor score, pupil reaction, CT classification, traumatic subarachnoid hemorrhage, hypoxia, hypotension, glycemia, and hemoglobin. However, their model has not been validated and they do not propose it as a method of developing prognostic bands for the sliding dichotomy approach. Instead, they simply propose measuring and adjusting for these baseline covariates in future studies as a way to increase power/reduce sample size. One year later, Murray et al [2007] essentially confirmed their predictive model results (strongest predictors were age, GCS motor score, pupil response, CT classification, and traumatic subarachnoid hemorrhage), this time with an eye toward applying the sliding dichotomy method. Additional analyses that year by Marmarou et al. [2007] and McHugh et al. [2007] looked more specifically at the predictive value of pupil reactivity and secondary insults such as hypoxia, hypotension, and hypothermia.

Despite the presence of many proposed predictive models for both TBI and stroke studies, it was not until 2005 that Murray et al. fully formalized the sliding dichotomy method, and specified two crucial questions - "... is it accepted that the concept of relating outcome for a given patient to that which would be expected, given the baseline prognosis, does give a clinically relevant estimate of treatment effect?" "... how many prognostic bands should be used and how should they be defined?" It is this latter question that is proving most challenging in both the stroke and TBI research communities. Counsell and Dennis

[2001] conducted a systematic review of predictive models for patients with acute stroke and concluded that "[n]one of the existing prognostic models have been sufficiently well developed and validated to be useful in either clinical practice or research." Perel et al [2006] conducted a systematic review of existing predictive models in the TBI literature covering 53 reports and 102 models and concluded that "...68% did not justify the rationale to include the predictors, 11% conducted an external validation and only 19% of the logistic models presented the results in a clinically user-friendly way." Overall, GCS, age, and pupil reactivity were the most commonly used predictors. Perel et al considered Hukkelhoven's [2005] model to be among the most clinically useful and methodologically sound, so his will be one of the predictive models compared to our prognostic score approach in chapter five.

Chapter 3

Confounding in Observational Studies: Comparing Propensity Score and Traditional Regression Analyses

3.1 Background

In non-randomized studies, groups may not be comparable due to systematic differences in the distribution of covariates unrelated to the treatment or exposure of interest. Traditionally, such confounding is addressed in an epidemiologic model by including these covariates in the final outcome regression model or by matching members of comparison groups of interest according to the sets of unbalanced covariates. Unfortunately, both solutions may be problematic under many real world circumstances (see section 1.3), and so the propensity score method was introduced by Rosenbaum and Rubin [1983a] as a potential alternative method for adjusting for confounding in observational studies.

The propensity score is the probability of treatment assignment (or exposure) given a set

of covariates, and is defined as (1.6) from section 1.3,

$$e(x_i) = P(Z_i = 1|X_i = x_i)$$

where Z represents exposure (or treatment assignment) and X represents a vector of covariates. Once the propensity score has been estimated, the probability that an individual observation received treatment can be used to adjust the final outcome model in a variety of ways (see 1.3.3).

Although Rosenbaum and Rubin introduced the propensity score more than two decades ago, and since then it has been used in a wide variety of fields [Imai and van Dyk, 2004, Grunkemeier et al., 2002], no consensus currently exists as to whether and how results based on propensity score analyses differ from those based on traditional regression adjustment.

Austin et al [2007] emphasize the importance of clearly specifying the performance of the two methods: "... several applied studies have estimated treatment effects using both propensity score and regression methods. Assuming that one of the analytic methods was added as a test of robustness of the findings of the initial analytic approach, this indicates that many applied researchers incorrectly believe propensity score methods and regression methods to be estimating the same treatment effect." More recently, Shadish et al [2008] point out that "... the practice of propensity score analysis in applied research may be yielding adjustments of unknown or highly variable accuracy. For a method as new as propensity score analysis, this is not surprising, and points to the need for more clarity about best propensity score practice." In particular, Drake's [1993] findings that the bias from propensity score analyses "... declines with increasing treatment effect and increases with increasing covariate effects on treatment as well as on the response" motivate a more thorough analysis of propensity score performance when covariate characteristics are varied.

Unfortunately, covariate characteristics are rarely varied in studies of the performance of propensity scores. As described briefly above and in more detail in section 2.1, recent propensity score research has focused on the parametric form of the propensity score and/or outcome regression models, different propensity score adjustment methods, and the poten-

tial sensitivity of such analyses to unmeasured confounders. The vast majority of these analyses assume continuous, normally distributed confounders and, with the exception of Drake [1993], do not assess the sensitivity of such analyses to varying levels of association between confounders and treatment assignment. In contrast, we are interested in the effect of the distribution of potential confounders on the estimated effect size from propensity score and traditional analyses, specifically when such confounders are dichotomous. By conducting two different types of simulations, we examined the performance of both traditional and propensity score regression adjustment methods when the distributions of potential confounders were varied to produce more and less similar comparison groups of interest (and thus stronger and weaker associations between confounders and treatment indicator). The first is referred to as a 'pseudo-simulation' in that existing data were repeatedly resampled to generate new samples with known covariate distributions of interest but an unknown treatment effect size. The second is referred to as a 'full simulation' in that all variables were generated from chosen distributions and datasets were created with a known treatment effect size. We hope that by examining the sensitivity of estimates of treatment effect size to varying distributions of confounders and specifically to dichotomous confounders this work further contributes to 'best propensity score practice.'

We are particularly interested in how these two methods perform when the confounders of interest are dichotomous variables, since this represents both a gap in the existing literature and a category of confounders frequently found in public health research. The majority of the theoretical results in this area assume normally distributed, or at least continuous, covariates, thus simplifying tests of similar variance structures (a key part of balance-checking techniques for propensity score analyses; see section 1.3.2). Rosenbaum and Rubin state [1983a] that traditional regression adjustment and propensity score regression adjustment should result in the same point estimate for treatment effect if the discriminant is a monotone function of the propensity score (i.e., if the covariance matrices in the treated and control groups are equal). Since this assumption is likely to be violated, to some degree, in observational studies, we are interested in determining if one method consistently produces a less biased estimate when the covariance matrices differ. Since the variance of dichoto-

mous variables depends upon the frequency of those covariates (i.e., variance = npq) we are particularly interested in how the two methods perform as the frequencies of dichotomous variables change, since this is inevitably also changing the variance structure of the two groups, which Rubin [1979] has clearly shown to increase the risk of adding bias to the final treatment effect estimate rather than reducing bias.

Lastly, most evaluations of the performance of propensity score methods report the percent of bias reduction, which is of course an important feature of the method. However, for analytical users of the method, bias reduction resulting from varying parameter estimates does not necessarily directly translate to the estimated effect size or the distribution of covariates. At the moment, data analysts are left to determine if their data resembles that used or generated in a given paper reporting bias results, but without any clear guidelines. We hope to develop a clear set of guidelines regarding the performance of traditional versus propensity score methods with regard to the distributions of dichotomous covariates and the estimated effect size. By focusing on the estimated effect size we are returning to important points made by Rubin [2001] and D'Agostino [1998] regarding the use of propensity scores in designing observational studies, an area of application that appears to be mostly neglected in the literature. If propensity scores do indeed produce more accurate estimates of effect size, this can help in the design of future studies by reducing the required sample size, increasing power, and generally reducing the cost of conducting a study.

3.2 Bias

As mentioned in section 1.3.3, bias reduction can be measured in terms of the remaining difference in x after adjustment or bias in the estimated treatment effect $\hat{\tau}$ (if true treatment effect τ is known).

It has been well established that both traditional regression adjustment methods and propensity score regression adjustment methods will produce unbiased estimates of treatment effect under a certain set of assumptions. Cochran [1973], Rosenbaum [1983a, 1984], and Rubin [1973, 1979, 1983a, 1984] all contributed to quantifying the bias that may remain

(and the possibility of increased bias) in an estimate of treatment effect from traditional regression adjustment when these assumptions are minorly violated. Rubin showed [1979] that the treatment effect estimate

$$\widehat{\tau} = (\overline{y}_t - \overline{y}_c) - \widehat{\beta}(\overline{x}_t - \overline{x}_c)$$

where $\widehat{\beta}$ is the traditional ordinary least squares estimate, has conditional bias given x_{ij}

$$\overline{w}_{t.} - \overline{w}_{c.} - (\overline{x}_{t.} - \overline{w}_{c.})\frac{S_{xw}}{S_{xx}}$$

where $w_{ij} = W(x_{ij})$ from $E(Y|x) = \alpha + W(x)$. In large samples, $\widehat{\tau} \approx E_t(W(X)) - E_c(W(X)) - \eta c$ where $E_z()$ is expectation taken across the distribution of x in treatment group z, and c is the pooled slope of $W(X)$ on the discriminant. Rubin defines X in the treated population as

$$\mathbf{X} \sim N\left(\begin{pmatrix}\eta\\0\end{pmatrix}, \begin{bmatrix}\sigma^2 & 0\\0 & \xi^2\end{bmatrix}\right)$$

and in the control

$$\mathbf{X} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right).$$

He then defines

$$W(\mathbf{X}) = W(u,v) = exp\Big[a\Big(\frac{2}{1+\sigma^2}\Big)^{1/2}(u - \eta/2) + b\Big(\frac{2}{1+\xi^2}\Big)^{1/2}v\Big]$$

where a and b are allowed to vary to generate moderately nonlinear response surfaces. Combining the above, Rubin writes

$$\eta c = aB[E_t(W(X))(\sigma^2/(1+\sigma^2)) + E_c(W(X))/(1+\sigma^2)]$$

where $B = \eta/\big(\frac{1+\sigma^2}{2}\big)^{1/2}$, showing that for $\sigma_t^2 \neq 1$ and/or $\xi^2 \neq 1$ combined with negative values of a, bias in the estimated treatment effect $\hat{\tau}$ may actually be increased.

As far as we are aware, there is no mathematical quantification of the bias remaining after

propensity score regression adjustment, though clearly it should be similar to that found in traditional regression but substituting a different function of x and a different parameter estimate.

Alternatively, bias can also be measured in terms of the remaining imbalance in x. After developing their propensity score, Rosenbaum and Rubin [1983a] showed that the initial bias in X is $B = E(X|Z = t) - E(x|Z = c)$ and that after adjusting for stratification based on propensity score subclass the remaining bias is:

$$B_s = \sum_{s=1}^{S} w_s \int E(X|b)[P(b|Z = t, b\epsilon I_s) - P(b|Z = c, b\epsilon I_s)]db$$

where b is some balancing score $b(x)$, subclasses I_s defined by the balancing score, and w_s is the direct adjustment weight for each subclass.

3.3 Variance Structure

As mentioned in previous sections, we are particularly interested in the performance of traditional and propensity score regression adjustment methods under varying distributions of dichotomous covariates since the difference in frequency of dichotomous covariates in treatment and control groups directly affects the variance structure of these covariates. Rubin typically limits his analyses of both traditional and propensity score methods to cases where the ratio σ_t^2/σ_c^2 is between 0.5 and 2. Love, in his 'Strategies for Using Propensity Methods Well' [2005] echoes this rule of thumb for both the variance of the logit of propensity scores themselves as well as the variance of the individual covariates. Tangential to our research into the estimated treatment effect size from these two methods, we were interested in assessing whether this rule of thumb also applies well to dichotomous variables. One way to formalize this question is to re-write the variance ratio above in terms of variance for dichotomous variables:

$$0.5 < \frac{n_t p_t(1 - p_t)}{n_c p_c(1 - p_c)} < 2$$

where n_z is the sample size for treatment group z and p_z is the probability of some dichotomous covariate outcome for treatment group z. For the simplified case where $n_t = n_c$ we can then write the question as for what value of a does the following hold true

$$0.5 < \frac{p_t(1-p_t)}{ap_t(1-ap_t)} < 2$$

where the probability in the control group is written in terms of the probability in the treated group ($p_c = ap_t$), and do these bounds on the ratio of variances result in 'good' implementations of propensity score methods with dichotomous variables? For example, in terms of the magnitude of the imbalance between comparison groups, a frequency of 5% in one group and 15% in another is obviously less imbalanced than 5% in one and 90% in the other, yet the former technically violates the variance ratio rule of thumb while the latter does not! (see figure 3.1) Of course, this is a simplified version of the question, and bounds would also need to be considered for a variety of values of r for $n_t = rn_c$.

	0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
0.05	1.00	0.53	0.37	0.30	0.25	0.23	0.21	0.20	0.19	0.19	0.19	0.20	0.21	0.23	0.25	0.30	0.37	0.53	1.00
0.10	1.89	1.00	0.71	0.56	0.48	0.43	0.40	0.38	0.36	0.36	0.36	0.38	0.40	0.43	0.48	0.56	0.71	1.00	1.89
0.15	2.68	1.42	1.00	0.80	0.68	0.61	0.56	0.53	0.52	0.51	0.52	0.53	0.56	0.61	0.68	0.80	1.00	1.42	2.68
0.20	3.37	1.78	1.25	1.00	0.85	0.76	0.70	0.67	0.65	0.64	0.65	0.67	0.70	0.76	0.85	1.00	1.25	1.78	3.37
0.25	3.95	2.08	1.47	1.17	1.00	0.89	0.82	0.78	0.76	0.75	0.76	0.78	0.82	0.89	1.00	1.17	1.47	2.08	3.95
0.30	4.42	2.33	1.65	1.31	1.12	1.00	0.92	0.88	0.85	0.84	0.85	0.88	0.92	1.00	1.12	1.31	1.65	2.33	4.42
0.35	4.79	2.53	1.78	1.42	1.21	1.08	1.00	0.95	0.92	0.91	0.92	0.95	1.00	1.08	1.21	1.42	1.78	2.53	4.79
0.40	5.05	2.67	1.88	1.50	1.28	1.14	1.05	1.00	0.97	0.96	0.97	1.00	1.05	1.14	1.28	1.50	1.88	2.67	5.05
0.45	5.21	2.75	1.94	1.55	1.32	1.18	1.09	1.03	1.00	0.99	1.00	1.03	1.09	1.18	1.32	1.55	1.94	2.75	5.21
0.50	5.26	2.78	1.96	1.56	1.33	1.19	1.10	1.04	1.01	1.00	1.01	1.04	1.10	1.19	1.33	1.56	1.96	2.78	5.26
0.55	5.21	2.75	1.94	1.55	1.32	1.18	1.09	1.03	1.00	0.99	1.00	1.03	1.09	1.18	1.32	1.55	1.94	2.75	5.21
0.60	5.05	2.67	1.88	1.50	1.28	1.14	1.05	1.00	0.97	0.96	0.97	1.00	1.05	1.14	1.28	1.50	1.88	2.67	5.05
0.65	4.79	2.53	1.78	1.42	1.21	1.08	1.00	0.95	0.92	0.91	0.92	0.95	1.00	1.08	1.21	1.42	1.78	2.53	4.79
0.70	4.42	2.33	1.65	1.31	1.12	1.00	0.92	0.88	0.85	0.84	0.85	0.88	0.92	1.00	1.12	1.31	1.65	2.33	4.42
0.75	3.95	2.08	1.47	1.17	1.00	0.89	0.82	0.78	0.76	0.75	0.76	0.78	0.82	0.89	1.00	1.17	1.47	2.08	3.95
0.80	3.37	1.78	1.25	1.00	0.85	0.76	0.70	0.67	0.65	0.64	0.65	0.67	0.70	0.76	0.85	1.00	1.25	1.78	3.37
0.85	2.68	1.42	1.00	0.80	0.68	0.61	0.56	0.53	0.52	0.51	0.52	0.53	0.56	0.61	0.68	0.80	1.00	1.42	2.68
0.90	1.89	1.00	0.71	0.56	0.48	0.43	0.40	0.37	0.36	0.36	0.36	0.37	0.40	0.43	0.48	0.56	0.71	1.00	1.89
0.95	1.00	0.53	0.37	0.30	0.25	0.23	0.21	0.20	0.19	0.19	0.19	0.20	0.21	0.23	0.25	0.30	0.37	0.53	1.00

Figure 3.1: Ratio of Variances for Dichotomous Variables, Highlighted Where $0.5 < \frac{p_t(1-p_t)}{ap_t(1-ap_t)} < 2$ violated; rows and columns correspond to potential values of p_t and $p_c = ap_t$

3.4 Motivating Example

This problem was originally motivated by analysis of data from the Coverdell Stroke Registry [Reeves et al., 2007], which was used as the basis for initial simulations. Wave I data

were used, which included four states (MI, MA, OH, and GA) treated as strata and multiple hospitals within each state treated as clusters. Although this complicated data structure does present additional analytical challenges, it also motivates the application of propensity score methods - the degrees of freedom for any analysis of complex survey data are determined by the number of primary sampling units (in our case, clusters), not the total number of individuals in the dataset. Although we are fortunate that this specific dataset contains many clusters, the number of covariates that it is reasonable to include in any analysis of complex survey data can become limited rather quickly.

The Coverdell Stroke Registry includes four stroke types (Ischemic, Transient Ischemic Attack, Intra-cranial Hemorrhage, and Sub-arachnoid Hemorrhage). Approximately 63% of patients had an Ischemic stroke, and this subpopulation forms the basis of our analyses. The population is predominantly white (75%) and we are interested in the causal effect of race on length of hospital stay following a stroke.

Some causal inference researchers object to the use of race, gender, and other non-manipulable attributes as factors of interest, based primarily on Paul Holland's argument that "each unit be *potentially exposable* to any one of the causes." [Holland, 1986] We argue that as studies are beginning to manipulate an individual's *perception* of race [Grogger and Ridgeway, 2006] and gender [Goldin and Rouse, 2000] this provides a conceptual framework for the estimation of the causal effect of these attributes. Another similar argument is that possible covariates for the propensity score model are typically limited to pre-treatment (or pre-exposure) variables, and technically there are no variables that were measured before race was determined. However, continuing with the idea of the perception of race, it is possible that other demographic and medical history variables could be noted on a medical chart while omitting race, therefore there is a pool of variables that could be known to a doctor prior to his/her perception of a patient's race. In this way, we fit Rubin's requirement of well-defined units, treatments, and outcomes (see section 1.2.1 - SUTVA). Throughout this chapter we hope the reader will tolerate a minor abuse of notation as we refer to race as the treatment Z of interest.

Of course, we also have to consider SUTVA, and in this case the clustering within hospitals

is probably a minor violation in the sense that there could be 'interference between units.' [Rubin, 1986] Since our 'treatment' or 'cause' of interest is race, it is plausible that we would expect to see a higher frequency of the same race clustered within any one hospital, and of course both patient characteristics and potential outcomes may be linked to hospitals. Given the large number of clusters in our dataset, we do not believe that this minor violation will strongly affect our analyses, however looking for ways to take this clustering effect into account should be considered for future analyses.

3.4.1 Methods - Pseudo-simulation

As mentioned in section 3.1, this first round of simulations is based on a re-sampling of existing data to create datasets with known covariate distributions of interest but unknown treatment effect sizes. Based on the pool of variables described above, propensity scores were calculated using logistic regression with race as the outcome. Possible covariates considered for this model included basic demographic and medical history information (age, residence, mode of arrival to emergency department (ED), diabetes, etc.), several of which were significantly associated with race (i.e., imbalanced between the two groups), see table 3.2. Covariates chosen for this model (using a combination of stepwise model selection and checking for covariate balance post-propensity score adjustment) were hypertension, hyperlipidemia, coronary artery disease (CAD), and afibrillation; these were used to calculate a propensity score for each observation.

Interestingly, the originally selected 'best model' based solely on stepwise model selection (which included age, HDL, hypertension, hyperlipidemia, triglycerides, diabetes, and LDL) left a non-trivial amount of imbalance between black and white patients in eight of the originally 11 imbalanced covariates. This emphasizes the need to evaluate propensity score model results on their balancing performance in addition to simply the overall quality of the model itself.

Balance was assessed in several ways - the binary race indicator was regressed on each covariate both before and after adjusting for the calculated propensity scores, and the

Wald F test values were compared (see Table 3.1). The standardized differences were calculated for each covariate before and after propensity score adjustment, the ratio of the variances of the propensity scores for black and white patients was calculated, and the ratio of the variances of the residuals from regressing each covariate on race after adjusting for propensity score was compared between black and white patients. Lastly, the frequency of the dichotomous covariates within each propensity score quintile was summarized using bar graphs and the overlap of the propensity score values among black and white patients was checked (see Figure 3.2). This is a critical step since a small amount of remaining imbalance in a variable that is strongly related to the outcome of interest can result in large bias in the final causal estimate [Ho et al., 2007].

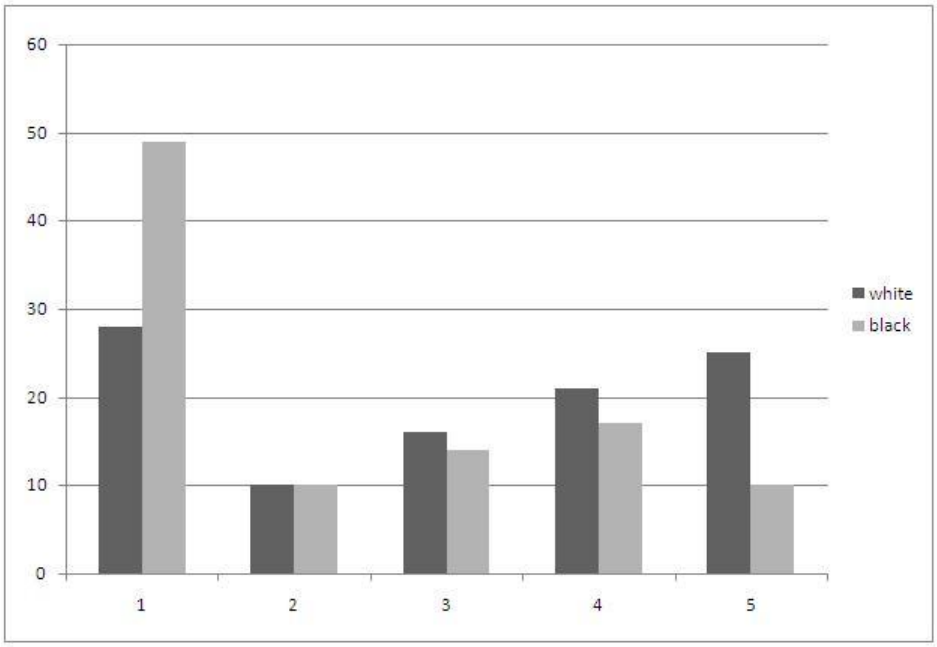

Figure 3.2: Distribution of black and white patients among propensity score quintiles

As with any model selection, our 'final' propensity model was not perfect. Of the 11 covariates originally significantly associated with race, adjusting for our propensity score based on hypertension, hyperlipidemia, CAD, and afibrillation left four covariates still unbalanced (see Table 3.1). However, based on the balance-checking methods described above, and the clinical relevance of the balanced versus unbalanced covariates, we decided that this model was our best choice. Additionally, two of the remaining significant covariates did indicate a

decrease in Wald F values, just not enough of a decrease to no longer be significantly associated with race. Last, the four remaining significantly associated covariates, diabetes, age, HDL, and triglycerides, all showed no association with length of stay in our initial univariate analyses. Therefore, we believe that despite the remaining imbalances the possibility of additional bias is low.

Table 3.1: Covariate Balance

	Pre-Adjustment Wald F	Post-Adjustment Wald F
Hypertension	11.8	1.9
Hyperlipidemia	8	0.8
CAD	26.8	0.3
Afibrillation	9.8	0.5
Diabetes	**8.6**	**12.3**
Myocardial Infarction	5.4	0.02
Age	**49.4**	**57.4**
HDL	**23.4**	**12.6**
LDL	2.2	2.4
Total cholesterol	3.2	3.2
Triglycerides	**12.7**	**10.3**

Using the covariates identified by propensity score analyses, we built a traditional linear regression model, with length of stay as the outcome, race as the primary covariate of interest, and controlling for hypertension, hyperlipidemia, CAD, and afibrillation. Stepwise model building indicated other significant variables - dysphagia screening, diagnosis of stroke by the emergency department team, diabetes medication, anti-thrombotics prescribed at discharge, stroke team present in the emergency department, prior stroke, and mode of arrival to emergency department.

Therefore two linear regression equations form the basis of our comparisons - the traditional model as described above, and a second model including most of the same covariates, except replacing hypertension, hyperlipidemia, CAD, and afibrillation with a single covariate representing quintiles of propensity score.

In essence, what we are interested in determining is if the estimated treatment effect $\widehat{\beta}_2$

from this model:

$$\begin{aligned}\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1(\widehat{e(x)}) + \widehat{\beta}_2\, race + \\ \widehat{\beta}_3\, dysphagia + \widehat{\beta}_4\, diagnosis + \widehat{\beta}_5\, diabmed + \widehat{\beta}_6\, antithrom + \\ \widehat{\beta}_7\, stroketeam + \widehat{\beta}_8\, prior\ stroke + \widehat{\beta}_9\, arrival\ mode\end{aligned} \tag{3.1}$$

(where $e(x) = logit(P(race = 1)) = \alpha_0 + \alpha_1 htn + \alpha_2 hlip + \alpha_3 CAD + \alpha_4 afib$) differs in a statistically significant way from the estimated treatment effect $\widehat{\beta}_5$ from this model:

$$\begin{aligned}\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1\, htn + \widehat{\beta}_2\, hlip + \widehat{\beta}_3\, CAD + \widehat{\beta}_4\, afib + \widehat{\beta}_5\, race + \\ \widehat{\beta}_6\, dysphagia + \widehat{\beta}_7\, diagnosis + \widehat{\beta}_8\, diabmed + \widehat{\beta}_9\, antithrom + \\ \widehat{\beta_{10}}\, stroketeam + \widehat{\beta_{11}}\, prior\ stroke + \widehat{\beta_{12}}\, arrival\ mode\end{aligned} \tag{3.2}$$

Another way to look at this is through the equation for the estimated treatment effect itself. Following our slightly modified version of Rubin's notation [1979] from section 3.2, the general form of the estimated treatment effect under traditional regression adjustment (which is unbiased if response surfaces are parallel) is:

$$\widehat{\tau} = (\overline{Y}_t - \overline{Y}_c) - \widehat{\beta}_{TR}(\overline{X}_t - \overline{X}_c) \tag{3.3}$$

and we wish to determine if that differs significantly from

$$\widehat{\tau} = (\overline{Y}_t - \overline{Y}_c) - \widehat{\beta}_{PS}(\overline{e}(X_t) - \overline{e}(X_c)) \tag{3.4}$$

Rosenbaum and Rubin show in their proof of Corollary 4.3 [1983a] that under large sample theory, 3.4 is unbiased when $E\{Y_t|Z = t, e(x)\} = E\{Y_t|e(x)\}$ (again with slightly modified notation).

Starting from the distributions of hypertension, hyperlipidemia, CAD, and afibrillation among black and white patients in our original dataset, we systematically increased and decreased the proportion of patients with each of these characteristics, to generate more

and less similar study populations in which to compare the results of our two models.

More specifically, we held the rate of each covariate steady within the black population (for ease of computation) and altered the rate in white patients from 5% to 95% by increments of 5%. Covariates were altered one at a time, independently of each other. Within each rate, 1,000 samples were generated, and new propensity scores were estimated for each sample, based on our previously selected 'best' model. Propensity score quintiles were used as a categorical variable in the linear regression model. The predicted marginal mean length of stay was used to estimate treatment effect size. First the predicted marginal mean length of stay was estimated for black and white patients from each model. Then the difference between black and white patients was calculated, and the difference from the propensity score model was subtracted from the difference from the traditional model. This 'difference of differences' was averaged across all 1,000 samples for each rate, generating a mean difference between methods for all 19 possible distributions for each of the four covariates. The 25^{th} and 975^{th} of the ordered differences were also recorded to form a 95% empirical probability interval for each difference.

A second set of analyses was also conducted, nearly identical to the first, except that instead of including the propensity score quintiles as a covariate in the linear regression model, five separate models were estimated, one for each quintile. A weighted average difference between black and white patients was calculated, based on the frequency distribution of the quintiles, and this weighted average was compared to the average estimated from the traditional analyses. Although the quintiles should have all had an equal distribution (20% of the sample in each category) due to the large variation across simulations, the propensity score did not always divide neatly in this manner (e.g., when very many or very few patients were recorded as having hypertension, some propensity score categories contained fewer than 20% of the observations).

3.4.2 Results

Prior to adjusting for any potential confounders, white patients spent an average of 5.3 days in the hospital (SE = 0.4) and black patients spent an average of 6.5 days in the hospital (SE = 0.4) following a stroke. This difference was statistically significant (p-value = 0.007).

In the original dataset, 72% of white patients had hypertension, 29% hyperlipidemia, 34% CAD, and 18% afibrillation. Among black patients 83% had hypertension, 21% hyperlipidemia, 19% CAD, and 10% afibrillation (Table 3.2).

Table 3.2: Potential Confounders

	Black	White	p-value
Hypertension	**83% (694/862)**	**72% (1762/2514)**	**0.0002**
Hyperlipidemia	**21% (169/862)**	**29% (706/2514)**	**0.008**
CAD	**19% (156/862)**	**34% (784/2514)**	**0.0001**
Afibrillation	**10% (80/849)**	**18% (433/2501)**	**0.003**
Diabetes	**40% (326/862)**	**30% (722/2514)**	**0.003**
Pre-Hospital GCS	58% (166/263)	64% (459/675)	0.4
Gender (male)	42% (383/860)	44% (1134/2513)	0.6
Resident (nursing home)	4% (35/852)	6% (153/2479)	0.2
Arrival mode (ambulance)	46% (358/792)	46% (1075/2348)	0.9
Prior stroke	36% (324/862)	36% (902/2514)	0.9
Smoker	24% (255/862)	24% (569/2514)	0.97
Myocardial Infarction	**10% (98/862)**	**14% (345/2514)**	**0.03**
Congestive Heart Failure	16% (124/862)	15% (355/2514)	0.5
Prosthetic Valve	1% (11/862)	2% (60/2514)	0.3
Age (SE)	**65 (1)**	**73 (1)**	**< 0.0001**
HDL (SE)	**49 (1)**	**44 (1)**	**< 0.0001**
LDL (SE)	**127 (4)**	**118 (5)**	**0.08**
Total cholesterol (SE)	**202 (4)**	**192 (4)**	**0.05**
Triglycerides (SE)	**132 (6)**	**178 (7)**	**< 0.0001**

Applying the two models to our original dataset, the difference in length of stay between black and white patients was estimated to be 1.53 days based on traditional linear regression, 1.54 days based on linear regression including the propensity score as a covariate, and 1.79 days when calculating a weighted average across individual linear regression analyses for each propensity score quintile. Stratifying by propensity score quintiles frequently resulted in a larger estimated effect size throughout simulations. This was primarily driven by a much larger estimated effect size in the highest quintile, and will be addressed further in

later sections.

As might be expected from the similar results in our original dataset, when propensity score was included as a covariate, few simulations indicated a statistically significant difference between traditional and propensity score analyses (i.e., 95% empirical probability intervals included 0) - the majority of significant differences were detected in the hypertension simulations (8 out of 19 rate comparisons produced significant results) and one significant difference was detected among the hyperlipidemia simulations. None of the simulations for the remaining two covariates (CAD and afibrillation) indicated a statistically significant difference between methods.

The differences within the hypertension simulations were consistently negative, indicating that the estimated effect size of race was consistently larger in the propensity score analyses (estimated difference between methods was one tenth of a day or less). Interestingly, significant differences were not detected among a single grouping of rates - i.e., it was not only when black and white patients were most or least similar along hypertension rates that propensity score methods differed from traditional. Instead, significant differences were detected when white patients were simulated to have hypertension rates of 5%, 10%, 50%, 55%, 60%, 75%, 80% and 85% (compared to the constant hypertension rate of 83% among black patients).

In contrast, when hypertension regression analyses were stratified by propensity score quintile, results were intuitive - significant differences were consistently detected when black and white patients were least similar - when white patients were simulated to have rates between 5% and 25%. Additionally, this represents a reduction in the number of significant differences between the two methods (eight when including propensity score as a covariate, five when stratifying by quintiles; see tables 3.3 and 3.4).

Interestingly, this reduction in the number of significant differences when analyzing hypertension is opposite from the effect found in the other three covariates. In general, more differences were observed when regression models were stratified by propensity score - the weighted average effect across propensity scores differed significantly from the tradition-

ally estimated effect size in approximately 30% of the simulations. These differences were observed in each of the four covariates and ranged from about half a day to a one day difference in estimated effect size between the two methods, with propensity score analyses consistently estimating a larger difference between black and white patients. The greatest number of significant differences were detected in the afibrillation simulations (10 out of 19; compared to zero in covariate adjustment analyses; see appendix for full simulation results) but these were distributed somewhat inconsistently. Black patients were held steady with a rate of 10% and significant differences between the methods were detected for rates within the white patient population of 30%, 35%, 50%, 55%, and 70%-95%.

For hyperlipidemia and CAD simulation results were intuitive - significant differences between the two methods were detected when black and white patients were most different. Black patients had a rate of hyperlipidemia of 21%, and significant differences between the methods were detected when white patients were simulated to have rates of 70% and above. Similarly, black patients had a rate of CAD of 19% and significant differences were detected when white patients were simulated to have rates of 70% to 90%. Interestingly, a significant difference was not detected when white patients were simulated to have a rate of 95%, however this empirical probability interval was the widest of the 19 simulations, so it is possible that the large difference between groups resulted in an unstable estimate.

3.4.3 Discussion

Typically, the success of a propensity score as a balancing score is assessed before proceeding with the next step in data analysis. Since implementing this step in every single simulation would have been time-prohibitive, several simulations were arbitrarily selected to assess balance between groups after controlling for propensity score. As described in section 3.4.1, we already know that ideal balance was not achieved for this dataset, so what we were checking here was if the already identified remaining imbalance was changing significantly between simulations. For this second round of balance checking we used two methods - first we estimated the association between race and each of the potential confounders after controlling for propensity score quintiles in logistic regression and second by estimating the

Table 3.3: Simulation results - Hypertension (83% among black patients) - Comparing traditional and propensity score regression adjustment

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	**-0.13**	**(-0.25, -0.03)**
10%	**-0.14**	**(-0.28, -0.02)**
15%	-0.1	(-0.24, 0.02)
20%	-0.05	(-0.17, 0.05)
25%	-0.06	(-0.18, 0.04)
30%	-0.1	(-0.22, 0.02)
35%	-0.03	(-0.17, 0.10)
40%	-0.02	(-0.14, 0.08)
45%	-0.03	(-0.12, 0.07)
50%	**-0.07**	**(-0.14, -0.01)**
55%	**-0.08**	**(-0.14, -0.03)**
60%	**-0.07**	**(-0.12, -0.03)**
65%	-0.04	(-0.09, 0.003)
70%	-0.03	(-0.11, 0.03)
75%	**-0.07**	**(-0.12, -0.04)**
80%	**-0.06**	**(-0.09, -0.03)**
85%	**-0.08**	**(-0.17, -0.03)**
90%	-0.05	(-0.12, 0.006)
95%	0.005	(-0.03, 0.05)

association between race and each of the potential confounders within each propensity score quintile. In some cases there was additional remaining imbalance as compared to our original dataset. Nevertheless, we consider the comparison of even these cases to be informative - a highly imbalanced sample, one that cannot be brought into balance by a propensity score, is also going to present complications to a traditional regression analysis. Furthermore, additional remaining imbalance was found in both simulations where a significant difference between the two methods was detected and simulations where a significant difference was not detected. Therefore, we do not believe that instances of failure in terms of balance scores affected the estimated difference between propensity score and traditional analysis methods. Lastly, as noted in section 1.3.2, the majority of balance-assessment rules of thumb are based on the assumption of normally distributed covariates, which was not the case in our dataset, so it is difficult to know precisely what cut-off values to use in concluding that a covariate is balanced or unbalanced. Throughout our original dataset and simulations, we do achieve adequate overlap of propensity score values between black and white patients, which further bolsters our belief that any remaining imbalance or added bias is minimally

Table 3.4: Simulation results - Hypertension (83% among black patients) - Comparing traditional regression adjustment to stratifying by propensity score

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	**-0.59**	**(-0.20, -0.60)**
10%	**-0.55**	**(-0.66, -0.16)**
15%	**-0.53**	**(-0.78, -0.07)**
20%	**-0.62**	**(-0.91, -0.17)**
25%	**-0.65**	**(-1.14, -0.06)**
30%	-0.38	(-1.29, 0.25)
35%	-0.38	(-0.82, 0.11)
40%	-0.41	(-0.85, 0.09)
45%	-0.51	(-1.10, 0.09)
50%	-0.4	(-1.08, 0.15)
55%	-0.32	(-0.93, 0.22)
60%	-0.06	(-0.76, 0.54)
65%	0.4	(-0.09, 0.81)
70%	-0.14	(-1.16, 0.81)
75%	-0.29	(-1.08, 0.43)
80%	-0.19	(-0.93, 0.45)
85%	-0.14	(-0.67, 0.52)
90%	-0.14	(-0.72, 0.38)
95%	0.33	(-0.50, 0.79)

affecting our conclusions regarding comparing traditional and propensity score regression adjustment methods.

The significant differences detected between methods in the initial simulations regarding hypertension warrant further examination. Within the rates where a significant difference was detected, the distribution of the propensity score differed substantially, leading us to believe that hypertension may be interacting with one (or more) of the other covariates. A necessary next step in this research would be to repeat simulations altering more than one covariate at a time, while accounting for covariance. This also makes good clinical sense, since a patient with hypertension may indeed also be more likely to have coronary artery disease or hyperlipidemia or afibrillation or some combination. Additionally, interaction terms should be considered between the potential confounders and race and alternative rate distributions should be considered, perhaps altering rates within both black and white populations.

Another potential modeling concern in our simulations is that after selecting the 'best'

propensity score and final regression models, these remained the same throughout simulations. It is feasible that as the distributions of covariates changed, were model building steps to be repeated, a different combination of covariates would be selected. However, since our primary interest here is in comparing propensity score and traditional regression methods, even if the performance of our models suffered across simulations, this should not have affected our overall comparison of the two methods, as we kept both models consistent for both methods throughout simulations.

The larger estimated effect sizes observed when stratifying by propensity score quintiles appear to be primarily driven by results in the last propensity score quintile. This warrants further investigation since even when simulating one of the potential confounders to have a low rate, the last quintile always contained the highest rates of the remaining potential confounders. Therefore, using propensity score analyses, the largest treatment effect was estimated when there was the highest incidence of potential confounding. Unfortunately, the question remains whether this indicates that propensity score analyses are correctly identifying a true treatment effect in the presence of increased noise (and traditional regression analyses controlling for potential confounders are underestimating the true treatment effect by mistaking it for noise) or propensity score analyses are giving too much weight to the larger estimated treatment effect in the last quintile. A common concern when this occurs is that lack of overlap between the two groups is to blame. Fortunately in our case, there are plenty of black and white patients in all five propensity score categories, so we do not believe lack of overlap to be contributing to a misleading estimate of treatment effect.

Lastly, in some ways this particular comparison is skewed slightly in favor of traditional analyses simply because of the limited number of covariates involved. One of the main advantages to propensity score techniques is that they are applicable in cases where the number of potential confounders makes traditional methods (either matching or regression adjustment) impossible.

3.4.4 Conclusion

Even when controlling for the same variables, propensity score analyses do produce significantly different results from traditional analyses, most frequently when treatment/exposure groups differ the most in terms of distribution of potential confounders. Although the final decision in terms of whether an effect (due to treatment or exposure) exists at all may be the same, the estimated size of that effect does differ between the two methods. More specifically, propensity score methods consistently estimate a larger effect size, at least within this example, and stratifying by propensity quintiles seems to produce a larger difference between the methods than simply controlling for quintiles in the regression equation. A critical next step will be to simulate data with a known effect size, to determine in cases where the two methods disagree, which is more closely approximating the correct effect size (see full simulation study in the next section).

3.5 Full Simulation Study

3.5.1 Introduction

As noted several times in the previous sections, model selection is an imperfect and subjective technique. There are numerous statistically defensible alternative models that could have been chosen at two different stages in our analyses (both the propensity score model with race as the outcome of interest and the final linear regression model with length of stay as the outcome of interest). None of the potential propensity score models achieved perfect balance between black and white patients. However, we believe we are justified in selecting our imperfect model since a) this is reflective of the kinds of decisions and trade-offs that must be made in real-world data analysis and b) based on the four balance-assessment techniques described in section 3.4.1 we are convinced that our imperfect model reduced more bias than it contributed. However, we did conduct identical analyses for alternative propensity score models and occasionally produced conflicting results to the ones presented here. Therefore, we believe that although this real world dataset has been useful in grounding our

comparison of two commonly used methods, it is now time to move to entirely simulated data where we can control the true treatment effect. This second round of simulations is therefore referred to as a full simulation; all variables are created from specifically chosen random distributions and treatment effect sizes are known.

These simulations compare results from propensity score regression adjustment and traditional regression adjustment methods under combinations of 'correct' propensity score and outcome models and 'incorrect' propensity score and outcome models, where treatment effect is known. In this case 'correct' models include all known confounders and 'incorrect' models omit one known confounder. Previous comparisons have focused on the correct parametric structure of the two models [Drake, 1993] but we are primarily interested in the performance of the two methods as the level of imbalance caused by confounders is increased or decreased. We include 'correct' and 'incorrect' versions of both models to further determine if one method is more sensitive to missing a potential confounder, as the imbalance caused by that confounder increases or decreases. Our goal is to use these simulations to better describe the performance of both propensity score and traditional regression adjustment methods in the presence of varying degrees of confounding, specifically when the confounders are dichotomous variables.

3.5.2 Methods

These simulations are still based loosely on the motivating example from the Coverdell Stroke Registry [Reeves et al., 2007]. First, covariates of interest were simulated at rates similar to those observed in the Coverdell dataset:

$$hypertension(htn) \sim Bin(n, 0.71)$$
$$hyperlipidemia(hlip) \sim Bin(n, 0.25)$$
$$age \sim N(69, 200).$$

The problem for this round of simulations was simplified by reducing the number of potential confounders. Additionally, although we are not interested in age as a potential confounder, it is included in the treatment assignment model below, with a very small coefficient, to generate a continuous propensity score. Essentially, this has the same effect as adding a small amount of normally distributed error on the end of the treatment assignment model. For conceptualization, we just considered this error to be the age of the study participants (throughout simulations it was confirmed that there was no significant difference in mean age between black and white patients). Alternatively, one could imagine this normally distributed 'error' as the remaining covariates in a more complicated propensity score model, where only two potential confounders are 'problematic' in terms of whether or not they should be included in the model. This was necessary since constructing a propensity score for only two dichotomous covariates results in a categorical propensity score, and adjusting for such a categorical propensity score would be equivalent to matching on values of those two dichotomous covariates, whereas we are interested in the performance of covariate adjustment via regression. Therefore, for simplicity, we are only focusing on the distribution and inclusion of two potential confounders, but are hopeful that our results could be generalized to models with more covariates.

The treatment assignment (in this case, race) was modeled:

$$logit(P(race = 1)) = \alpha_0 + \alpha_1 htn + \alpha_2 hlip + \alpha_3 age$$

which is also the 'true' propensity score model. The actual race indicator was determined by

$$race \sim Bin\Big(n, \frac{exp(\alpha_0 + \alpha_1 htn + \alpha_2 hlip + \alpha_3 age)}{1 + exp(\alpha_0 + \alpha_1 htn + \alpha_2 hlip + \alpha_3 age)}\Big). \tag{3.5}$$

As mentioned in section 3.4, it is somewhat inaccurate/controversial terminology within causal inference methods to refer to race as a treatment. However, for consistency and brevity of terminology we hope the reader will tolerate this minor abuse.

The primary outcome of interest (length of stay) was simulated as $Y \sim N(\mu, \sigma^2)$ where two μs were generated, one with and one without a true treatment effect

$$\mu_{trt} = \beta_0 + \beta_1 htn + \beta_2 hlip + \beta_3 race$$
$$\mu_{notrt} = \beta_0 + \beta_1 htn + \beta_2 hlip.$$

These also represent the 'correct' outcome models - $E(Y|X) = \beta_0 + \beta_1 htn + \beta_2 hlip + \beta_3 race$ and $E(Y|X) = \beta_0 + \beta_1 htn + \beta_2 hlip$. Values of α were varied to achieve differing levels of imbalance in terms of confounders between black and white patients and values of β were varied to achieve differing levels of association between those confounders and length of stay. The former is the primary problem of interest in this chapter whereas the latter allows us to test Drake's [1993] claim that propensity score bias decreases as treatment effect size and covariate effect on treatment as well as response increase. For the initial set of results presented in the following section, values of β were held constant as α values were varied. Subsequently, individual β values were increased while holding α values constant.

For each set of simulations, datasets with 1,000 observations were generated according to the above variable values. Once the simulated datasets were generated, four models were compared - the two 'correct' models

$$logit(P(race = 1)) = \alpha_0 + \alpha_1 htn + \alpha_2 hlip + \alpha_3 age \tag{3.6}$$
$$E(Y|X) = \beta_0 + \beta_1 htn + \beta_2 hlip + \beta_3 race \tag{3.7}$$

and two 'incorrect' models, either

$$logit(P(race = 1)) = \alpha_0 + \alpha_1 htn + \alpha_3 age \tag{3.8}$$

$$E(Y|X) = \beta_0 + \beta_1 htn + \beta_3 race \tag{3.9}$$

or

$$logit(P(race = 1)) = \alpha_0 + \alpha_1 hlip + \alpha_3 age \tag{3.10}$$

$$E(Y|X) = \beta_0 + \beta_1 hlip + \beta_3 race \tag{3.11}$$

each used twice, once with length of stay (Y) simulated to have a true treatment effect and once without. Hyperlipidemia was somewhat arbitrarily chosen as the potential confounder to be dropped in the initial 'incorrect' model, so analyses were repeated keeping hyperlipidemia and dropping hypertension in the 'incorrect' models to confirm our findings. This will be further addressed in later sections.

Performance of the two methods was determined two different ways. Traditionally, simulations estimate some parameter of interest $\hat{\theta}$, how far that estimate is from the true value of that parameter θ, and the coverage probability of that estimate, i.e., how often a 95% confidence interval around the point estimate $\hat{\theta}$ includes the true value θ. For this problem the parameter of interest is β_3, the regression coefficient for race in the final outcome model. Estimates of $\hat{\beta}_3$ along with coverage probabilities from propensity score and traditional regression methods are presented in the following section. Variance estimates for $\hat{\beta}_3$ from the propensity score method are based on bootstrap samples, since the estimate of the standard error of $\hat{\beta}_3$ from each propensity score model is slightly underestimated by ignoring the additional variability caused by estimating the propensity score itself.

Additionally, we estimated the treatment effect size based on the difference in marginal mean predicted outcome values for black and white patients from both methods. This

estimate of treatment effect was averaged across all 1,000 simulations and the estimated effects were sorted with the 25^{th} and 975^{th} forming a 95% empirical probability interval for the estimated treatment effect generated by each model.

The difference of these marginal means provides a comparable point estimate $\hat{\theta}$ to the estimate of $\hat{\beta}_3$ used above. Marginal means for each comparison group of interest are calculated by multiplying L, a column vector containing the appropriate contrast values, and b, the vector of parameter estimates $\hat{\beta}$, and the difference in these marginal means is simply $\hat{\beta}_3$. The advantage to this second measure is that it provides an estimate of the actual marginal means of the comparison groups of interest, not just the difference between them (i.e., the treatment effect size may be 2, but it may be clinically relevant whether that difference arises because the adjusted mean in one group is 4 and the other is 6 or the adjusted means are 200 and 202). Individual marginal means are also more closely related to our original conceptualization of the causal inference problem as formalized in section 1.2. The marginal mean length of stay for black and white patients can be written as $\mu_c = E(Y_c)$ and $\mu_t = E(Y_t)$ respectively with the average causal effect of interest

$$\mu_t - \mu_c = E(Y_t) - E(Y_c).$$

As shown in section 1.3 if we have successfully estimated the propensity score, the outcome Y is conditionally independent of treatment assignment Z given propensity score $e(x)$, making it possible to estimate the desired average causal effect:

$$E\{Y_t|e(x), Z = t\} - E\{Y_c|e(x), Z = c\} = E\{Y_t - Y_c|e(x)\}$$

This second measure was included both for its clinical relevance and to provide comparisons to the pseudo-simulation in the previous section. Since that analysis was based on resampling a real dataset (rather than simulating a new dataset with known characteristics) there is no known β_3 to which to compare estimated $\hat{\beta}_3$ values. Although similar analyses could have been conducted estimating the difference in $\hat{\beta}_3$ values between the two methods the marginal mean predicted values were considered more readily clinically interpretable.

Again, although point estimates of $\hat{\beta}_3$ and the difference in marginal mean values are comparable, empirical probability intervals for the latter estimate of treatment effect are slightly different. Overall results based on this estimation of treatment effect are consistent with those based on estimates of $\hat{\beta}_3$ so tables in the following section summarize $\hat{\beta}_3$ and tables summarizing marginal mean estimates can be found in the appendix. The slight difference in empirical probability intervals will be addressed in the results and discussion sections.

3.5.3 Results

Overall, all four models ('correct' and 'incorrect' traditional and propensity score) produced very similar estimates of $\hat{\beta}_3$ across a range of levels of imbalance. What differed drastically across methods was coverage probabilities. Results were consistent across simulations, so the following tables present the most extreme cases. Tables 3.5, 3.6, 3.7, and 3.8 show estimated treatment effect when the primary source of imbalance is hypertension and dropping either hyperlipidemia or hypertension in the 'incorrect' models. Tables 3.9, 3.10, 3.11, and 3.12 show estimated treatment effect when the primary source of imbalance is hyperlipidemia, and again dropping either hyperlipidemia or hypertension in the 'incorrect' models. These sets of tables were chosen to clearly demonstrate the trend when imbalance between the two groups was severe - similar patterns exist for alternative imbalances. The magnitude of the imbalance between black and white patients in terms of hyperlipidemia rates is smaller compared to cases where hypertension rates were altered; this magnitude is limited by the lower overall rate of hyperlipidemia in the simulated sample. However, we believe the results reported here are still indicative of the performance of the two methods under cases of extreme imbalance.

As shown in tables 3.5, 3.7, 3.9, and 3.11, the true simulated value of β_3 is two, and all models consistently underestimate this parameter. Correctly specified propensity score and traditional models more closely resemble each other in terms of point estimate, as do incorrectly specified propensity score and traditional models. This indicates that including the correct set of potential confounders in a model has a stronger affect on estimates of treatment effect than the method employed for estimation. However, both correct and

incorrect propensity score models have troublingly low coverage probabilities.

Both propensity score and traditional methods perform much worse when hyperlipidemia is the source of the imbalance. This is somewhat surprising since at this stage of the analysis both hypertension and hyperlipidemia were simulated to have the same strength of association with length of stay (i.e., $\beta_1 = \beta_2$). Additionally, despite their different distributions within the general simulated population, similar α values were used to generate the high levels of imbalance presented in the tables. However, in the hypertension simulations α_1 was slightly larger in magnitude (4.5) than α_2 used in the hyperlipidemia simulations (-4). It is possible that at extreme values, stronger associations between confounders and treatment assignment do produce slightly better estimates. This will be addressed further in later sections examining Drake's [1993] claims.

Similar patterns can be seen in tables 3.6, 3.8, 3.10, and 3.12 where no treatment effect exists ($\beta_3 = 0$). Again, propensity score models have troublingly low coverage probability and all models overestimate treatment effect.

When the difference between marginal mean outcomes was used to estimate treatment effect, the estimated 95% empirical probability intervals require a slightly different interpretation. Coverage probabilities in the tables previously presented refer to the proportion of the 1,000 simulated datasets that resulted in a 95% confidence interval around $\hat{\beta}_3$ that included the true parameter value β_3. In contrast, 95% empirical probability intervals in tables 3.13, 3.14 below, and tables A.7, A.8, A.9, A.10, A.11, and A.12 in the appendix indicate the range of estimated treatment effect sizes covered by 95% of the 1,000 simulations. These empirical probability intervals provide a measure of the variability of the point estimate of treatment effect across all simulations rather than the variability of each point estimate within each simulation. They tell a similar story - across every simulation, 95% empirical probability intervals for traditional and propensity methods (with either correct or incorrect model specifications) overlapped, indicating that even in the most extreme cases of confounding, the two methods provided statistically similar estimates of treatment effect (albeit sometimes statistically similar *incorrect* estimates). Additionally, when hypertension is the primary source of imbalance all methods result in empirical probability intervals

that include the true value of β_3 (both when $\beta_3 = 2$ and when $\beta_3 = 0$) whereas simulations where hyperlipidemia is the primary source of imbalance and no treatment effect exists some empirical probability intervals exclude this point estimate entirely.

Table 3.5: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.8)	2	1.54	0.61	0.70	84.5%
'Correct' PS (eq. 3.6)	2	1.55	0.62	0.70	85%
'Incorrect' Trad. (eq. 3.9)	2	1.54	0.86	0.70	95.3%
'Correct' Trad. (eq. 3.7)	2	1.55	0.86	0.70	95.6%

Table 3.6: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.8)	0	0.33	0.63	0.65	91.3%
'Correct' PS (eq. 3.6)	0	0.34	0.63	0.65	90.6%
'Incorrect' Trad. (eq. 3.9)	0	0.33	0.81	0.65	97.2%
'Correct' Trad. (eq. 3.7)	0	0.34	0.81	0.65	97.2%

Table 3.7: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.10)	2	1.61	0.50	0.61	84.1%
'Correct' PS (eq. 3.6)	2	1.55	0.62	0.70	85%
'Incorrect' Trad. (eq. 3.11)	2	1.61	0.68	0.61	93.3%
'Correct' Trad. (eq. 3.7)	2	1.55	0.86	0.70	95.6%

Table 3.8: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.10)	0	0.41	0.49	0.56	82.4%
'Correct' PS (eq. 3.6)	0	0.34	0.63	0.65	90.6%
'Incorrect' Trad. (eq. 3.11)	0	0.41	0.65	0.56	94.4%
'Correct' Trad. (eq. 3.7)	0	0.34	0.81	0.65	97.2%

In cases of very little confounding (hypertension rates of 73% vs. 69% and hyperlipidemia rates of 27% vs. 24% among black and white patients respectively) and a true treatment

Table 3.9: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_2 = -4.0$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.8)	2	1.13	0.35	0.33	27.6%
'Correct' PS (eq. 3.6)	2	1.30	0.41	0.34	62.8%
'Incorrect' Trad. (eq. 3.9)	2	1.13	0.46	0.33	55.1%
'Correct' Trad. (eq. 3.7)	2	1.30	0.54	0.34	84.9%

Table 3.10: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_2 = -4.0$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.8)	0	0.41	0.34	0.34	76.2%
'Correct' PS (eq. 3.6)	0	0.77	0.39	0.31	49.4%
'Incorrect' Trad. (eq. 3.9)	0	0.41	0.45	0.34	91%
'Correct' Trad. (eq. 3.7)	0	0.77	0.53	0.31	80.7%

Table 3.11: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_2 = -4.0$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.10)	2	1.31	0.41	0.34	63.7%
'Correct' PS (eq. 3.6)	2	1.30	0.41	0.34	62.8%
'Incorrect' Trad. (eq. 3.11)	2	1.31	0.54	0.34	85.1%
'Correct' Trad. (eq. 3.7)	2	1.30	0.54	0.34	84.9%

Table 3.12: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_2 = -4.0$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.10)	0	0.78	0.39	0.32	48.5%
'Correct' PS (eq. 3.6)	0	0.77	0.39	0.31	49.4%
'Incorrect' Trad. (eq. 3.11)	0	0.78	0.53	0.32	79.7%
'Correct' Trad. (eq. 3.7)	0	0.77	0.53	0.31	80.7%

effect all methods and model specifications performed quite well in terms of coverage probabilities (see table 3.15), but the similar quantity of bias for each method and model is quite troubling for what would likely be considered a small amount of confounding (all estimates of β_3 are off by nearly -0.5). Both traditional and propensity score regression adjustment methods should (and do) produce unbiased estimates in the absence of confounding, but

Table 3.13: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS (eq. 3.8)	1.542	(0.152, 2.939)
'Correct' PS (eq. 3.6)	1.557	(0.197, 2.951)
'Incorrect' Trad. (eq. 3.9)	1.543	(0.153, 2.935)
'Correct' Trad. (eq. 3.7)	1.555	(0.190, 2.926)
β_3	2	

Table 3.14: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS (eq. 3.8)	0.325	(-0.941, 1.563)
'Correct' PS (eq. 3.6)	0.338	(-0.956, 1.572)
'Incorrect' Trad. (eq. 3.9)	0.325	(-0.950, 1.569)
'Correct' Trad. (eq. 3.7)	0.336	(-0.959, 1.584)
β_3	0	

are known to produce biased estimates if adjustment fails to bring comparison groups into balance. This demonstration of the relatively small amount of initial imbalance that can still result in biased estimates from both methods, specifically in the case with dichotomous confounders, presents a nontrivial analytical challenge.

In both cases of a true simulated treatment effect and absent a true treatment effect we again see that the propensity score method results in lower coverage probabilities compared to the traditional adjustment method, however these are even more extreme in the absence of a true treatment effect (see table 3.16). Additionally, in the case of no treatment effect all methods and model specifications overestimate the true value of β_3 by more than 0.5, again highlighting the rapidity with which estimates become biased in the presence of confounding.

Although the primary question of interest is the performance of propensity score and traditional regression models under varying degrees of imbalance, in answering this question alpha values were varied, so we took the opportunity to also examine Drake's [1993] claims that propensity score bias decreases with increasing treatment effect and increasing covari-

Table 3.15: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 73% hypertension rates among black patients vs. 69% among white patients ($\alpha_1 = -0.2$) and 27% hyperlipidemia rates among black patients vs. 24% among white patients ($\alpha_2 = -0.2$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.8)	2	1.50	0.35	0.12	93.6%
'Correct' PS (eq. 3.6)	2	1.51	0.35	0.12	94.9%
'Incorrect' Trad. (eq. 3.9)	2	1.50	0.46	0.12	100%
'Correct' Trad. (eq. 3.7)	2	1.51	0.46	0.11	100%

Table 3.16: Comparing parameter estimates from 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 73% hypertension rates among black patients vs. 69% among white patients ($\alpha_1 = -0.2$) and 27% hyperlipidemia rates among black patients vs. 24% among white patients ($\alpha_2 = -0.2$)

Model	β_3	$\hat{\beta}_3$	**Avg SD**	**SD Across Sims**	**Coverage**
'Incorrect' PS (eq. 3.8)	0	0.62	0.34	0.12	66.1%
'Correct' PS (eq. 3.6)	0	0.63	0.34	0.12	62.7%
'Incorrect' Trad. (eq. 3.9)	0	0.62	0.45	0.11	99.8%
'Correct' Trad. (eq. 3.7)	0	0.63	0.45	0.11	99.5%

ate effect on treatment as well as response. Applying this to the models described above, increasing values of α_1 and α_2 should produce 'better' (less biased) estimates from the propensity score models (where increasing α_1 and α_2 translates into an increased covariate effect on treatment and increased levels of imbalance, i.e., confounding). Increasing β_1 and β_2 should also produce 'better' propensity score estimates (increasing covariate effect on response). Lastly, increasing β_3 (treatment effect) should produce 'better' propensity score estimates. Overall, we found none of these relationships to hold true in our models. Bias from both traditional and propensity score methods remained nearly equal across increasing values of α_1 and α_2 in our original analyses, although bias from both methods does appear to decrease as α values are increased. Subsequent simulations increasing β_3 (see table 3.17) similarly resulted in nearly equal biases between the two methods. Since the magnitude of the bias increases as β_3 increases, a measure of relative bias is also included in table 3.17. When increasing β_1 and β_2 (see table 3.18), a very minor decrease in bias was seen in the 'correct' propensity score estimates, however this decrease was arguably trivial in size and particularly in the case of β_2 was not seen until parameter values much larger than typically encountered in practice. It is important to remember in table 3.18 that although the true

simulated values of β_1 and β_2 are being altered, the estimated parameter of interest is still β_3 (and in these simulations is held constant at $\beta_3 = 2$).

Table 3.17: Estimated Bias Under Increasing Treatment Effect

Model	**'Correct'**	**'Incorrect'**	β_3
Propensity Score	0.06 (6%)	0.07 (7%)	1
Traditional	0.06 (6%)	0.07 (7%)	
Propensity Score	0.45 (22%)	0.46 (23%)	2
Traditional	0.45 (22%)	0.46 (23%)	
Propensity Score	0.81 (27%)	0.82 (27%)	3
Traditional	0.81 (27%)	0.82 (27%)	
Propensity Score	1.13 (28%)	1.15 (29%)	4
Traditional	1.13 (28%)	1.15 (29%)	
Propensity Score	1.43 (29%)	1.44 (29%)	5
Traditional	1.43 (29%)	1.44 (29%)	

Table 3.18: Estimated Bias Under Increasing Covariate Effect on Outcome

Model	**'Correct'**	**'Incorrect'**	β_1
Propensity Score	0.44	0.46	1
Traditional	0.45	0.46	
Propensity Score	0.43	0.45	2
Traditional	0.44	0.45	
Propensity Score	0.42	0.44	3
Traditional	0.42	0.44	
Propensity Score	0.41	0.42	4
Traditional	0.41	0.42	
Propensity Score	0.39	0.41	5
Traditional	0.40	0.41	
			β_2
Propensity Score	0.43	0.46	1
Traditional	0.43	0.46	
Propensity Score	0.40	0.45	2
Traditional	0.40	0.45	
Propensity Score	0.36	0.45	3
Traditional	0.37	0.45	
Propensity Score	0.33	0.44	4
Traditional	0.34	0.44	
Propensity Score	0.30	0.44	5
Traditional	0.32	0.44	

3.5.4 Discussion

As mentioned in the introduction, hypertension and hyperlipidemia were simulated with their respective frequencies based on the original Coverdell dataset, but this also provided a convenient comparison of potential confounders occuring with high frequency versus low frequency in the general population. Of course, it is worth noting that hyperlipidemia was somewhat arbitrarily selected as the confounder to be omitted in both sets of 'incorrect' models. Since hyperlipidemia was also simulated with a lower frequency in the general population, it is possible that switching and dropping hypertension from the models would

produce slightly different results. Therefore we repeated the most extreme examples of imbalance with incorrect models dropping hypertension (tables 3.7, 3.8, 3.11, and 3.12) and found conclusions to be similar. However, as noted in the previous section, when hyperlipidemia is the primary source of confounding, all models perform more poorly, with the most bias and lowest coverage probabilities when hyperlipidemia is also the confounder dropped in the 'incorrect' models. Which is not surprising, since a model failing to adjust for the primary source of confounding is bound to produce a rather inaccurate estimate. However, this performance was consistent across traditional and propensity score methods of adjustment.

It is also important to note that given the simulation methods described in section 3.5.2, the frequency of black and white patients was also affected by varying α parameter values. Therefore our results are linked to both an increasing or decreasing imbalance between simulated black and white patients in terms of hypertension and hyperlipidemia and an increasing or decreasing imbalance in the number of black vs. white patients. Given this same imbalance in our original dataset (which was approximately 75% white), and that observational studies frequently involve an uneven distribution among comparison groups, we believe this side effect of our simulations is still realistic. However, this is worth investigating further, and future simulations should certainly examine modeling structures that maintain a more consistent and even distribution of patients into comparison groups of interest.

3.5.5 Conclusion

The simulation results presented here indicate that both traditional and propensity score regression adjustment methods produce comparably biased estimates of treatment effect. The primary source of the remaining bias in estimates from both propensity score regression adjustment and traditional regression adjustment is the reliance of the distribution of race on the distributions of hypertension and hyperlipidemia (as indicated in equation 3.5). There is the possibility of additional bias in the propensity score regression adjustment estimate resulting from the estimation of the propensity score itself - even when the correct

confounders are included in the model, we are still only estimating the distribution of the race variable.

One of the primary concerns when estimating a propensity score is failing to account for any remaining imbalance in a variable that is strongly related to the outcome of interest (hence the balance-checking step is critical). Ho [2007] has shown that even a small amount of remaining imbalance can result in large bias in the final causal estimate. These simulations indicate that estimates from traditional regression adjustment methods are neither more nor less sensitive to remaining imbalance.

Not only did both adjustment methods consistently provide statistically similar estimates of treatment effect regardless of simulated levels of imbalance, 'correct' propensity score and traditional models resulted in estimates of treatment effect closer to each other than to their 'incorrect' counterparts. In other words, whether or not the 'correct' list of confounders is included in models adjusting for confounding has a stronger influence on the estimated treatment effect than the method employed to carry out the adjustment.

However, the coverage probability results presented in section 3.5.3 indicate that traditional adjustment methods should still be preferred over propensity score regression adjustment, regardless of level of imbalance. We find ourselves drawing a similar conclusion to Kang and Schafer [2007] (among other researchers in the field) that propensity score analyses provide useful additional information that careful statisticians would be remiss to dismiss, but directly adjusting for this information in an outcome model does not necessarily result in better estimates of treatment effect than more traditional methods.

Additionally, our findings do not confirm Drake's [1993] findings that propensity score regression adjustment bias will decrease as treatment effect and covariate effect on treatment as well as outcome increase. Several additional simulations were conducted varying treatment effect and covariate effect on outcome and the bias of the estimated treatment effect remained consistent across methods. Additionally, as α values increased (corresponding to a larger covariate effect on treatment, and specifically for these simulations, higher rates of hyperlipidemia or hypertension among black patients) the size of the bias remained similar

between propensity score and traditional regression adjustment methods.

It is important to note that Drake's [1993] findings are based on continuous, normally distributed confounders, and indeed much of the propensity score literature assumes potential confounders of this type. To continue to develop 'best propensity score practice' [Shadish et al., 2008] more research must involve a wider variety of potential confounders, both to examine the overall performance of the method and to develop useful rules of thumb for applying propensity score analyses to a variety of types of data.

Overall the set of simulations in section 3.5 appear to contradict results from the pseudo-simulations in the previous section (3.4), where significant differences in estimated treatment effect were found between the two methods. Given the additional complications included in the pseudo-simulations (complex survey design structure with stratification and clustering, additional potential confounders), it is difficult to determine precisely the cause of these measured differences.

Lastly, all of the above simulations focus only on comparing traditional regression adjustment to propensity score regression adjustment. There are many other propensity score adjustment methods - stratification, matching, and weighting. Regression adjustment was chosen for these comparisons since it is most often used by clinical researchers due to its ease of use and ready interpretability. Other propensity score adjustment methods arguably perform better than regression adjustment (most recently shown by Shadish et al.), and future work should repeat the above simulations with these other propensity score adjustment methods.

Chapter 4

Assessing Causal Effects with Truncation Due to Death and Missing Mortality Status

4.1 Background

Missing data represent a prevalent problem in all types of research. Frequently, missing observations are the result of loss to follow-up, in which case the data exist and would be theoretically possible to collect if we had more time and money for follow-up. In contrast, another type of 'missingness' (which will be more accurately referred to as truncation) results when some post-treatment variable ($S_i(Z)$) affects our ability to collect the primary outcome of interest. As mentioned in section 1.4, examples of this include school enrollment status, when the outcome of interest is a final test score, or employment status, when the outcome of interest is salary, or mortality status, when the outcome of interest is quality of life or functional status. In contrast to the type of missingness that results from loss to follow-up, this sort of truncation results in outcome measures that do not exist, at least not on the same measurement scale as the outcome of interest. It is critical that analyses account for this truncation in ways that clearly discriminate between it and other types of

missingness.

One such type of analysis is the application of principle stratification, which refers to stratifying data into homogenous groups of potential post-treatment variables $S_i(Z)$, which are not the primary outcome of interest, but may be affected by the treatment. Within this framework, causal inferences are only valid within a single strata, and in the specific instance where $S_i(Z)$ refers to mortality status, we are interested in estimating the treatment effect within the stratum where $S_i(t) = S_i(c) =$ alive. See section 1.4 for more details.

Both principle stratification itself and its specific application to truncation due to death problems are new techniques, so there are few examples of this type of analysis available [Zhang and Rubin, 2003, Hayden et al., 2005, Mattei and Mealli, 2007, Egleston et al., 2007, Frangakis et al., 2007, MacKenzie et al., 2008]. However, the application is much more prevalent in the compliance literature (see section 2.2) and in fact principle stratification language can be found in Bayesian analyses of the compliance problem [Imbens and Rubin, 1997] prior to Frangakis and Rubin's [2002b] formalization of the method.

In each of the above referenced studies, truncated outcome data are the only type of 'missingness' considered. Many longitudinal studies that model survival include both death as a competing risk and missing data due to loss to follow-up, but in all cases that we are aware of, missing data due to loss to follow-up refers to missing *outcome* data. Mortality status (or other general post-randomization variable $S_i(Z)$ used to determine principle strata) is always assumed completely known. In contrast, one of the things that is unique to our dataset is the combination of missing outcome data caused by truncation due to death coupled with missing mortality indicators due to loss to follow-up. The principle strata are defined by mortality status, so when these indicators are missing, additional assumptions are required to estimate the proportions of observations in each principle strata. Since all types of data are susceptible to loss to follow-up, we believe that our comparison of the affect of missing mortality status data on causal inference using principle stratification is a new and important contribution to the field. As far as we are aware, this is the first work to compare the sensitivity of causal estimates to principle strata assumptions versus missing data pattern assumptions.

4.2 Motivating Example

In the ProTECT study patients who had suffered a traumatic brain injury (TBI) were randomized to receive either progesterone or a placebo. This was a small pilot study designed primarily to determine the safety (rather than efficacy) of progesterone to treat acute TBI. Seventy seven patients received the treatment (progesterone) and 23 received placebo. Thirty days post-injury, two primary outcome measures were used to assess recovery - the Glasgow Outcome Score Extended (GOSE) and the Disability Rating Scale (DRS).

Patients were followed for one year, and DRS was measured again at this later follow-up time. The one year data were used for this paper and DRS was our outcome of interest for this secondary analysis.

One year post injury, seven of the control patients and fourteen of the treatment patients were known to be deceased. Five control patients and twenty five treatment patients were of unknown status.

4.2.1 Original Analyses

Due to the small sample size of the pilot study, original analyses included Fisher's exact test (for GOSE scores) and Wilcoxon's rank sum test (for DRS scores). All analyses were stratified by the initial severity of the injury (moderate vs. severe). At 30 days 40% of the severely injured patients randomized to placebo were deceased whereas only 13.2% of those randomized to treatment were deceased. Mortality rates among the moderately injured were similar across the treatment arms (14.3% among the control group versus 16.7% among the treated). Among the survivors with severe initial injury severity, 26.7% of patients in the control group had a GOSE score at 30 days indicative of moderate or good recovery versus 21.2% of the treated. Among the moderately injured survivors, none in the control group attained moderate to good recovery versus 55.6% in the treated. Additionally, although the confidence intervals overlap, the mean estimated DRS score among severely injured patients was lower (indicating more improvement) in the placebo group, whereas DRS results among

the moderately injured showed a significant gain for treatment patients [Wright et al., 2007].

It is worth noting that repeating the original study's Wilcoxon rank sum test on the one year follow-up data fails to find a statistically significant difference in DRS scores between the groups (p-value = 0.5, see Table 4.3 below), although patients in the treatment group do have a lower mean DRS (lower is better).

If the worst DRS score (29 - extreme vegetative state) is substituted for all patients who died before the one year follow-up (a common alternative to survivor-only analyses), Table 4.1 results,

Table 4.1: Mean DRS Assuming DRS = 29 for Deceased

	Mean DRS (SD)
Progesterone	9.1 (12.3)
Placebo	13.0 (13.7)

which still indicates improvement among the treated patients, but much worse recovery overall. Additionally, this is clearly misleading since it adds back potentially inaccurate outcome values for 30% of the control patients and 18% of the treated patients. However, even this amount of data replacement fails to produce a statistically significant difference between groups using a Wilcoxon rank sum test (p-value = 0.2).

Traditional methods are hampered by the significant amount of missing data resulting from the two different sources - by the one year follow-up 50% of outcome data are missing, either due to loss to follow-up or truncation due to death. Both survivor-only analyses and the alternative of substituting the worst possible outcome for all deceased patients are likely to provide misleading estimates of treatment effect. Therefore, alternative methods are needed.

4.3 Principle Stratification

To conduct the principle stratification analysis described in section 1.4, we would ideally be able to fill out a table such as Table 4.2.

Table 4.2: Principle Strata

Prob. of Principle Stratum Membership	Principle Stratum	$S_i(t)$	$S_i(c)$	$Y_i(t)$	$Y_i(c)$
π_{LL}	LL	1	1	$\in \Re$	$\in \Re$
π_{LD}	LD	1	0	$\in \Re$	*
π_{DL}	DL	0	1	*	$\in \Re$
π_{DD}	DD	0	0	*	*

Where $S_i(Z)$ is observed mortality status under treatment or control, $Y_i(Z)$ is observed outcome (DRS) under treatment and control (which is sometimes a valid DRS score on the real number line and sometimes truncated by death, indicated by *), and LL, LD, DL, and DD refer to the principle strata - those who would live under both treatments, those who would live under control but die under treatment, those who would die under control but live under treatment, and those who would die regardless of treatment, respectively. Instead, what we observe is Table 4.3.

Table 4.3: Principle Strata - Observed

Observed Group	% of population	Principle Stratum	Z_i	$S_i(Z)$	$\overline{Y}_i(Z)$ (SD)
OBS(tL)	55%	LL or LD	1	1	1.9 (2.8)
OBS (tD)	20%	DD or DL	1	0	*
OBS (cL)	15%	LL or DL	0	1	2.8 (5.3)
OBS (cD)	10%	DD or LD	0	0	*

The principle stratification approach is an attempt to tease out what proportion of the 70% of the sample that consists of a mixture of LL, DL, and LD patients are actually LL, and estimate a treatment effect within this principle strata.

4.4 Methods

Based on work by Zhang and Rubin [2003], the causal effect of treating TBI patients with the drug progesterone was estimated, taking into account both missing data due to loss to follow-up and missing outcome data caused by truncation due to death. Zhang and Rubin's work recommends a bound for the causal effect estimate, based on a weighted average of approximate principle strata membership. These four strata can never be directly observed, but we do know that the group we observe to survive under treatment is actually

a combination of those who would survive regardless of treatment group and those who would survive under treatment but die under control. Similar observations can be made about combinations of the principle strata in the four other observed groups (those who die under treatment, those who survive under control, and those who die under control, see Table 4.3). Depending on the characteristics of the data, and whether or not two important assumptions hold, these bounds may be narrowed (see monotonicity and stochastic dominance assumptions in section 1.4.2).

In addition to missing DRS information for those patients who died, our data also contained missing information due to loss to follow-up. Therefore we also calculated the causal effect intervals under each combination of monotonicity and stochastic dominance assumptions under four extreme boundary options for the pattern of missing mortality status:

- ignoring the missing data
- assuming everyone with missing data was dead by the one year follow-up
- assuming everyone with missing data was alive at the one year follow-up
  - with the lowest possible DRS of zero
  - with the highest observed DRS, 15 for progesterone patients, 18 for control.

Under these assumptions we observe Tables 4.4, 4.5, and 4.6.

Table 4.4: Ignoring Missing

Observed Strata	N	Proportion of Observed Group
OBS(t, L)	39	$P_{tL} = 39/53 = 0.74$
OBS(t, D)	14	$1 - P_{tL} = 14/53 = 0.26$
OBS(c, L)	11	$P_{cL} = 11/18 = 0.61$
OBS(c, D)	7	$1 - P_{cL} = 7/18 = 0.39$

Table 4.5: Assuming Missing Are Dead

Observed Strata	N	Proportion of Treatment Group
OBS(t, L)	39	$P_{tL} = 39/78 = 0.5$
OBS(t, D)	39	$1 - P_{tL} = 39/78 = 0.5$
OBS(c, L)	11	$P_{cL} = 11/23 = 0.48$
OBS(c, D)	12	$1 - P_{cL} = 12/23 = 0.52$

Table 4.6: Assuming Missing Are Alive

Observed Strata	N	Proportion of Treatment Group
OBS(t, L)	64	$P_{tL} = 64/78 = 0.82$
OBS(t, D)	14	$1 - P_{tL} = 14/78 = 0.18$
OBS(c, L)	16	$P_{cL} = 16/23 = 0.7$
OBS(c, D)	7	$1 - P_{cL} = 7/23 = 0.3$

More sophisticated methods for handling missing data (such as multiple imputation) were not considered following the results of the sensitivity analysis.

We also conducted two extensions of Zhang and Rubin's basic method, stratifying causal effect estimation by covariates (age group and initial severity) and propose one extension using Bayesian analyses.

For all 16 possible combinations plus two extensions, we calculated the causal effect bounds based on the equations provided by Zhang and Rubin (see Table 4.7), where π_g indicates the probability of an observation belonging to a given principle stratum g, $P_{z,s}$ is the proportion of observed individuals receiving treatment z with mortality status s and $\overline{Y}_{z,s}$ is the mean DRS in a given observed group (see section 1.4.1 for derivation of large sample bounds).

Our primary interest in this paper is to compare causal effects estimates under a variety of combinations of reasonable assumptions regarding the structure of the principle strata themselves and the pattern of missingness of the mortality data. In other words, we are interested in the sensitivity of causal effects estimates to these assumptions and therefore hope to make recommendations regarding which should be prioritized when conducting analyses.

4.5 Results

The large sample bounds for the sixteen possible combinations of principle strata and missingness assumptions described in section 4.2 are presented in Table 4.8. The intervals may appear misleading at first, but keep in mind that the Disability Rating Scale works like a golf score - lower is better, and 0 indicates no disability whatsoever. So all of the negative

Table 4.7: Large Sample Bounds for the Average Causal Effect on Y in the LL Principle Stratum - Table 6, Zhang and Rubin, 2003, slightly modified notation

A1	A2	Lower Bound, Upper Bound
No	No	$min_{\pi_{DL}}[\overline{Y}_{tL}(min\|P_{cL}/P_{tL} - \pi_{DL}/P_{tL}) - \overline{Y}_{cL}(max\|1 - \pi_{DL}/P_{cL})]$, $max_{\pi_{DL}}[\overline{Y}_{tL}(max\|P_{cL}/P_{tL} - \pi_{DL}/P_{tL}) - \overline{Y}_{cL}(min\|1 - \pi_{DL}/P_{cL})]$
Yes	No	$\overline{Y}_{tL}(min\|P_{cL}/P_{tL}) - \overline{Y}_{cL}, \overline{Y}_{tL}(max\|P_{cL}/P_{tL}) - \overline{Y}_{cL}$
No	Yes	$\overline{Y}_{tL} - max_{\pi_{DL}}[\overline{Y}_{cL}(max\|1 - \pi_{DL}/P_{cL})]$, $max_{\pi_{DL}}[\overline{Y}_{tL}(max\|P_{cL}/P_{tL} - \pi_{DL}/P_{cL})] - \overline{Y}_{cL}$
Yes	Yes	$\overline{Y}_{tL} - \overline{Y}_{cL}, \overline{Y}_{tL}(max\|P_{cL}/P_{tL}) - \overline{Y}_{cL}$

intervals indicate that the point estimate of $DRS_{trt} - DRS_{control}$ is expected to be negative, across a plausible range of proportions of observations falling in the LL strata, which in turn implies a beneficial treatment effect, among those who would be expected to live regardless of treatment assignment. The most extreme analysis, stacked against treatment, is the one that involves neither the monotonicity nor stochastic dominance assumptions (hence a broader causal effect interval) and assuming that all of the missing are alive with the worst observed DRS. Even in this case, treatment was found to be neutral at worst and given how skewed the interval is toward a large negative difference between treatment and control, indicative of a potential positive treatment effect at best. When both (plausible) assumptions are allowed to hold, producing the narrowest causal effect intervals, in all four possible missing data cases a positive treatment effect is estimated.

One of the boundary estimates did not produce a 'reasonable' result - when neither assumption is considered to hold and all missing observations are assumed dead, our sample size is too small to produce an interval because the second half of the boundary calculations indicated in Table 4.7 produce no observations. Therefore the point estimate listed as the second entry in Table 4.8 is actually the upper limit of that causal estimate. Additionally, the interval estimated when only assumption two is considered to hold and all missing observations are assumed dead is equivalent to the mean DRS among those observed to be alive in the treated group (as the upper bound) and the negative mean DRS among those observed to be alive in the control group (as the lower bound).

Table 4.8: Large Sample Bounds for Causal Effect Estimates

		Assumptions	DRS Causal Effect (t - c)
1	*2*	*Missing*	
no	no	ignored	(-0.79, -5.67)
no	no	assumed dead	-0.92*
no	no	assumed alive - DRS = 0	(-0.57, -2.58)
no	no	assumed alive - DRS = 15 (trtment) or 18 (control)	(0.77, -7.85)
yes	no	ignored	(-0.51, -1.88)
yes	no	assumed dead	(-0.82, -1.39)
yes	no	assumed alive - DRS = 0	(-0.60, -1.52)
yes	no	assumed alive - DRS = 15 (trtment) or 18 (control)	(0.6, -1.85)
no	yes	ignored	(1.90, -2.82)
no	yes	assumed dead	(1.90, -2.82)
no	yes	assumed alive - DRS = 0	(0.91, -1.94)
no	yes	assumed alive - DRS = 15 (trtment) or 18 (control)	(2.94, -6.01)
yes	yes	ignored	(-0.92, -1.88)
yes	yes	assumed dead	(-0.92, -1.49)
yes	yes	assumed alive - DRS = 0	(-0.78, -1.52)
yes	yes	assumed alive - DRS = 15 (trtment) or 18 (control)	(-0.54, -1.85)

Repeating analyses stratified by initial severity confirms the original study's findings at 30 days that moderate traumatic brain injury survivors had an improved outcome at one year when treated with progesterone (causal effect = (-4.2, -2.92)) whereas severely injured patients experienced a neutral effect (-3, 2.02) resulting in an overall neutral effect when combining these estimates based on the sample sizes in both groups (-3.37, 0.49). Lastly, when age is taken into account, progesterone again appears to cause an improvement in outcome, driven primarily by older patients (over 25 years old) (see Table 4.9).

Table 4.9: Causal Effects Estimates Stratified by Age Group

Age Group	DRS Causal Effect (treatment - control)
18-25	(0.92, -1.03)
26-42	(-0.05, -2.33)
43-82	-5.17*

In the age stratification we again see the case where we are missing a lower bound due to our small sample size.

It is also important to keep in mind that these are bounds on the *point estimate* of the causal effect based on an attempt to bound the likely proportion of observations that are

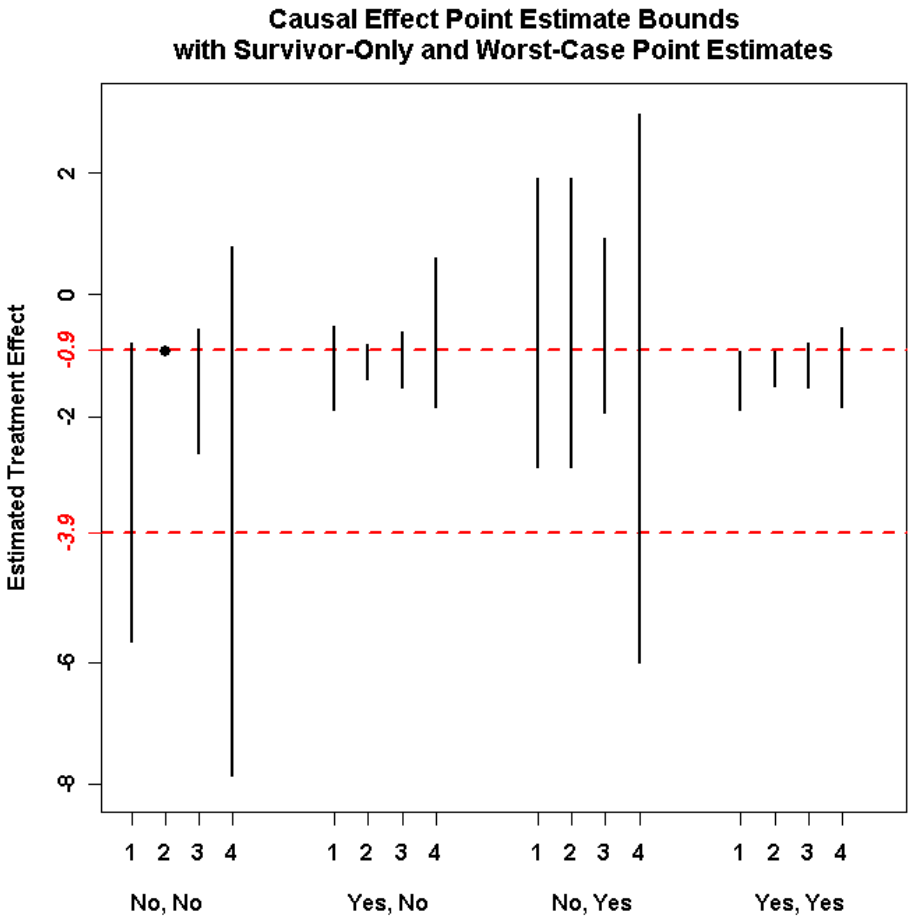

Figure 4.1: Bounds on Point Estimate of Causal Effect of Treatment; '1' = Ignore missing, '2' = Assume missing are dead, '3' = Assume missing alive with DRS = 0, '4' = Assume missing alive with DRS = 15 (progesterone) or 18 (control); 'No, No' → Neither monotonicity nor stochastic dominance assumption

in fact members of the LL principle strata. Additional variability is present, and is not accounted for in these intervals (see proposed confidence interval calculations in section on future work).

4.6 Conclusion

Overall, causal effect estimates appear to be more sensitive to assumptions about principle strata structure rather than missingness patterns. A positive treatment effect was estimated in seven out of eight analyses under the monotonicity assumption, regardless of assumed

missingness pattern. A null effect was estimated in five out of seven analyses without the monotonicity assumption, regardless of missingness pattern. This can also be seen in Figure 4.1, since the variability between principle stratification assumptions is clearly greater than within ('yes, no' groups refer to assumptions A1 and A2 in Table 4.8 and numbers 1 - 4 refer to the four possible missingness patterns; the horizontal dotted line at -0.9 is the survivor-only estimated treatment effect, and the horizontal dotted line at -3.9 is the worst case estimated treatment effect). Covariate analyses produced mixed results, with stratification by age resulting in a positive treatment effect and stratification by initial severity resulting in an estimated null effect.

Not only did causal inference analyses confirm the general direction of the original traditional analyses, we note that with causal inference a positive treatment effect was detected more often (even without boundary-narrowing assumptions). In terms of pilot studies and planning future studies this is critical - the ability to detect small indications of improvement, in the context of small sample sizes and a treatment that affects not only functional status but mortality as well, is needed in many medical research fields. Additionally, the traditional one-year analysis of the survivor-only group underestimated treatment effect ($\overline{Y}_t - \overline{Y}_c$ = -0.9) whereas substituting the lowest DRS value for those who died by the one-year follow-up overestimated treatment effect ($\overline{Y}_t - \overline{Y}_c$ = -3.9) (from Tables 4.3 and 4.1), both of which would have resulted in inaccurate power and sample size calculations for a phase III clinical trial of the affect of progesterone on TBI.

Of course, we must also keep in mind that these estimates are based on large-sample derivations, and obviously we are applying them to a dataset with a small sample size. This has led to some odd results, where the estimated proportion of observations of which to take the mean was zero, which clearly indicates that our data do not always match the large-sample theory. We must be cautious in interpreting these results, but nonetheless we believe this is an important first step toward taking mortality (and missing mortality status) into account in a more satisfying way than traditional analyses. Future work (proposed in the following section) suggests a Bayesian approach that may handle some of the small sample challenges better than the preceding application of large-sample theory.

4.7 Future Work

4.7.1 Confidence Intervals

As mentioned in the results section, the large sample bounds presented in this chapter are bounds on possible values of the point estimate itself, they do not represent a confidence interval. To draw conclusions regarding the potential statistical significance of the causal estimates calculated in this chapter we would need to account for variability in both the estimate of the probability that an individual is a member of a given principle strata g and the estimate of the average DRS outcome $\overline{Y}$.

One place to start in calculating confidence bounds would be to simply calculate standard deviations for the proportion of $Y_{z,s}$ used in each causal effect estimate. For example, when neither assumption is considered true, the lower point estimate bound is

$$min_{\pi_{DL}}[\overline{Y}_{tL}(min|P_{cL}/P_{tL} - \pi_{DL}/P_{tL}) - \overline{Y}_{cL}(max|1 - \pi_{DL}/P_{cL})] \tag{4.1}$$

(from Table 4.7). If missing mortality status data are ignored, the minimum π_{DL} value is 0. Recall that π_{DL} is bounded

$$max(0, P_{cL} - P_{tL}) \leq \pi_{DL} \leq min(P_{cL}, 1 - P_{tL}),$$

$P_{cL} - P_{tL} = 0.61 - 0.74 = -0.13$, $P_{cL} = 0.61$, and $1 - P_{tL} = 1 - 0.74 = 0.26$ from section 4.4, so

$$max(0, -0.13) \leq \pi_{DL} \leq min(0.61, 0.26) \Rightarrow 0 \leq \pi_{DL} \leq 0.26.$$

Then the left hand side of equation 4.1 is $\overline{Y}_{tL}(min|P_{cL}/P_{tL} - 0)$, which indicates to take the average of the lowest [1] 32 DRS values (a proportion $P_{cL}/P_{tL} = 0.61/0.74 = 0.82$ of the total of 39) in the observed treated and alive group. The right hand side then indicates to subtract the average of the highest 11 DRS values (a proportion $1 - \pi_{DL}/P_{CL} \Rightarrow 100\%$, for $\pi_{DL} = 0$) in the observed control and alive group. We could then treat this like a

[1] 'highest' and 'lowest' must be interpreted carefully, since DRS = 0 is actually the best, or 'highest' outcome value

traditional difference of means and calculate the lower confidence interval bound by subtracting either $t^{n_1+n_2-2}_{1-\alpha/2}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ or $t^{n_1+n_2-2}_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ (depending on assumptions about underlying variance structures) from the estimated difference of means, to achieve an approximate lower confidence bound. Of course, this implies an assumption of normality, or a large enough sample size to invoke the central limit theorem, which may or may not be applicable.

Similarly, the upper point estimate bound is

$$max_{\pi_{DL}}[\overline{Y}_{tL}(max|P_{cL}/P_{tL}-\pi_{DL}/P_{tL})-\overline{Y}_{cL}(min|1-\pi_{DL}/P_{cL})] \tag{4.2}$$

(from Table 4.7). From the calculations above, this indicates to take the average of the highest 15 DRS values (a proportion $P_{cL}/P_{tL}-\pi_{DL}/P_{tL}=0.82-0.26/0.74=0.39$ of the total of 39) in the observed treated and alive group and subtract the average of the lowest 6 DRS values in the observed control and alive groups (a proportion $1-0.26/0.61=0.57$ of the total of 11). Again, treating this as a traditional difference of means, and calculating the appropriate SD as above, but this time adding to the upper point estimate bound to find the upper confidence bound.

Although this method would provide a starting point in estimating confidence intervals for the estimated causal effect, it only takes one source of variability into account. The conditional distribution of observations belonging in the LL principle stratum given that they were observed in the tL group could be modeled as a binomial distribution (similarly membership in the LL principle stratum given observed in the cL group). More traditional estimates of variability (i.e., $\sigma^2=npq$ for binomially distributed random variables) could then be used to estimate this source of variability. Of course, the next analytical question would then be how best to combine these two sources of variability into a single estimate of confidence interval limits.

4.7.2 Bayesian Methods

We propose extending Zhang and Rubin's (2003) initial analysis to include a Bayesian approach. This is essentially a two-step hierarchical process - defining the distribution of patients into each of the four possible principle strata and, given that, defining the probable outcome within each strata. In other words, modeling the potential outcomes

$$f(S(t), S(c), Y(t), Y(c)|\mathbf{X}) = f(S(t), S(c)|\mathbf{X}) f(Y(t), Y(c)|S(t), S(c), \mathbf{X})$$

In their study of the effect of job training on wages, Zhang, Rubin, and Mealli [2006] start with a multinomial distribution of probabilities of principle strata membership and assume that their outcome of interest (wages) is normally distributed.

$$P(G_i = g|\mathbf{X}_i, \theta) = \frac{exp(\alpha_g + \mathbf{X}_i^T \beta_g)}{\sum_{g'} exp(\alpha_{g'} + \mathbf{X}_i^T \beta_{g'})} \tag{4.3}$$

and

$$(log(Y_i(z))|G_i = g, \mathbf{X}_i, \theta) \sim N(\mu_{g,z} + \eta_{g,z}^T \mathbf{X}_i, \sigma_{g,z}^2)$$

where g indicates principle strata and z treatment.

We will assume the same model for our principle strata, but with the addition of a simplifying assumption - no 'denier' group (i.e., that no patients died under treatment but survived under control), leaving three possible categories for our multinomial distribution.

Instead of a normally distributed outcome, we propose transforming the pseudo-continuous outcome measure (DRS) into a dichotomous outcome (probability of being disabled, which is equivalent to DRS > 0 vs. DRS $= 0$) and assume a binomial distribution with a stratum-specific probability p_g assigned to each stratum g. The collection of unknown parameters θ is then $\theta = \{(\alpha_g, \beta_g, p_g), g \in \{LL, LD, DD\}\}$

Under our model, the likelihood function is

$$L(\mathbf{X}, \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{Y}^{obs}|\theta) \propto \prod_{i \in O(1,1)} [\pi_{LL:i} Bin_i(p_{LL}) + \pi_{LD:i} Bin_i(p_{LD})] \times \prod_{i \in O(1,0)} \pi_{DD:i} \times \prod_{i \in O(0,1)} [\pi_{LL:i} Bin_i(p_{LL})] \times \prod_{i \in O(0,0)} (\pi_{LD:i} + \pi_{DD:i}). \quad (4.4)$$

Where $\pi_{g,i} = P(G_i = g|\mathbf{X}_i, \theta)$ for $g \in \{LL, LD, DD\}$. The first and third terms account for those who received treatment and were alive at one year and those who received placebo and were alive at one year respectively. The second and last terms account for those who received treatment and placebo respectively and were not alive at one year, and hence are lacking the binomial component for disability since their probability of disability has been truncated. Since we are maintaining Zhang, Rubin, and Mealli's (2006) principle stratum distribution, we will also assume the same prior distribution

$$(\alpha_g, \beta_g) \sim N(\mathbf{0}, K_0\mathbf{I}) \quad (4.5)$$

for each $g \in \{LL, LD, DD\}$

Equation 4.4 indicates that it would be challenging to draw our parameters of interest from their posterior distribution. Fortunately, if the principle strata could be treated as known and we could condition on them, the posterior distribution of θ has a much preferable structure:

$$f(\theta|\mathbf{G}, \mathbf{X}, \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{Y}^{obs}) \propto L(\mathbf{G}, \mathbf{X}, \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{Y}^{obs}|\theta)p(\theta). \quad (4.6)$$

Continuing to follow Zhang, Rubin, and Mealli [2006], we will then use their suggested data augmentation approach to iteratively impute $\mathbf{G}$ given θ and draw θ given $\mathbf{G}$. For starting values, we will randomly draw $\theta_+ = \{\alpha_g, \beta_g\}$ from 4.5 and then compute $\mathbf{G}$ from equation

4.3 (using $\mathbf{X}$ = initial severity) and then draw an updated θ_+ from

$$f(\theta_+|\mathbf{p}_g, \mathbf{G}, \mathbf{X}, \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{Y}^{obs}) \propto p(\theta_+) \times \prod_{i \epsilon O(1,1) \bigcap LL} \pi_{LL:i} \times \prod_{i \epsilon O(1,1) \bigcap LD} \pi_{LD:i} \times \prod_{i \epsilon O(1,0) \bigcap DD} \pi_{DD:i} \times \prod_{i \epsilon O(0,1) \bigcap LL} \pi_{LL:i} \times \prod_{i \epsilon O(0,0) \bigcap LD} \pi_{LD:i} \times \prod_{i \epsilon O(0,0) \bigcap DD} \pi_{DD:i} \quad (4.7)$$

where $p(\theta_+)$ is the prior distribution defined in 4.5. The conditional posterior mode and covariance matrix of θ_+ will then be $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ respectively and can be used to construct a multivariate t distribution $t_v(\mathbf{\Lambda}, \mathbf{\Omega})$. θ_+ will then be updated with values from $t_v(\mathbf{\Lambda}, \mathbf{\Omega})$ with probability

$$min\left(1, \frac{f(\theta_+^{new}|p_g, G, X, Z, S^{obs}, Y^{obs}) t_v(\theta_+^{cur}|\Lambda, \Omega)}{f(\theta_+^{cur}|p_g, G, X, Z, S^{obs}, Y^{obs}) t_v(\theta_+^{new}|\Lambda, \Omega)}\right).$$

At this stage, Zhang, Rubin, and Mealli [2006] assume their normally distributed outcome and define subsequent prior distributions accordingly. In contrast, we assume a dichotomous outcome with a binomial distribution, so we will assume a conjugate beta prior distribution for the binomial probability p_g

$$p_g \sim Beta(a, b). \quad (4.8)$$

The beta parameters a and b could be chosen such that a higher frequency of low p values (i.e., lower probability of disability) is generated among the LL strata and a higher frequency of high p values (i.e., higher probability of disability) among the DD strata, to incorporate the assumption that the LL strata includes healthier patients overall, and therefore patients with a lower probability of disability.

Alternatively, we could model the probability of disability, p, as

$$p_g = \beta_{0,g} + \beta_{1,g} age$$

and then the prior for $\{\beta_{0,g}, \beta_{1,g}\}$ would be $N_v(\mu_g, \sigma_g^2)$. Either way, we will eventually be

working toward a posterior estimate of the causal effect of treatment:

$$P(disability|LL, treatment\ group) - P(disability|LL, control\ group).$$

Although Zhang, Rubin, and Mealli provide the initial starting point for this problem with their Bayesian analysis of the employment and wage problem, we are not aware of any other studies that take the particular Bayesian approach we have with a dichotomous outcome variable to the truncation due to death problem.

Additionally, we could develop a prior for the possible distribution of missing mortality status indicators ($S_i(Z)$) and include that in our model (unknown parameter θ would then include a factor for $P(S_i(z) = s|Z_i = z, S_i^{obs} = .)$). An obvious starting value for the Markov Chain Monte Carlo method would then be $P(S_i(t) = 1) = p_t = 0.26$ (the observed proportion of deceased individuals in the treatment group) and $P(S_i(c) = 1) = p_c = 0.52$ (the observed proportion of deceased individuals in the control group). The prior distribution for mortality status, given missing observed mortality status, would then be $Bin(p_Z)$ and this information could be used to improve the estimation of the distribution of observations among the principle strata ($P(G_i = g|X_i, \theta)$).

Chapter 5

Prognostic Scores and Sliding Dichotomy

5.1 Background

Many studies rely on a measure of functional status as an outcome of interest. In stroke and traumatic brain injury (TBI) research, a commonly used scale is the Glasgow Outcome Scale (GOS), which is composed of five categories - dead, vegetative state, severe disability, moderate disability, and good recovery. Traditionally, a patient is categorized as having a favorable outcome if he or she achieves a categorization of moderate disability or good recovery at some post-treatment time point (say, three or six months later). Clinical trials evaluating potential new treatments then define a successful treatment as one that results in a ten percentage point increase in favorable outcomes among patients in the treatment group as compared to those in the control group. Unfortunately, a successful treatment has not been identified in either field (with the exception of tissue plasminogen activator for stroke) in many years. This has led some researchers to hypothesize that the problem is not a lack of successful treatment options but rather a highly heterogeneous patient population resulting in frequently under-powered trials. In particular, many argue that the current standard definition of a successful treatment implies that every patient has the same

probability of achieving a favorable outcome, which is simply clinically untrue - probabilities of favorable outcomes are clearly affected by patient characteristics and the initial severity of injury [Barer, 1998, Maas et al., 1999].

Many research fields encounter heterogenous patient populations and there are numerous analytical methods available to adjust for a patient population with varying probabilities of favorable outcomes. For example, polytomous regression may be used, with the outcome of interest being whether or not a patient improves one or more categories rather than applying a single threshold to the entire patient population. This approach is typically not used in TBI research, potentially due to Choi's [2002] findings that a categorical response measure does not necessarily translate into higher powered trials, primarily due to the increased risk of misclassification error. Another way to address a heterogeneous patient population is to adjust for baseline characteristics in a final regression model. This is commonly done in TBI research, but it is unclear how often this analytical approach is taken into account in power and sample size calculations during the design stage of a clinical trial.

A newer approach gaining popularity in the TBI field is one called the sliding dichotomy. Instead of adjusting the final analysis for variables that may affect the relationship between outcome and treatment assignment, the sliding dichotomy method attempts to make a portion of this adjustment up front by more specifically tailoring the definition of a favorable outcome to an individual patient's recovery prognosis. More specifically, the sliding dichotomy slides the cutpoint for favorability up or down the outcome scale depending on which prognostic category a patient has been grouped into. Currently in the TBI field, this has only been well-defined for outcomes based on the GOS.

Murray [2005] suggests grouping patients into 'worst,' 'intermediate,' or 'best' prognostic categories, determined by the tertiles of predicted probabilities from a predictive model. Among those with the 'worst' prognosis, a GOS of severe disability, moderate disability, or good recovery may all be considered a favorable outcome. Those with an 'intermediate' prognosis maintain the traditional definition of a favorable outcome (moderate disability or good recovery) and those with the 'best' prognosis are given a slightly higher bar to clear with only good recovery counting toward a favorable outcome (see figure 1.1). This definition

of a favorable outcome based on the sliding dichotomy method will be used throughout this chapter. Machado [1999], Maas [1999], and Murray [2005] suggest that such an approach will lead to higher powered clinical trials.

Initially, the research in this chapter was motivated by an interest in developing better predictive models for defining prognostic categories for the sliding dichotomy approach as defined above. Several predictive models currently exist in both TBI and stroke literature (see section 2.3), however, few have been validated. A systematic review of prognostic models for TBI [Perel et al., 2006] found that only 38% of current models were validated as part of the development process, and that only 11% of those were validated in an external population. Additionally, there does not exist a general method for developing statistically sound predictive models for a variety of study characteristics in either field. Therefore, we considered this to be a vital gap in the literature and hoped to advance TBI research by applying Hansen's prognostic score theory (2006, 2008) to the development of a general method for predictive modeling. Results from this method would then be compared to results from two validated predictive models from the TBI literature [Hukkelhoven et al., 2005, Murray et al., 2005].

However, in the process of evaluating these predictive models, another gap in the literature became clear - although Machado [1999], Maas [1999], and Murray [2005] suggest that the sliding dichotomy approach results in higher powered clinical trials, only two formal comparisons of the sliding dichotomy approach and the traditional definition of a favorable outcome exist in the current literature. Young's [2003] work relates to stroke patients, so although her research lays the conceptual groundwork for an alternative definition of a favorable outcome, it is not directly comparable to this research since Young's work involved different outcome measures. Therefore, Machado [1999] provides the only example of a power analysis using the siding dichotomy as defined for the GOS. Machado concludes that "...a strategy of recruiting only patients with an intermediate prognosis allows the sample size to be reduced by the order of 30% with no loss of statistical power [1999]." Unfortunately, it is unclear how these power calculations were conducted. Machado [1999] uses data from the European Brain Injury Consortium (EBIC) Core Data Survey to develop

a predictive model linking age, Glasgow Coma Scale (GCS) motor score, and computed tomography (CT) classification to a traditional favorable outcome based on GOS at six months (see Eq. 5.1, see Machado et al. [1999] for categorical variable breakdown). He then formalizes a range of ways that a treatment effect may modify this association and then simulates datasets based on the EBIC survey within which to conduct power analyses. Although the original dataset contained 689 patients who met the exclusion/inclusion criteria, simulated datasets included 1,000 hypothetical TBI patients. A uniform potential treatment effect was incorporated into logistic regression models as an additional covariate with a simulated parameter estimate of 0.557, translating into an odds ratio of 1.75 (see Eq. 5.2). This resulted in an increase in the proportion of observed favorable outcomes from 51% to 61% across the entire simulated population. Machado [1999] claims that

> [t]his model means, for example, that a patient with an intermediate prognosis, say aged 36-50, with motor score 4 and CT class 5/6 (mass lesion), has a predicted probability of 0.48 of having a favorable outcome on placebo, which increases to 0.62 on active treatment. The same odds ratio for a patient with a very favorable prognosis, say aged 16-25, motor score 5/6 and CT class 1 (no visible pathology) corresponds to a probability of having a favorable outcome, which increases from 0.95 to 0.97 on active treatment.

Additional simulations also included alternative treatment effects, including a treatment that provided greater benefit to those with an intermediate prognosis and a treatment that benefited only those with a specific CT classification.

$$\begin{aligned} logit(P(fav=1)) &= 2.979 - 0.196age1 - 0.709age2 - 1.634age3 - 2.442age4 \\ &- 1.912GCS1 - 0.981GCS2 - 1.162GCS3 - 1.060CT1 - 2.046CT2 - 1.361CT3 \end{aligned} \tag{5.1}$$

$$\begin{aligned} logit(P(fav=1)) &= 2.979 - 0.196age1 - 0.709age2 - 1.634age3 - 2.442age4 - 1.912GCS1 \\ &- 0.981GCS2 - 1.162GCS3 - 1.060CT1 - 2.046CT2 - 1.361CT3 + 0.557treatment \end{aligned} \tag{5.2}$$

Analyses in this chapter include a general method for developing a predictive model (loosely based on Hansen's prognostic score technique) and a formal comparison of the power achieved by sliding dichotomy and traditional definitions of favorable outcomes. The latter analysis concludes with guidelines regarding under what conditions each method results in higher power.

5.2 Motivating Example

Research was motivated by data from the National Acute Brain Injury Study: Hypothermia (NABISH), a multicenter randomized trial that recruited 392 patients, 16-65 years of age, with severe head trauma, from October 1994 through May 1998. The original aim of the study was to determine the effect of induced hypothermia on functional status six months post-injury. A traditional favorable outcome was used (good recovery or moderate disability vs. severe disability, vegetative state, or death according to the GOS). Patient enrollment was stopped in May 1998 by the patient safety and monitoring board because it was determined "...that the probability of detecting a treatment effect was less than 0.01 if the trial expanded to include 500 patients." The original study used the Wilcoxon Rank-Sum test to assess a difference in favorable outcomes between treatment groups [Clifton et al., 2001].

Our goal is to use prognostic scores (as described in section 1.6) to divide patients from the NABISH dataset into prognostic tertiles with corresponding definitions of favorable outcomes as recommended by Murray [2005]. The performance of prognostic scores in assigning patients to these groups will then be compared to two existing predictive models from the literature [Hukkelhoven et al., 2005, Murray et al., 2005]. We hope to make recommendations regarding our approach versus those from the literature in terms of their application to the sliding dichotomy method of designing future studies, however as mentioned in the previous section, an important question regarding power and sample size must be answered first.

A formal analysis of the power of the two methods to detect a potential treatment effect

will begin with simulations. Starting values for these simulations will be based on initial analyses using Hansen's [2006] prognostic score approach and the two predictive models from the literature [Hukkelhoven et al., 2005, Murray et al., 2005].

5.3 Methods - Developing Predictive Models

The next section will elaborate on all of the assumptions necessary to estimate the power of an analysis using a traditionally defined outcome versus a sliding dichotomy outcome. However, one of the first requirements of such an analysis is an estimate of the probability of a favorable outcome under either definition. Simulations will vary this probability, but we must define a reasonable parameter space within which to run the simulations. This reasonable parameter space will be defined based on initial results from two validated predictive models from the TBI literature [Hukkelhoven et al., 2005, Murray et al., 2005] and Hansen's prognostic score approach [2006].

Hukkelhoven [2005] models the probability of a favorable outcome (dichotomized using the traditional definition based on GOS) using the covariates age (as both a main effect and squared term), GCS motor score, pupil reactivity, and CT score.

$$logit(P(fav = 1)) = \beta_0 + \beta_1 age + \beta_2 GCS + \beta_3 pupil + \beta_4 CT + \beta_5 age^2 \tag{5.3}$$

As originally published, Hukkelhoven's model also included indicators for hypotension, hypoxia, and traumatic subarachnoid haemorrhage (SAH).

$$\begin{aligned} logit(P(fav = 1)) =& \beta_0 + \beta_1 age + \beta_2 GCS + \beta_3 pupil + \beta_4 CT + \beta_5 hypotension \\ &+ \beta_6 hypoxia + \beta_7 SAH + \beta_8 age^2 \end{aligned}$$

Unfortunately, the NABISH dataset does not include information on these covariates, so they were omitted from the model. This is one of the challenges of using a predictive model from the literature - not all studies collect data on the same set of covariates. Although there is a movement within the TBI field to develop a standard set of predictive covariates

that all studies will collect, this has not yet been formalized. Additionally, even when data are collected on a common set of covariates, the distribution of those covariates will not always be similar across studies and therefore researchers may make individual decisions regarding how best to analyze data - for example, age may be treated as continuous or categorical, GCS motor score categories may be combined due to low sample size, etc.

Hukkelhoven [2005] suggests using his logistic regression model to generate deciles of predicted probabilities of favorable outcomes (see table 5.1 for deciles of predicted probability of favorable outcome versus observed favorable/unfavorable outcome in the NABISH dataset). How best to break down these categories of predicted probabilities into prognostic regions is left up to the reader/researcher/clinician. However, one can see from table 5.1 that although there does appear to be a natural division around either the fourth or seventh decile, the predicted probabilities still do not appear to discriminate very well between the observed favorable and unfavorable outcomes.

Table 5.1: Deciles of Predicted Probabilities from Hukkelhoven Model (5.3) versus Observed Outcome

Decile of Pred. Prob.	**Observed Favorable Outcome**	
	No	**Yes**
0	30 (88%)	4 (12%)
1	28 (80%)	6 (17%)
2	27 (77%)	6 (17%)
3	27 (77%)	7 (20%)
4	22 (65%)	11 (32%)
5	29 (83%)	5 (14%)
6	26 (74%)	9 (26%)
7	17 (50%)	17 (50%)
8	17 (47%)	19 (53%)
9	12 (35%)	21 (62%)

To match Murray's [2005] sliding dichotomy definition, we instead divided predicted probabilities from Hukkelhoven's model into tertiles, with favorable outcomes defined for each tertile as described in section 5.1. Within the group identified as having the 'worst' prognosis (i.e., the lowest tertile of predicted probabilities, according to Hukkelhoven's model), 15% were observed to attain the traditionally defined favorable outcome and 39% were observed to attain the sliding dichotomy definition of a favorable outcome. Within the 'moderate'

group 22% attained a favorable outcome (defined as the same GOS categories for both traditional and sliding dichotomy) and within the 'best' group 53% attained a traditional favorable outcome while 46% attained the sliding dichotomy version of a favorable outcome (see Table 5.2).

Murray [2005] proposed two proportional odds models, both with GOS (treated as ordinal; not dichotomized into favorable and unfavorable) as the outcome of interest. Model A included age, GCS motor score, and pupil reactivity, model B included the same covariates with the addition of CT score.

$$logit(P(Y = k)) = \alpha_k + \beta_1 age + \beta_2 GCS + \beta_3 pupil \tag{5.4}$$

$$logit(P(Y = k)) = \alpha_k + \beta_1 age + \beta_2 GCS + \beta_3 pupil + \beta_4 CT \tag{5.5}$$

These proportional odds models produce predicted probabilities for each of the five GOS categories. Therefore the category with the maximum predicted probability was taken to be the predicted GOS outcome for each individual. Model A resulted in only two predicted GOS categories, (GOS = 1 (dead) or 3 (severely disabled)) and so was discarded in favor of model B, which resulted in three predicted GOS categories. GOS = 1 was treated as the 'worst' prognostic category, and 34% of patients in this group were observed to attain a traditional favorable outcome while 62% attained the sliding dichotomy favorable outcome. GOS = 3 was treated as the 'moderate' prognostic category, with 27% of patients in this group attaining a favorable outcome. Lastly, GOS = 4 (moderately disabled) was considered the 'best' prognostic group and 67% of these patients attained a favorable outcome, using either the traditional or sliding dichotomy definition (see Table 5.2).

Next we used Hansen's [2006] suggestion of modeling outcome measures looking only at the subpopulation of control patients. Stepwise model selection indicated that GCS motor score, pupil reactivity, and age were significantly associated with a dichotomous favorable outcome and GCS score and pupil reactivity were significantly associated with the ordinal GOS (in control patients only). Since Hukkelhoven's model also included age as a squared term this was checked in our logistic regression model, but adding this higher order term

did not appear to improve our model.

$$logit(P(fav_c = 1)) = \beta_0 + \beta_1 GCS + \beta_2 pupil + \beta_3 age \tag{5.6}$$

$$logit(P(Y_c = k)) = \alpha_k + \beta_1 GCS + \beta_2 pupil \tag{5.7}$$

According to Hansen [2006], the next step in this procedure would be to estimate a formal prognostic score $\Psi(X)$, such that outcomes in the control patients are conditionally independent of a set of covariates X given some prognostic score $\Psi(X)$, as defined in section 1.6

$$Y_c \perp X | \Psi(X), X \epsilon A.$$

Similar to how the propensity score $e(x)$ was estimated in chapter three, $\Psi(X)$ is the linear combination of the covariates chosen in the above model, using the parameter estimates from the model including only the control patients, but applied to the entire patient population (see appendix for parameter estimates used to calculate $\Psi(X)$).

The final outcome model should then adjust for $\Psi(X)$ as a balancing technique, similar to the propensity score adjustment from chapter three. It is possible that such an adjustment, rather than using $\Psi(X)$ to define a new outcome, would improve analysts' abilities to detect a statistically significant treatment effect in otherwise under-powered clinical trials. Exploration of this possibility is suggested for future work.

Instead, we used parameter estimates from this model, based only on control patients, to estimate predictive probabilities for the entire study population. In this way our method more closely resembles that suggested by Peters [1941] and Belson [1956], however with a slight modification. In the Peters-Belson method, individual prediction values (typically from linear rather than logistic regression) would be used to estimate $\overline{D}$, the adjusted treatment effect.

$$\overline{D} = \frac{1}{n_t} \sum_{i=1}^{n_t} (Y_{ti} - \hat{Y}_{ti})$$

where Y_{ti} are the observed responses from the treatment group and $\hat{Y}_{ti}$ are the predicted responses, based on parameter estimates from a regression model fit to the control group only. Whether or not this difference is statistically significantly different from zero requires a specific test based on the asymptotics of $\overline{D}$ (see Gastwirth and Greenhouse [1995] for derivation). In contrast, we are not interested in whether or not this difference is statistically significantly different from zero, but rather in using the predicted values $\hat{Y}_{ci}$ and $\hat{Y}_{ti}$ (or probabilities, as they are in this case) to define prognostic categories of patients within which to define a new outcome variable. This represents a new method of developing predictive models in TBI research.

Similar to the use of predicted probabilities from the Hukkelhoven and Murray models above, we divided the predicted probabilities from the first Hansen model (with dichotomous GOS as outcome, Eq 5.6) into tertiles and selected the GOS value with the highest predicted probabilities (from the model with ordinal GOS as the outcome, EQ 5.7) as the predicted GOS grouping. Based on the first model, 19% of those grouped in the 'worst' prognostic category attained a traditionally defined favorable outcome while 46% attained a favorable outcome defined by the sliding dichotomy. Of those in the 'moderate' category 26% attained a favorable outcome, and 47% of those in the 'best' prognostic group attained a traditional favorable outcome while 40% of them attained the sliding dichotomy definition of favorability. For the model with GOS as the ordinal outcome, the 'worst' prognostic group were those with a predicted GOS of 1, and 25% of these attained a traditional favorable outcome while 47% attained a favorable outcome on the sliding dichotomy scale. The 'moderate' group were those with a predicted GOS of 3 and 15% of them attained a favorable outcome (on either scale). Lastly, the 'best' prognostic category were those with a predicted GOS of 5, and 45% of them attained a traditional favorable outcome while 38% of them attained favorability as defined by the sliding dichotomy approach. All model results are summarized in table 5.2.

These observed proportions were treated as the probability of achieving a favorable outcome under either definition for each prognostic group and were used to define the initial parameter space for the simulations in the following section.

Table 5.2: Observed Proportion of Patients Achieving Favorable Outcome Under Either Definition

Prognostic Group	Traditional Favorable	Sliding Dichotomy Favorable
Hukkelhoven Model (5.3)		
Worst	15%	39%
Moderate	22%	22%
Best	53%	46%
Murray Model (5.5)		
Worst	34%	62%
Moderate	27%	27%
Best	67%	67%
Hansen-Hukkelhoven Model (5.6)		
Worst	19%	46%
Moderate	26%	26%
Best	47%	40%
Hansen-Murray Model (5.7)		
Worst	25%	47%
Moderate	15%	15%
Best	45%	38%

5.4 Methods - Simulations

We used the results from the four predictive models described in the previous section as starting points to simulate data to compare power and sample size calculations when favorable outcomes are defined traditionally versus by a sliding dichotomy.

Several assumptions are required in these simulations. First is the assumption that researchers suggesting the need for a sliding dichotomy are correct, that TBI and stroke patient populations are heterogeneous and that traditional definitions of favorable outcomes fail to take this heterogeneity into account. This implies that were we able to identify patients by prognosis (or prognostic group) we could specify their probability of a favorable outcome, which would be different from members of other prognostic groups. For simplicity, we start with three prognostic groups - 'best', 'moderate,' and 'worst.'

Within each prognostic category we must define the probability that any individual patient achieves a favorable outcome, either according to the traditional definition or the sliding dichotomy definition. Initial values for these probabilities were based on the observed attainment of a favorable outcome (under either definition) for patients in the NABISH

dataset categorized according to each of the predictive models described in the previous section (see Table 5.2). Depending on which predictive model was referenced, the probability of a favorable outcome according to the sliding dichotomy varied from 39% to 62% for patients with the 'worst' prognosis, 15% to 27% for patients with a 'moderate' prognosis, and 38% to 67% for patients with the 'best' prognosis. The distribution of probabilities looks counterintuitive at first, as those patients categorized with the worst prognosis appear to have the highest probability of achieving a favorable outcome. However, this is also the patient category for whom the definition of 'favorable' has been moved the lowest on the sliding dichotomy scale (see Figure 1.1). The probability of a traditional favorable outcome in the 'worst' prognostic group varied from 15% to 34%, 'moderate' varied from 15% to 27%, and 'best' from 45% to 67%.

These provided the starting probability values for simulated control patients; the next step was to choose the size of the simulated treatment effect. Some clinicians suggest that an improvement as small as 2% between treatment and control patients could translate into a highly successful treatment [Narayan et al., 2002], so the probability of a favorable outcome (either traditionally defined or based on the sliding dichotomy) was increased by two to ten percentage points (the current standard definition for a 'successful' treatment) in the simulated treatment patients. For example, one round of simulations to detect a two percentage point increase would use the values in Table 5.3.

Table 5.3: Simulated Probabilities of Favorable Outcome by Prognostic Group, Treatment Assignment, and Definition of Favorable

	Sliding			Traditional	
Prognostic Group	Treatment	Control		Treatment	Control
Worst	41%	39%		17%	15%
Moderate	17%	15%		17%	15%
Best	40%	38%		47%	45%

More generally, outcomes were simulated as

$$favorable \sim Bin(p_{gdz})$$

with a different probability for each prognostic group g, definition of favorability d, and

treatment group z.

Of course, this also assumes a constant treatment effect across prognostic groups, something else that many researchers believe is not reflective of reality. The simulations varied from a constant treatment effect across all prognostic groups to a treatment that only benefitted one prognostic group at a time.

This also interacted with another assumption - the distribution of patients into the prognostic categories. Initially patients were simulated to be evenly distributed into three prognostic categories. This distribution was eventually varied to more realistically reflect the patient population that might be available for a phase III clinical trial (majority 'worst' category, since often severely injured patients are recruited for trials); alternatively, simulations were also conducted with a simulated majority 'moderate' prognosis with fewer patients in either the 'best' or 'worst' categories to reflect Machado's [1999] suggestion to recruit primarily patients with an intermediate prognosis. Additionally, for simulations with a treatment effect in only one prognostic group, the proportion of patients in the prognostic group experiencing a treatment effect was increased. For example, one simulation scheme would implement a treatment effect only among those in the 'worst' prognostic category, but with patients still evenly distributed among the three categories. Subsequent simulations would then increase the proportion of patients in the 'worst' category to 40% and 80%, in an attempt to determine if a known treatment effect only affected a subset of the patient population, and clinicians were able to conduct targeted recruitment of patients in that particular prognostic group, could a higher powered clinical trial be achieved? And if so, which definition of favorable outcome would be most advantageous? Of course, this is an extreme example, since if the treatment effect only affects one prognostic group, that group is the only one contributing to the power of the study.

Since sample size is also a critical component of power calculations, the above described simulations were conducted with individual simulated samples of size 400. This was reflective of the approximate size of the original NABISH dataset and also falls well within Perel's [2006] systematic review of prognostic models for TBI research in which he found a median sample size of 319 patients and 75% of studies included less than 500 patients.

Lastly, prognostic groups were simulated such that there was an even 50%/50% split between treatment and control patients within each prognostic category, though in future analyses this too could be varied.

Whether or not a significant treatment effect was identified using each method was determined two different ways - logistic regression models were fit with favorable (either traditional ('tradfav') or sliding dichotomy ('fav')) as the outcome and treatment as the only covariate. For the sliding dichotomy definition of favorability, prognostic group was also included in the logistic regression model, since the point of the sliding dichotomy approach is to take prognostic group into account. The two models were

$$logit(P(\text{tradfav} = 1)) = \beta_0 + \beta_1 \text{treatment} \tag{5.8}$$

and

$$logit(P(\text{fav} = 1)) = \beta_0 + \beta_1 \text{treatment} + \beta_2 \text{prognostic group} \tag{5.9}$$

This slightly favors the sliding dichotomy approach, since Choi [1998] and Hernandez [2004] show that including a significant covariate in analyses with a dichotomous outcome increases power. This is a simplified version of this problem, since presumably final analyses to detect a potentially significant treatment effect would also adjust for other covariates.

Datasets with the above characteristics were simulated 1,000 times and the two logistic regression models were fit to each simulated dataset. The odds ratios and corresponding confidence intervals and p-values for the treatment coefficient (i.e., $\hat{\beta}_1$) were recorded from each model and the number of 'significant' treatment effects using each method was counted across simulations. Significance was defined as either an odds ratio confidence interval that excludes one or a p-value of less than 0.05. The proportion of significant results out of all 1,000 simulations indicated the power of each method to detect a potentially significant treatment effect of varying size.

5.5 Results - Sliding Dichotomy Power Analysis

A total of 295 simulations were conducted, varying probabilities of favorable outcome, distribution of patients into prognostic categories, and treatment effect size across prognostic groups. Since the four prognostic models used to set initial values in the simulations resulted in a range of probabilities of favorable outcomes, simulations somewhat arbitrarily increased probabilities by 5 percentage points across the suggested ranges (i.e., for the traditional definition of a favorable outcome, probabilities within the 'best' group included 45%, 50%, 55%, 60%, 65%, and 67%). After a clear pattern emerged in the simulation results, it was determined that finer variations in predicted probabilities were probably not necessary.

The traditional favorable outcome resulted in higher power in 196 simulations (66.4%), the sliding dichotomy produced higher power in 95 simulations (32.2%) and the two methods tied in 4 simulations (1.4%) (see Table 5.4).

Simulations also included examples of a treatment effect in only one prognostic group. When broken down this way, the traditional definition of a favorable outcome still clearly provides more power. Of the 90 simulations with a treatment effect only in the 'worst' prognostic group, 65 simulations (72.2%) had higher power with a traditionally defined outcome, 24 (26.7%) with a sliding dichotomy outcome, and 1 (1%) was tied. When treatment effect was simulated to only affect those in the 'moderate' prognostic group 58 simulations (64.4%) had higher power with a traditionally defined outcome versus 30 (33.3%) with a sliding dichotomy outcome, and 2 (2.2%) were tied. The sliding dichotomy approach appears to work best when treatment effect is restricted to those in the 'best' prognostic category, however traditional outcomes still more often resulted in higher power - 51 (56.7%) versus 38 (42.2)%, with one tie. Lastly, when treatment effect was constant across the prognostic groups (25 total simulations) 21 (84%) had higher power with a traditional outcome versus 4 (16%) with a sliding dichotomy outcome. These results are summarized in Table 5.4. It is also important to note that neither method ever achieved 80% power in any simulation scheme.

Again, at first these results appear to be counterintuitive, since the sliding dichotomy has

Table 5.4: Simulation Results Comparing Power

	% of Simulations Achieving Higher Power		
Simulation Scheme	**Traditional**	**Sliding**	**Tied**
Overall (n = 295)	66.4%	32.2%	1.4%
Constant (n = 25)	84%	16%	0%
'Worst' (n = 90)	72.2%	26.7%	1%
'Moderate' (n = 90)	64.4%	33.3%	2.2%
'Best' (n = 90)	56.7%	42.2%	1.1%

the potential to make the most gains in the 'worst' prognostic category by lowering the threshold for defining a favorable outcome. However, by lowering the threshold, the probability of a favorable outcome is increased, for both control and treatment patients, regardless of treatment effect. Traditional sample size calculations for the difference of two proportions show that the largest sample size (holding all else constant) is required when both proportions approach 0.5. By increasing the probability of a favorable outcome, the sliding dichotomy definition results in probabilities closer to 0.5 and thus requires a larger sample size/produces lower power than the traditional definition.

Although these simulations are by no means exhaustive, the clear pattern they indicate coincides with basic properties of sample size calculations for differences of proportions. To generalize these results to scenarios beyond those covered in the above simulations, our next step was to attempt a closed-form solution for sample size for both methods.

The most simplified version of the traditional favorable outcome is simply a comparison of two proportions:

$$n_i = \left(\frac{\sqrt{2\overline{pq}}Z_{1-(\alpha/2)} + \sqrt{p_t q_t + p_c q_c}Z_{1-\beta}}{ES}\right)^2 \tag{5.10}$$

Where p_t is the probability of a favorable outcome in the treatment group, p_c is the probability of a favorable outcome in the control group, Effect Size $= ES = |p_c - p_t|$, $\overline{p} = \frac{p_t + p_c}{2}$, and $\overline{q} = 1 - \overline{p}$. Using this equation one can estimate the sample size required for each group for a variety of values of p_t and p_c, once α, β, and ES have been chosen. For example, using the traditional cutpoint of a ten percentage point improvement, $\alpha = 0.05$, and $\beta = 0.1$, the table of sample sizes in Figure 5.1 could be calculated.

Currently, no closed-form solution exists for sample size calculations using a sliding di-

p1/p2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1		327.5							
0.2	327.5		483						
0.3		483		586.7					
0.4			586.7		638.6				
0.5				638.6		638.6			
0.6					638.6		586.7		
0.7						586.7		483	
0.8							483		327.5
0.9								327.5	

Figure 5.1: Sample Size Calculations for ES = 0.1, alpha = 0.05, beta = 0.1 with maximum n required per group highlighted

chotomy approach. However, one can exploit the fact that although in the case of three prognostic groups we are interested in six different proportions, it is actually three comparisons of two proportions (rather than directly comparing six proportions).

For either the traditional outcome or sliding dichotomy outcome, if we accept the assumption of a heterogeneous population with different probabilities of favorable outcome per prognostic category, the problem of under-powered clinical trials arises because we are not actually detecting a ten percentage point difference between treatment and control groups, but rather some other difference $p_t - p_c$ that is a weighted average of the differences within each prognostic group. Where p_t could be written as

$$p_t = (n_{t1}/N_t)(p_{t1}) + (n_{t2}/N_t)(p_{t2}) + (n_{t3}/N_t)(p_{t3}) \tag{5.11}$$

where n_{tk} is the number of treatment patients in prognostic group k, N_t is the total number of treatment patients, and p_{tk} is the probability of a favorable outcome for those who receive treatment in prognostic group k. A similar weighted average can be calculated for p_c. A few algebraic steps then show that for the simplified case where the proportions $\frac{n_{tk}}{N_t} = \frac{n_{ck}}{N_c} = w_k$, the difference in p_t and p_c that a study is actually powered to detect is:

$$p_t - p_c = w_1(p_{t1} - p_{c1}) + w_2(p_{t2} - p_{c2}) + w_3(p_{t3} - p_{c3}) \tag{5.12}$$

(for the simplified case of only three prognostic groups - an important next step would be to generalize to k prognostic groups). The traditional sample size calculation described in equation 5.10 can then be carried out for this new (and more accurate) estimate of the two proportions.

Although this formulation of the problem still requires an unreasonable number of assumptions to reach a closed-form solution, one can begin to get a handle on the problem by choosing the final difference $p_t - p_c$ that a clinical trial needs to be powered to detect, say, 0.1. The possible combinations of differences within each prognostic group that could lead to a difference of 0.1 could be written as

$$0.1 = w_1(p_{t1} - p_{c1}) + w_2(p_{t2} - p_{c2}) + w_3(p_{t3} - p_{c3}) = w_1 d_1 + w_2 d_2 + w_3 d_3 \tag{5.13}$$

where $d_k = p_{tk} - p_{ck}$. The question then becomes what combinations of values of $w_k * d_k$ sum to 0.1, for a given w_k? Of course, there is more than one solution to this problem, and d_k values may vary across prognostic groups according to whether favorable is defined traditionally or using the sliding dichotomy. But there are a finite number of solutions, and for each of those solutions there are a finite number of combinations of p_t and p_c.

We further simplify the problem by assuming the same distribution of patients into prognostic groups regardless of definition of outcome (reasonable, if comparing the two methods within the same patient population), i.e., w_k the same whether estimating p_t and p_c for traditional or sliding dichotomy approaches. Given the current definition of the sliding dichotomy approach, p_t and p_c for the 'worst' prognostic category (referred to as p_{t1} and p_{c1} for the remainder of this paper) will be higher for the sliding dichotomy defined outcome versus the traditional outcome (since the sliding dichotomy *includes* one lower GOS category as 'favorable' for those with the 'worst' prognosis; see Figure 5.2) and p_t and p_c for the 'best' prognostic category (referred to as p_{t3} and p_{c3} for the remainder of this paper) will be lower for the sliding dichotomy defined outcome versus the traditional outcome (since the sliding dichotomy *excludes* one GOS category in defining 'favorable' for those with the 'best' prognosis; see Figure 5.2). Lastly, assume that the 'moderate' prognostic group has

the same probability of achieving a favorable outcome under the sliding dichotomy model as under the traditional model. This is because favorability for this group is defined to be the same GOS categories under both schemes (see Figure 1.1 and pattern in Table 5.2). Therefore, any gain in power can be estimated by determining the trade-off between p_{t1} and p_{c1} using the two methods versus p_{t3} and p_{c3} using the two methods and the proportion of patients allocated to each category. The question then becomes what difference $(p_{t1} - p_{c1})$ versus $(p_{t3} - p_{c3})$ results in a smaller detectable overall effect size $ES = p_t - p_c$ without increasing n_i?

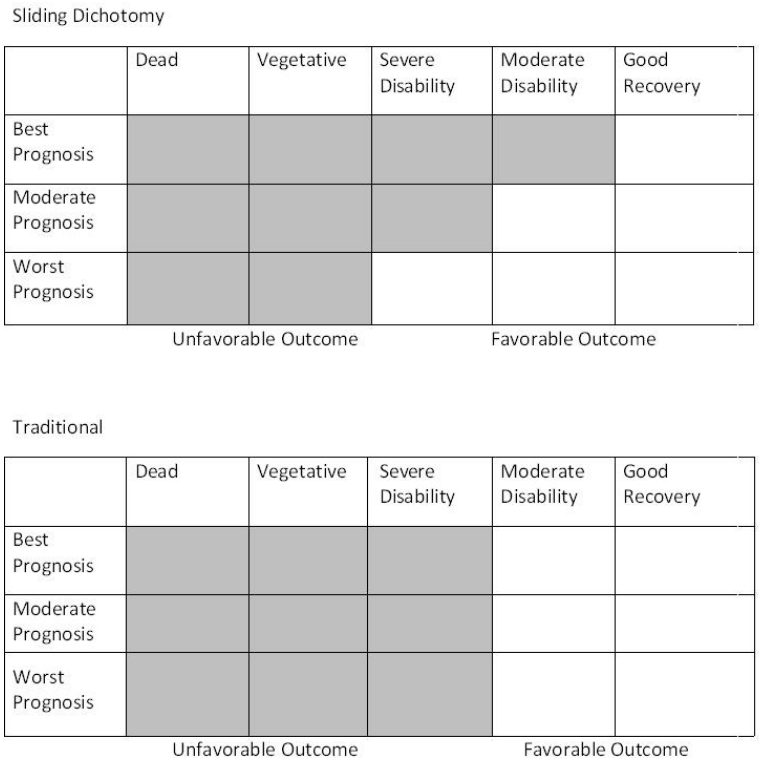

Figure 5.2: Graphical comparison of sliding dichotomy and traditional definitions of favorable outcomes as defined by GOS

Alternatively, what difference $(p_{t1} - p_{c1})$ versus $(p_{t3} - p_{c3})$ results in a smaller sample size for a given ES? Returning to equation 5.13, except this time only considering d_1 and d_3 (since d_2 does not contribute to any differences in sample size calculations between traditional and sliding dichotomy methods), and choosing a 65%/35% split of patients into 'worst' and 'best' prognostic groups, we can construct Figure 5.3 to summarize possible solutions to

equation 5.13. Highlighted cells in Figure 5.3 indicate possible values of d_1 and d_3 that would result in overall differences $p_t - p_c$ of 0.1 or greater. Assuming more 'room' for improvement in the 'worst' prognostic group, let's select $d_1 = 0.15$ and $d_3 = 0.05$ for an overall $p_t - p_c = w_1 d_1 + w_3 d_3 = 0.65 * 0.15 + 0.35 * 0.05 = 0.115$. Figure 5.4 shows possible combinations of $p_{1t} - p_{1c}$ resulting in a difference of 0.15 or greater. Similarly, Figure 5.5 shows possible combinations of $p_{3t} - p_{3c}$ resulting in a difference of 0.05 or greater. From Figures 5.4 and 5.5 we could select, for example, the probabilities in Table 5.5. Note that the probability of favorability in the worst category is higher for the sliding dichotomy method and the probability of favorability in the best category is higher for the traditional method, due to the definition of the sliding dichotomy method based on GOS outlined previously.

d1/d3	0	0.025	0.05	0.075	0.1	0.125	0.15	0.175	0.2	0.225	0.25	0.275	0.3	0.325	0.35	0.375	0.4	0.425
0	0	0.009	0.018	0.026	0.035	0.044	0.053	0.061	0.07	0.079	0.088	0.096	0.105	0.114	0.123	0.131	0.14	0.149
0.025	0.016	0.025	0.034	0.043	0.051	0.06	0.069	0.078	0.086	0.095	0.104	0.113	0.121	0.13	0.139	0.148	0.156	0.165
0.05	0.033	0.041	0.05	0.059	0.068	0.076	0.085	0.094	0.103	0.111	0.12	0.129	0.138	0.146	0.155	0.164	0.173	0.181
0.075	0.049	0.058	0.066	0.075	0.084	0.093	0.101	0.11	0.119	0.128	0.136	0.145	0.154	0.163	0.171	0.18	0.189	0.198
0.1	0.065	0.074	0.083	0.091	0.1	0.109	0.118	0.126	0.135	0.144	0.153	0.161	0.17	0.179	0.188	0.196	0.205	0.214
0.125	0.081	0.09	0.099	0.108	0.116	0.125	0.134	0.143	0.151	0.16	0.169	0.178	0.186	0.195	0.204	0.213	0.221	0.23
0.15	0.098	0.106	0.115	0.124	0.133	0.141	0.15	0.159	0.168	0.176	0.185	0.194	0.203	0.211	0.22	0.229	0.238	0.246
0.175	0.114	0.123	0.131	0.14	0.149	0.158	0.166	0.175	0.184	0.193	0.201	0.21	0.219	0.228	0.236	0.245	0.254	0.263
0.2	0.13	0.139	0.148	0.156	0.165	0.174	0.183	0.191	0.2	0.209	0.218	0.226	0.235	0.244	0.253	0.261	0.27	0.279
0.225	0.146	0.155	0.164	0.173	0.181	0.19	0.199	0.208	0.216	0.225	0.234	0.243	0.251	0.26	0.269	0.278	0.286	0.295
0.25	0.163	0.171	0.18	0.189	0.198	0.206	0.215	0.224	0.233	0.241	0.25	0.259	0.268	0.276	0.285	0.294	0.303	0.311
0.275	0.179	0.188	0.196	0.205	0.214	0.223	0.231	0.24	0.249	0.258	0.266	0.275	0.284	0.293	0.301	0.31	0.319	0.328
0.3	0.195	0.204	0.213	0.221	0.23	0.239	0.248	0.256	0.265	0.274	0.283	0.291	0.3	0.309	0.318	0.326	0.335	0.344
0.325	0.211	0.22	0.229	0.238	0.246	0.255	0.264	0.273	0.281	0.29	0.299	0.308	0.316	0.325	0.334	0.343	0.351	0.36
0.35	0.228	0.236	0.245	0.254	0.263	0.271	0.28	0.289	0.298	0.306	0.315	0.324	0.333	0.341	0.35	0.359	0.368	0.376
0.375	0.244	0.253	0.261	0.27	0.279	0.288	0.296	0.305	0.314	0.323	0.331	0.34	0.349	0.358	0.366	0.375	0.384	0.393
0.4	0.26	0.269	0.278	0.286	0.295	0.304	0.313	0.321	0.33	0.339	0.348	0.356	0.365	0.374	0.383	0.391	0.4	0.409
0.425	0.276	0.285	0.294	0.303	0.311	0.32	0.329	0.338	0.346	0.355	0.364	0.373	0.381	0.39	0.399	0.408	0.416	0.425
0.45	0.293	0.301	0.31	0.319	0.328	0.336	0.345	0.354	0.363	0.371	0.38	0.389	0.398	0.406	0.415	0.424	0.433	0.441
0.475	0.309	0.318	0.326	0.335	0.344	0.353	0.361	0.37	0.379	0.388	0.396	0.405	0.414	0.423	0.431	0.44	0.449	0.458
0.5	0.325	0.334	0.343	0.351	0.36	0.369	0.378	0.386	0.395	0.404	0.413	0.421	0.43	0.439	0.448	0.456	0.465	0.474

Figure 5.3: Possible values of d_1 and d_3 combining with 65% patients in 'worst' prognostic group and 35% patients in 'best' prognostic group to generate $ES \geq 0.1$ (shaded region)

Table 5.5: Possible Probabilities of Favorability Resulting in Overall $ES = 0.115$

	Sliding			Traditional	
Prognostic Group	Treatment	Control		Treatment	Control
Worst (p_{1zd})	35%	20%		25%	10%
Best (p_{3zd})	15%	10%		20%	15%
Overall (p_{zd})	28%	16.5%		23.3%	11.8%

ES for both methods is 0.115 but the sample size required for the sliding dichotomy method

	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
0.1	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85
0.15	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
0.2	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75
0.25	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7
0.3	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65
0.35	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6
0.4	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55
0.45	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.5	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
0.55	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.6	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35
0.65	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3
0.7	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25
0.75	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2
0.8	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15
0.85	-0.75	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1
0.9	-0.8	-0.75	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05
0.95	-0.85	-0.8	-0.75	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0

Figure 5.4: Possible values of p_{1t} and p_{1c} resulting in difference of 0.15 or greater (shaded region)

	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
0.1	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85
0.15	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
0.2	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75
0.25	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7
0.3	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65
0.35	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6
0.4	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55
0.45	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.5	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
0.55	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.6	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35
0.65	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25	0.3
0.7	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2	0.25
0.75	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15	0.2
0.8	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1	0.15
0.85	-0.75	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05	0.1
0.9	-0.8	-0.75	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0	0.05
0.95	-0.85	-0.8	-0.75	-0.7	-0.65	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.1	-0.05	0

Figure 5.5: Possible values of p_{3t} and p_{3c} resulting in difference of 0.05 or greater (shaded region)

is 336 versus 280 for the traditional method (assuming $\alpha = 0.05$ and 80% power).

$$n_i = \left(\frac{\sqrt{2 * 0.22 * 0.78} * 1.96 + \sqrt{(0.28 * 0.72) + (0.165 * 0.835)} * 1.64}{0.115}\right)^2 = 336$$

$$n_i = \left(\frac{\sqrt{2 * 0.0.18 * 0.82} * 1.96 + \sqrt{(0.233 * 0.767) + (0.118 * 0.882)} * 1.64}{0.115}\right)^2 = 280$$

This is because the overall p_t and p_c for the sliding dichotomy method are greater (and more specifically, closer to 0.5) than the overall values for the traditional method, and n_i is maximized when p_t and p_c are both closest to 0.5 (see Figure 5.1). Although we have been

using equation 5.12, we must return to equation 5.11 to calculate the individual overall p_t and p_c values needed in the sample size equation 5.10 used above. Therefore, for the sliding dichotomy, $p_t = (0.35*0.15)+(0.65*0.35) = 0.28$ and $p_c = (0.35*0.1)+(0.65*0.2) = 0.165$. For the traditional method, $p_t = (0.35*0.2)+(0.65*0.25) = 0.233$ and $p_c = (0.35*0.15)+(0.65*0.1) = 0.118$

Within this example, as shown in Figure 5.6, for the sliding dichotomy approach to result in a smaller sample size, the probability of a favorable outcome would either have to far surpass 50% within one prognostic category (say, $p_{1t} = 0.9$ and $p_{1c} = 0.75$) or the difference between probabilities of a favorable outcome would have to be quite large between the two methods (see Figure 5.8). The former seems unreasonable given the results of past clinical trials and the latter seems unreasonable since currently the two methods only differ by one GOS category (see Figure 5.2).

Figure 5.6 shows the difference in required sample size for the sliding dichotomy and traditional methods as probabilities of a favorable outcome are varied among the 'worst' and 'best' prognostic groups. Several assumptions are included in this figure:

- A detectable treatment effect of 0.15 among the 'worst' prognostic group
- A detectable treatment effect of 0.05 among the 'best' prognostic group
- $\alpha = 0.05$ and power = 80%
- 65% of patients in 'worst' prognostic group
- 35% of patients in 'best' prognostic group
- The probability of a traditionally defined favorable outcome in the 'worst' prognostic group is 0.1 less than the probability of a favorable outcome under the sliding dichotomy definition
- The probability of a traditionally defined favorable outcome in the 'best' prognostic group is 0.05 more than the probability of a favorable outcome under the sliding dichotomy definition.

The diagonal line labeled '0' indicates that the two methods require the same sample size when the probability of a favorable outcome is approximately 70% among the 'worst' prognostic group (combined with any probability in the 'best' group) or when the probability of a favorable outcome is over 80% among the 'best' prognostic group (combined with any probability in the 'worst' group). For higher combinations of probabilities the sliding dichotomy method results in a smaller required sample size (indicated by the negative difference labeled on the lines in the upper righthand portion of Figure 5.6) and for lower combinations of probabilities the traditional method results in a smaller required sample size.

To confirm these findings, we could repeat the simulations described in section 5.4 with values indicated by Figure 5.6. Specifically, Figure 5.6 indicates that the sliding dichotomy should require a smaller sample size when the probability of a favorable outcome is greater than 70% among those in the worst prognostic category. Repeating simulations with $p_{c1} = 0.8$ under the sliding dichotomy definition, and $p_{c1} = 0.7$ under the traditional definition indeed confirms higher power achieved under the sliding dichotomy definition (where power is estimated as defined in section 5.4). If we do not adjust for prognostic category in either model (comparing results from equation 5.8 for both definitions of favorability) we find that the sliding dichotomy method results in slightly less than 50% power, whereas the traditional definition results in slightly less than 40% power. If we adjust for prognostic group (comparing results from equation 5.9 for both definitions of favorability) we find that the sliding dichotomy definition results in almost 65% power whereas the traditional definition results in slightly more than 45% power (both were underpowered overall, due to a relatively small simulated sample size). As one might expect, adjusting for prognostic group favors the sliding dichotomy definition slightly, but failing to adjust for it in both models still confirms findings that for large probabilities of a favorable outcome the sliding dichotomy definition results in higher power to detect a significant treatment effect. Conversely, and again as indicated in Figure 5.6, if $p_{c1} = 0.3$ under the sliding dichotomy definition, and $p_{c1} = 0.2$ under the traditional definition, the traditional method results in higher power, regardless of adjustment for prognostic group.

As listed above, these results rely on several assumptions. Varying them one at a time produces both similar and different patterns. A similar pattern to that seen in Figure 5.6 holds if the distribution of patients into prognostic groups is flipped - 35% in the 'worst' category and 65% in the 'best', however in this case the differences in sample sizes are much smaller, since both methods are simulated to have more similar probabilities of a favorable outcome among the 'best' category and the assumed treatment effect size is smaller in the 'best' category (see Figure 5.7). Varying these assumptions as well can produce different patterns. For example, leaving the patient distribution with more patients in the 'best' category, allowing for a larger treatment effect within the 'best' category ($ES_3 = 0.1$), and a larger difference in the probability of a favorable outcome in the 'best' category as defined by the sliding dichotomy versus the traditional definition (delta = 0.15 instead of 0.05 above), flips the observable pattern in the difference of required sample sizes (see Figure 5.8). For a study population fitting this set of assumptions the traditional method would require a smaller sample size at higher combinations of probabilities of a favorable outcome and the sliding dichotomy would require a smaller sample size at lower combinations of probabilities of a favorable outcome.

The observed data in our motivating example fit the assumptions and patterns of lower combinations of probabilities in Figures 5.6 and 5.7, so we are reluctant to conclude that the sliding dichotomy method will result in higher-powered clinical trials. However, the figures presented here are produced by a program written in R (see appendix), and researchers could certainly enter their own values for the assumptions listed above and draw their own conclusions regarding the benefits of one method over the other for a specific set of data.

5.6 Discussion

5.6.1 Power and Sample Size

The probabilities used for both definitions of a favorable outcome in the above simulations are slightly lower than those assumed by Machado [1999] in his simulations. To detect

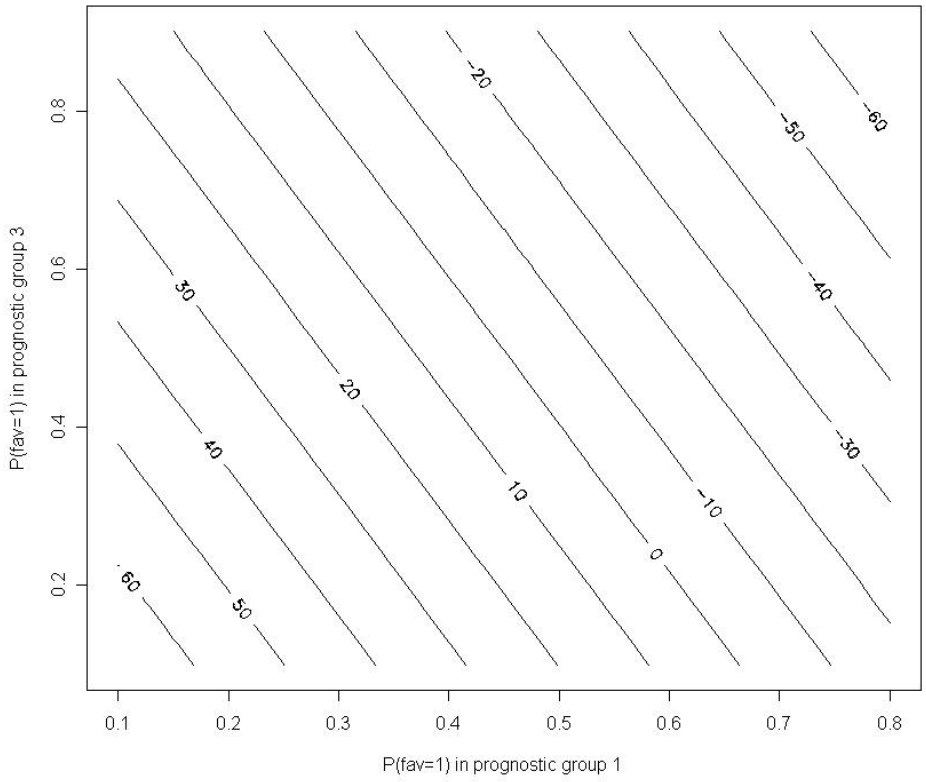

Figure 5.6: Difference in sample size required by sliding dichotomy vs. traditional methods to detect an overall treatment effect of 0.115 as probability of favorable outcome varies in 'worst' and 'best' prognostic groups; 65% of patients in 'worst' category, 35% in 'best'

an overall treatment effect corresponding to a ten percentage point increase in favorable outcomes among treatment versus control patients Machado simulated odds ratios that translated to a probability of a favorable outcome of 62% among treated patients in the 'intermediate' group versus 48% among control patients and 97% among treated patients in the 'best' prognostic category versus 95% among control patients. Although Machado does not provide enough information to duplicate his calculations, one possibility is a distribution of approximately 70% of patients into the 'intermediate' group and 30% into the 'best' group, which would result in a difference $p_t - p_c = 0.1$, with the individual prognostic group probabilities reported by Machado - $p_{1t} = 0.97$, $p_{1c} = 0.95$, $p_{2t} = 0.62$, and $p_{2c} = 0.48$. This would require a sample size of 636 (per treatment arm) to achieve 80% power (Machado [1999] alleges 90% power to detect this treatment effect with 500 patients per arm).

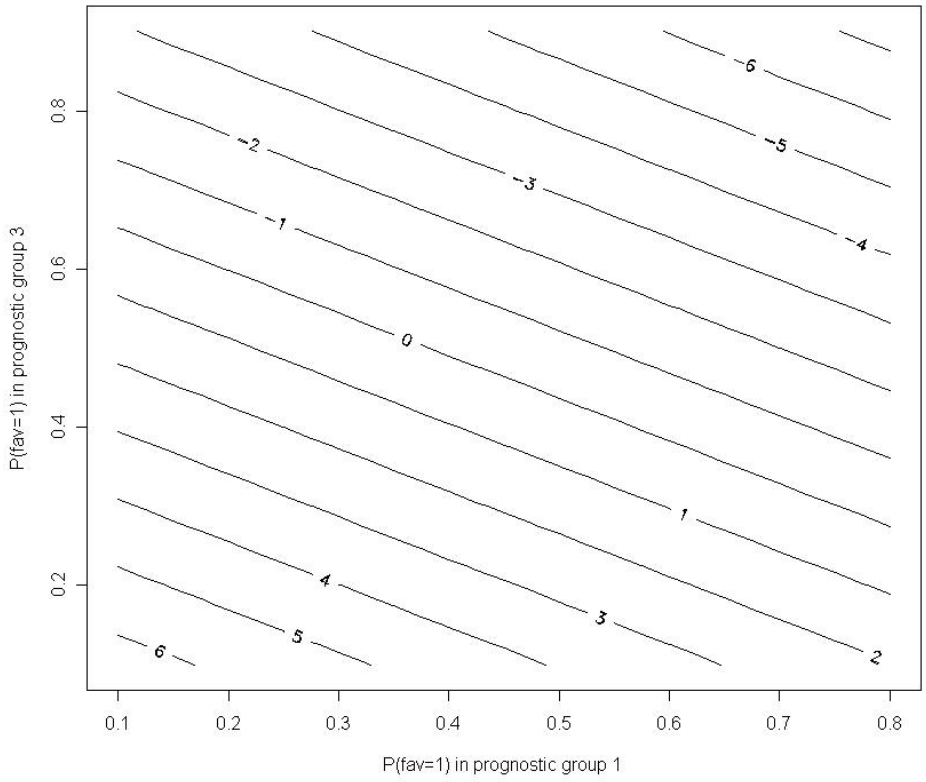

Figure 5.7: Difference in sample size required by sliding dichotomy vs. traditional methods to detect an overall treatment effect of 0.085 as probability of favorable outcome varies in 'worst' and 'best' prognostic groups; 35% of patients in 'worst' category, 65% in 'best'

Similarly, Mendelow [2005] reports that approximately 60% of all TBI patients achieve a favorable outcome across numerous studies, which is again a higher proportion than that observed in the NABISH dataset. If it is indeed true that the majority of TBI patients have a higher probability of a favorable outcome (regardless of treatment assignment) then perhaps the elevated probabilities needed to achieve gains in power via the sliding dichotomy method are reasonable (see Figures 5.6, 5.7, and 5.8).

A remaining complication to the existing literature on the sliding dichotomy approach and predictive models in general is the conflation of models designed to aid clinicians in assigning patients to treatment regimens and models designed to improve the efficiency of clinical trial design. The models discussed in this chapter and the recommendations based

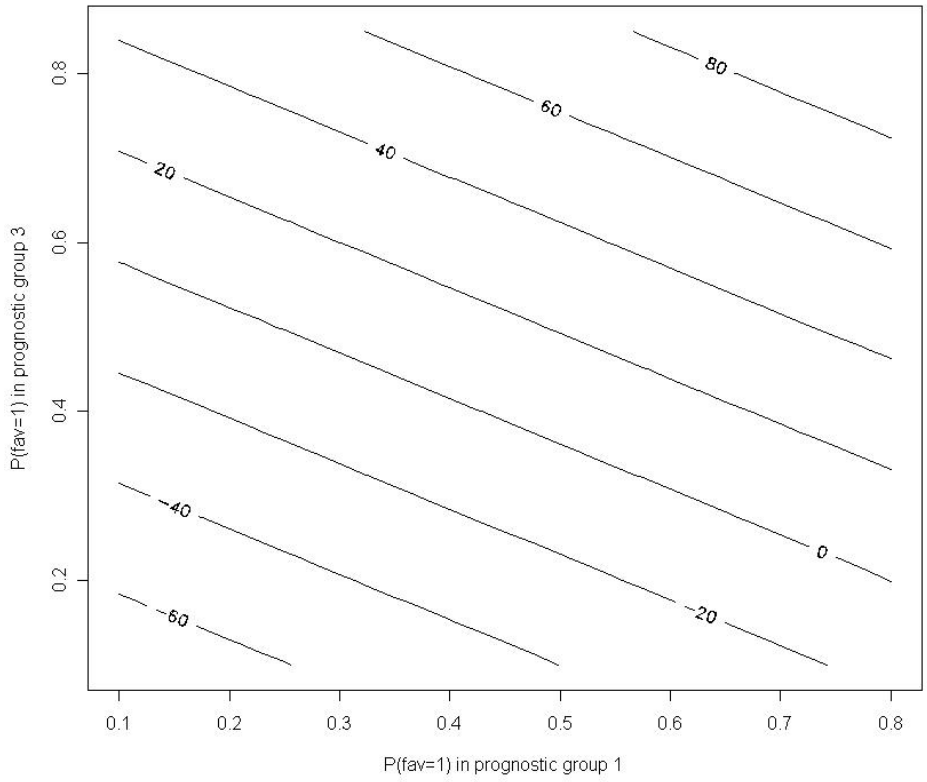

Figure 5.8: Difference in sample size required by sliding dichotomy vs. traditional methods to detect an overall treatment effect of 0.12 as probability of favorable outcome varies in 'worst' and 'best' prognostic groups; 35% of patients in 'worst' category, 65% in 'best'

on the simulations developed in this chapter are aimed at the latter problem. In particular, the Hansen prognostic score approach relies on fully collected data (albeit only from control patients) and therefore cannot be used to assist clinicians in estimating the prognosis of a specific patient prior to the completion of the study. Rather the predicted probabilities from models based on Hansen's approach can be used to categorize patients post data collection but prior to data analysis in hopes of closely matching the planned power of a study (assuming the study was designed with a sliding dichotomy approach). Inclusion and exclusion criteria for any specific clinical trial will still have to rely on predictive models developed in the existing literature.

5.6.2 Traditional versus Alternative Predictive Models

In much the same way that the propensity score $e(x)$ in chapter three enables multiple checks of covariate balance without revealing the primary outcome of interest, the development of a prognostic score $\Psi(X)$ enables checks of 'prognostic' balance without revealing any potential treatment effect. We believe that estimating predicted probabilities of a favorable outcome based on the subpopulation of control patients is a better general method than the models suggested by Hukkelhoven [2005] and Murray [2005] (or others in the literature). Although predictive models that have been internally and externally validated are a valuable addition to the literature, as pointed out in section 5.3 there are a variety of challenges associated with using predictive models developed on external datasets - failure to collect data on the same covariates, differing distributions among common covariates, etc. A general method for developing a new predictive model, tailored to each dataset, such as the new application of Hansen [2008] and Peters-Belson [1995] to ideas used in this chapter, should result in more accurate predictive probabilities (the Hansen modeling approach correctly identified nearly twice as many patients with favorable outcomes as either Hukkelhoven or Murray models). However, as noted in the previous section, this applies only to predictive models intended to improve final statistical analyses - predictive models intended to provide treatment guidance to clinicians would, obviously, have to be developed on external datasets and applied to new studies.

It is also worth noting that the prognostic scoring method is still susceptible to vast differences among treatment and control groups. In other words, developing a predictive model based on the control patients of a clinical trial when the randomization scheme has severely failed in some way could potentially be worse than applying a predictive model from the literature. Hansen's [2008] simulations show that "... when comparison groups differ substantially on X, adjustment based on same-sample estimation of prognostic scores can be much worse than no adjustment at all."

5.7 Conclusion

Although there do appear to be situations in which the sliding dichotomy approach would produce a higher powered clinical trial, the results of these simulations indicate that the majority of the time the traditional approach is higher powered. Additionally, for the sliding dichotomy approach to provide higher power, the probability of a favorable outcome would have to be quite high (greater than 0.5) and/or the change in overall probability of a favorable outcome as defined by the sliding dichotomy versus the traditional definition would have to be quite large (more than one would reasonably expect across one GOS category).

The formal power analysis presented here is especially needed in the TBI field since as recently as last fall scientists convened a workshop to '... outline the steps needed to develop a reliable, efficient and valid classification system.' [Saatman et al., 2008] Of course, this is a laudable goal when such a classification system is aimed at enabling physicians to make therapeutic decisions. But as mentioned in the previous section, often these goals are conflated with analytical goals, and as this chapter has shown that the sliding dichotomy approach is unlikely to produce significant gains in power, it seems wasteful to continue to contribute resources to developing better predictive models, if the goal of identifying 'better' definitions of favorable outcomes does not in fact improve the design of clinical trials.

It is possible that using predictive models to better map the probability of a favorable outcome (however it may be defined) within prognostic groups could help to achieve more efficient clinical trials by more closely approximating the true treatment effect $p_t - p_c$ as the weighted average of a series of differences as outlined in section 5.5. However, assigning patients to prognostic groups and estimating probabilities of a favorable outcome within those groups requires a nontrivial amount of guesswork and assumptions. Therefore more potential gains in efficiency can probably be found in the work of Choi [1998], Hernandez [2004], and others tackling the problem of adjusting for covariates when modeling dichotomous outcomes.

5.8 Future Work

Choi [1998] formalizes the problem of adjusting for a covariate in logistic regression with a dichotomous outcome. In linear regression problems the reduction of residual variance resulting from covariate adjustment plays a key role in estimations of sample size reduction. However, with a dichotomous outcome it is challenging to determine the appropriate measure of variability. Choi [1998] suggests three different potential measures of variance reduction, which he calls R (involving sums of squares, maximized log-likelihood values, and prediction rates) and suggests that the "...required sample size [may] be simply approximated by multiplying the usual asymptotic sample size by $(1 - R)$."

It appears that the most potential for gains in power/reductions in sample size can be achieved by covariate adjustment, so more research should focus on Choi's [1998] methods and in particular in examining if there is a difference in estimations of R if adjusting for traditional covariates or some combination of covariates summarized in the prognostic score $\Psi(X)$.

Chapter 6

Conclusions

Shadish et al [2008] warn that "...the practice of propensity score analysis in applied research may be yielding adjustments of unknown or highly variable accuracy. For a method as new as propensity score analysis, this is not surprising, and points to the need for more clarity about best propensity score practice." Currently, the literature on propensity scores includes comparisons of different parametric structures for both the propensity score and outcome models, different propensity score adjustment methods, and estimates of the sensitivity of propensity score analyses to potentially omitted variables. The majority of this research, and particularly the majority of theoretical work on propensity scores, assumes normally distributed covariates. In contrast, we chose to focus on dichotomous confounders, since this type of covariate is often neglected in the literature and yet often occurs in applied research, especially in public health. Additionally, dichotomous covariates present unique analytical challenges to the traditional propensity score rules of thumb regarding successful balancing techniques.

Specifically, we chose to compare the performance of propensity score regression adjustment to traditional regression adjustment in the presence of varying degrees of imbalance due to dichotomous covariates by conducting two different types of simulations - one called a 'pseudo-simulation' and one a 'full' simulation. The first involved re-sampling an existing dataset to create new datasets with more and less similiar confounder distributions. Al-

though this 'pseduo-simulation' provided some insight into the performance of propensity score and traditional regression adjustment methods, it did not provide a way to know the true treatment effect, and which method was providing a more accurate estimate of this effect. Therefore a 'full' simulation was conducted using datasets composed of variables with known distributions and a known treatment effect size. This latter set of simulations confirmed that propensity score and traditional regression adjustment methods suffer from the same quantity of bias in the presence of dichotomous confounders. However, the traditional method provides better coverage probabilities, so it should be preferred, even in the presence of high levels of confounding. Lastly, this set of simulations contradicted Drake's [1993] findings, based on normally distributed confounders, that propensity score regression adjustment has decreasing levels of bias as the association between confounders and treatment assignment increases.

Although propensity score regression adjustment analyses may still offer an alternative in the presence of many confounders and a small sample size, this is the only situation, with dichotomous confounders, that these simulations indicate a propensity score regression adjustment method would be preferred over traditional regression adjustment. This represents an important advancement in propensity score research, as many applied analyses are currently implementing propensity score regression adjustments with the false belief that propensity scores represent an improvement over traditional regression adjustment methods, or at the very least provide a confirmation of traditional regression adjustment results. In particular, many studies claim that propensity score regression adjustment analyses provide a more conservative estimate of treatment effect, based on the lower frequency of statistically significant findings, when in fact the consistently lower coverage probabilities presented in chapter three indicate that propensity score regression adjustment analyses may in fact be simply missing the true treatment effect.

Chapter four presented a unique application of principle stratification in the presence of missing values for the covariate used to define the principle strata. Up to this point all examples of principle stratification in the literature assume that all missing data are contained in the outcome variable, and that the post-treatment covariate used to define the

principle strata is known completely. Clearly this assumption will be frequently violated in many applied research problems, so we presented a sensitivity analysis to bound causal effect estimates according to assumptions about both missing outcome data and missing post-treatment covariate data. Our findings indicate that crude estimates about the latter source of missing data are sufficient, as assumptions about the structure of the principle strata themselves are much more influential over the bounds on the causal effect estimate. Hopefully these findings will advance the principle stratification literature by guiding researchers in prioritizing the many assumptions necessary to calculate a point estimate of causal effect in the presence of competing sources of missing data.

Additionally, the bounds proposed by the existing literature do not take into account two possible sources of variability in the estimate - both from the outcome variable itself and from the estimate of the proportion of observations belonging to each principle strata. We proposed an initial method for estimating these sources of variability to enable to calculation of confidence intervals around the bounds presented in chapter four, thus making it possible to draw conclusions regarding the statistical significance of the estimated causal effect.

Lastly, these bounds on the causal effect estimate draw on large sample theory, and we propose a Bayesian approach that will not only take multiple sources of missing data into account, but will potentially reflect the true distribution of the data more accurately than the large sample asymptotic assumptions.

In chapter five we tackled the question of whether or not the sliding dichotomy method is truly the analytical silver bullet that some in the traumatic brain injury (TBI) and stroke research fields desire it to be. Although it is conceptually appealing to more specifically tailor a patient's outcome classification into 'favorable' and 'unfavorable' based on his or her initial prognosis, such definitions based on current outcome scales only provide an improvement in power under very specific, and likely to be rare, circumstances. Additionally, estimating an individual patient's initial prognosis and accurately assigning an outcome 'goal' based on such a prognosis is both highly complicated and likely to be inaccurate. Our findings in chapter five indicate that such resources would be potentially better applied to alternative analytical techniques, such as including baseline covariates in final analyses and

designing studies with the assumption that final analyses will include such variables.

The power analyses presented in chapter five are the first formal comparison of the sliding dichotomy and traditional methods and advance the TBI literature by outlining the circumstances under which the sliding dichotomy method results in a higher powered trial. This occurs either when the probability of a favorable outcome within one prognostic group is much higher than 50% (say 90% among the treated and 75% among the control) or when a favorable outcome as defined by one cutpoint on the Glasgow Outcome Scale versus another results in dramatically different probabilities of achieving a favorable outcome.

Bibliography

Pater J.D. Andrews, Derek H. Sleeman, Patrick F.X. Statham, Andrew McQuatt, Vincent Corruble, Patricia A. Jones, Timothy P. Howells, and Carol S.A. MacMillan. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: A comparison between decision tree analysis and logistic regression. *Journal of Neurosurgery*, 97:326–336, 2002.

Peter C. Austin and Muhammad M. Mamdani. A comparison of propensity score methods: A case-study estimating the effectiveness of post-ami statin use. *Statistics in Medicine*, 25:2084–2106, 2006.

Peter C. Austin, Paul Grootendorst, and Geoffrey M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in Medicine*, 26:734–753, 2007.

D. Barer. Could stroke mega-trials be missing important treatment effects? 1998. poster presented at European Stroke Conference, Edinburgh.

William A. Belson. A technique for studying the effects of a television broadcast. *Applied Statistics*, 5:195–202, 1956.

Xun Chen, Minzhi Liu, and Ji Zhang. A note on postrandomization adjustment of covariates. *Drug Information Journal*, 39:373–383, 2005.

Jing Cheng and Dylan S. Small. Bounds on causal effects in three-arm trials with noncompliance. *Journal of the Royal Statistical Society B*, 68:815–836, 2006.

Sung C. Choi. Sample size in clinical trials with dichotomous endpoints: Use of covariables. *Journal of Biopharmaceutical Statistics*, 8:367–375, 1998.

Sung C. Choi, J.P. Muizelaar, T.Y. Barnes, Anthony Marmarou, D.M. Brooks, and Fiona B. Young. Prediction tree for severely head-injured patients. *Journal of Neurosurgery*, 75: 251–255, 1991.

Sung C. Choi, Guy L. Clifton, Anthony Marmarou, and Emmy R. Miller. Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. *Journal of Neurotrauma*, 19:17–22, 2002.

Guy L. Clifton, Emmy R. Miller, Sung C. Choi, Harvey S. Levin, Stephen McCauley, Kenneth R. Smith, J. Paul Muizelaar, Franklin C. Wagner, Donald W. Marion, Thomas G. Luerssen, Randall M. Chesnut, and Michael Schwartz. Lack of effect of induction of hypothermia after acute brain injury. *The New England Journal of Medicine*, 344:556–563, 2001.

W.G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24:295–313, 1968.

W.G. Cochran. The use of covariance in observational studies. *Applied Statistics*, 18: 270–275, 1969.

W.G. Cochran. *Observational Studies*. Ames: Iowa State University Press, 1972.

William G. Cochran and Donald B. Rubin. Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics: Series A*, 35:417–446, 1973.

Francis E. Cook and Lee Goldman. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *Journal of Clinical Epidemiology*, 42:317–324, 1989.

Carl Counsell and Martin Dennis. Systematic review of prognostic models in patients with actute stroke. *Cerebrovascular Diseases*, 12:159–170, 2001.

Ralph B. D'Agostino Jr. Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17:2265–2281, 1998.

A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41:5–7, 1979.

Rajeev H. Dehejia and Sadek Wahba. Propensity score matching methods for non-experimental causal studies. Technical Report 6829, National Bureau of Economic Research, 1998.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Staistical Association*, 94:1053–1062, 1999.

M. Soledad Depeda, Ray Boston, John T. Farrar, and Brian L. Strom. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158:280–287, 2003.

Paula Diehr, Donald L. Patrick, Susan Hedrick, Margaret Rothman, David Grembowski, Trivellore E. Raghunathan, and Shirley Beresford. Including deaths when measuring health status over time. *Medical Care*, 33:AS164–AS172, 1995.

Paula Diehr, Laura Lee Johnson, Donald L. Patrick, and Bruce Psaty. Methods for incorporating death into health-related variables in longitudinal studies. *Journal of Clinical Epidemiology*, 58:1115–1124, 2005.

Christiana Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49:1231–1236, 1993.

Carole Dufouil, Carol Brayne, and David Clayton. Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine*, 23:2215–2226, 2004.

Brian L. Egleston, Daniel O. Scharfstein, Ellen E. Freeman, and Sheila K. West. Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, 8:526–545, 2007.

R.A. Fisher. The causes of human variability. *Eugenics Review*, 10:213–220, 1918.

Constantine E. Frangakis and Donald B. Rubin. The defining role of 'principle stratification and effects' for comparing treatments adjusted for posttreatment variables: from treatment noncompliance to surrogate endpoints. *Biometrics*, 58:191–199, 2002a.

Constantine E. Frangakis and Donald B. Rubin. Principle stratification in causal inference. *Biometrics*, 58:21–29, 2002b.

Constantine E. Frangakis and Donald B. Rubin. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-complaince and subsequent missing outcomes. *Biometrika*, 86:365–379, 1999.

Constantine E. Frangakis, Donald B. Rubin, and Xiao-hua Zhou. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics*, 3:147–164, 2002.

Constantine E. Frangakis, Donald B. Rubin, Ming-Wen An, and Ellen MacKenzie. Principle stratification designs to estimate input data missing due to death. *Biometrics*, 63:641–662, 2007.

J.L. Gastwirth and S.W. Greenhouse. Biostatistical concepts and methods in the legal setting. *Statistics in Medicine*, 14:1641–1653, 1995.

Robert J. Glynn, Sebastian Schneeweiss, and Til Sturmer. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic and Clinical Pharmacology and Toxicology*, 98:253–259, 2006.

C. Goldin and C. Rouse. Orchestrating impartiality: The impact of 'blind' auditions on female musicians. *The American Economic Review*, 90:715–741, 2000.

J. Grogger and G. Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101:878–887, 2006.

Gary L. Grunkemeier, Nicola Payne, Ruyun Jin, and John R. Handy Jr. Propensity score

analysis of stroke after off-pump coronary artery bypass grafting. *Annals of Thoracic Surgery*, 74:301–305, 2002.

Xing Sam Gu and Paul R. Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2:405–420, 1993.

Ben B. Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99:609–616, 2004.

Ben B. Hansen. Matching with prognosis scores: A new method of adjustment for comparative studies. 2006. oral presentation at Winemiller 2006 Conference, University of Missouri.

Ben B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95:481–488, 2008.

Ofer Harel, Scott M. Hofer, Lesa Hoffman, and Nancy L. Pederson. Population inference with mortality and attrition in longitudinal studies on aging: a two-stage multiple imputation method. *Experimental Aging Research*, 33:187–203, 2007.

Douglas Hayden, Donna K. Pauler, and David Schoenfeld. An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*, 61:305–310, 2005.

Adrian V. Hernandez, Ewout W. Steyerberg, and J. Dik F. Habbema. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*, 57:454–460, 2004.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189, 2003.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236, 2007.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–970, 1986.

Chantal W.P.M. Hukkelhoven, Ewout W. Steyerberg, Dik F. Habbema, Elana Farace, Anthony Marmarou, Gordan Murray, Lawrence F. Marshall, and Andrew I.R. Maas. Predicting outcome after traumatic brain injury: Development and validation of a prognostic score based on admission characteristics. *Journal of Neurotrauma*, 22:1025–1039, 2005.

D. Hume. *An Enquiry Concerning Human Understanding*. 1748.

Kosuke Imai. Sharp bounds on the causal effects in randomized experiments with 'truncation-by-death'. *Statistics & Probability Letters*, 78:144–149, 2008.

Kosuke Imai and David A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99: 854–866, 2004.

Guido Imbens and Donald B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25:305–327, 1997.

Hui Jin and D. Rubin. Principle stratification for causal inference with extended partial compliance. *Journal of the American Statistical Assocation*, 103:101–111, 2008.

Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:523–539, 2007.

Joseph T. King, Patricia M. Carlier, and Donald W. Marion. Early glasgow outcome scale scores predict long-term functional outcome in patients with severe traumatic brain injury. *Journal of Neurotrauma*, 22:947–954, 2005.

Brenda Kurland, Laura Lee Johnson, and Paula Diehr. Longitudinal data with follow-up truncated by death: Finding a match between analysis method and research aims. Technical Report 319, University of Washington, 2007.

Brenda F. Kurland and Patrick J. Heagerty. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6:241–258, 2005.

Tobias Kurth, Alexander M. Walker, Robert J. Glynn, K. Arnold Chan, J. Michael Glaziano, Klaus Berger, and James M. Robins. Results of a multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163:262–270, 2006.

Dennis V. Lindley. Seeing and doing: the concept of causation. *International Statistical Review*, 70:191–214, 2002.

Thomas E. Love. Strategies for using propensity scores well. 2005. 6th International Conference for Health Policy Research.

Andrew I.R. Maas, Ewout W. Steyerberg, Gordon D. Murray, Ross Bullock, Alexander Baethmann, Lawrence F. Marshall, and Graham M Teasdale. Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? a pragmatic analysis and theoretical considerations. *Neurosurgery*, 44:1286–1298, 1999.

S.G. Machado, G.D. Murray, and G.M. Teasdale. Evaluation of designs for clinical trials of neuroprotective agents in head injury. *Journal of Neurotrauma*, 16:1131–1138, 1999.

Ellen J. MacKenzie, Frederick P. Rivara, Gregory J. Jurkovich, Avery B. Nathens, Katherine P. Frey, Brian L. Egleston, David S. Salkever, Sharada Weir, and Daniel O. Scharfstein. The national study on costs and outcomes of trauma. *The Journal of Trauma Injury, Infection, and Critical Care*, 63:S54–S67, 2007.

Ellen J. MacKenzie, Frederick P. Rivara, Gregory J. Jurkovich, Avery B. Nathens, Brian L. Egleston, David S. Salkever, Katherine P. Frey, and Daniel O. Scharfstein. The impact of trauma-center care on functional outcoems following major lower-limb trauma. *The Journal of Bone and Joint Surgery, Incorporated*, 90:101–109, 2008.

Anthony Marmarou, Juan Lu, Isabella Butcher, Gillian S. McHugh, Gordon D. Murray, Ewout W. Steyerberg, Nino A. Mushkudiani, Sung Choi, and Andrew I.R. Maas. Prog-

nostic value of the glasgow coma scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: An impact analysis. *Journal of Neurotrauma*, 24:270–280, 2007.

Yutaka Matsuyama and Satoshi Morita. Estimation of the average causal effect among subgroups defined by post-treatment variables. *Clinical Trials*, 3:1–9, 2006.

A. Mattei and F. Mealli. Application of the principal stratification approach to the faenza randomized experiment on breast self-examination. *Biometrics*, 63:437–446, 2007.

Sheena McConnell, Elizabeth A. Stuart, and Barbara Devaney. The truncation-by-death problem: What to do in an experimental evaluation when the outcome is not always defined. *Evaluation Review*, 32:157–186, 2008.

Gillian S. McHugh, Doortje C. Engel, Isabella Butcher, Ewout W. Steyerberg, Juan Lu, Nino Mushkudiani, Adrian V. Hernandez, Anthony Marmarou, Andrew I.R. Maas, and Gordon D. Murray. Prognostic value of secondary insults in traumatic brain injury: Results from the impact study. *Journal of Neurotrauma*, 24:287–293, 2007.

Fabrizia Mealli and Donald B. Rubin. Assumptions allowing the estimation of direct causal effects: Commentary on 'health, wealth, and wise?' test for direct causal paths between health and socioeconomic status. *Journal of Econometrics*, 112:79–87, 2003.

A. David Mendelow, Barbara A. Gregson, Helen M. Fernandes, Gordon D. Murray, Graham M. Teasdale, D. Terence Hope, Abbas Karimi, M. Donald M. Shaw, and David H. Barer. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the international surgical trial in intracerebral haemorrhage (stich): A randomized trial. *Lancet*, 365:387–397, 2005.

A.D. Mendelow, G.M. Teasdale, D. Barer, H.M. Fernandes, G.D. Murray, and B.A. Gregson. Outcome assignment in the international surgical trial of intracerebral haemorrhage. *Acta Neurochirurgica*, 145:679–681, 2003.

Peter Menzies. *Stanford Encyclopedia of Philosophy*, chapter Counterfactual Theories of Causation. Metaphysics Research Lab, CSLI, 2008.

Peter Menzies. Stanford encycopedia of philosophy: Counterfactual theories of causation, 2001. http://plato.stanford.edu/entries/causation-counterfactual, accessed 2/7/08.

Olli Miettinen. Stratification by a multivariate confounder score. *American Journal of Epidemiology*, 104:609–620, 1976.

K.K. Mukherjee, B.S. Sharma, S.M. Ramanathan, N. Khandelwal, and V.K. Kak. A mathematical outcome prediction model in severe head injury: A pilot study. *Neurology India*, 48:43–48, 2000.

Gordan D. Murray, David Barer, Sung Choi, Helen Fernandes, Barbara Gregson, Kennedy R. Lees, Andrew I.R. Maas, Anthony Marmarou, A. David Mendelow, Ewout W. Steyerberg, Gillian S. Taylor, Graham M. Teasdale, and Christopher J. Weir. Design and analysis of phase iii trials with ordered outcome scales: The concept of the sliding dichotomy. *Journal of Neurotrauma*, 22:511–517, 2005.

Gordan D. Murray, Isabella Butcher, Gillian S. McHugh, Juan Lu, Nino A. Mushkudiani, Andrew I.R. Maas, Anthony Marmarou, and Ewout W. Steyerberg. Multivariable prognostic analysis in traumatic brain injury: Results from the impact study. *Journal of Neurotrauma*, 24:329–337, 2007.

Raj K. Narayan, Mary Ellen Michel, and The Clinical trials In Head Injury Study Group. Clinical trials in head injury. *Journal of Neurotrauma*, 19:503–557, 2002.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Annals of Agricultural Sciences, translated in Statistical Science*, 5: 465–480, 1990.

James A. O'Malley and Sharon-Lise T. Normand. Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. *Biometrics*, 61:325–334, 2005.

Donna K. Pauler, Sheryl McCoy, and Carol Moinpour. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine*, 22:795–809, 2003.

Pablo Perel, Phil Edwards, Reinhard Wentz, and Ian Roberts. Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics and Decision Making*, 6, 2006.

Charles C. Peters. A method of matching groups for experiments with no loss of population. *The Journal of Educational Research*, 34:606–612, 1941.

A. Posner, Michael, Arlene S. Ash, Karen M. Freund, Mark A. Moskowitz, and Michael Shwartz. Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis. *Health Services and Outcomes Research Methodology*, 2:279–290, 2001.

MJ Reeves, J Gargano, J Broderick, M Frankel, KA LaBresh, L Schwamm, and CJ Moomaw. Patient- and hospital-level determinants of the quality of acute stroke care: The paul coverdell national acute stroke prototype registry. *Stroke*, 38:478, 2007.

James M. Robins. Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17:269–302, 1998.

James M. Robins, Donald Blevins, Grant Ritter, and Michael Wulfsohn. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of aids patients. *Epidemiology*, 3:319–336, 1992a.

James M. Robins, Steven D. Mark, and Whitney K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48: 479–495, 1992b.

James M. Robins, Miguel A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000.

Paul R. Rosenbaum. From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, 79:41–48, 1984.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983a.

Paul R. Rosenbaum and Donald B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society B*, 45:212–218, 1983b.

Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1984.

Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, pages =, 1985.

Donald B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2:169–188, 2001.

Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–332, 2005.

Donald B. Rubin. Causal inference through potential outcomes and principle stratification: Application to studies with 'censoring' due to death. *Statistical Science*, 21:299–309, 2006.

Donald B. Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29:185–203, 1973.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychiatry*, 66:688–701, 1974.

Donald B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2:1–26, 1977.

Donald B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74: 318–328, 1979.

Donald B. Rubin. *William G. Cochran's Contributions to the Design, Analysis, and Evaluation of Observational Studies*. New York: Wiley, 1984.

Donald B. Rubin. Comment: Which ifs have causal answers [comment on statistics and causal inference]. *Journal of the American Statistical Association*, 81:945–970, 1986.

Donald B. Rubin and Neal Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95:573–585, 2000.

Donald B. Rubin and Neal Thomas. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52:249–264, 1997.

Kathryn E. Saatman, Ann-Christine Duhaime, Ross Bullock, Andrew I.R. Maas, Alex Valadka, Geoffrey T. Manley, Workshop Scientific Team, and Advisory Panel Members. Classification of traumatic brain injury for targeted therapies. *Journal of Neurotrauma*, 25:719–738, 2008.

Jeffrey L. Saver and Banafsheh Yafeh. Confirmation of tpa treatment effect by baseline severity-adjusted end point reanalysis of the ninds-tpa stroke trials. *Stroke*, 38:414–416, 2007.

William R. Shadish, M.H. Clark, and Peter M. Steiner. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random to nonrandom assignment. 2008.

B. Shah, A. Laupacis, J. Hux, and P. Austin. Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *Journal of Clinical Epidemiology*, 58:550–559, 2005.

David G. Sherman, Richard P. Atkinson, Thomas Chippendale, Kenneth A. Levin, Ken Ng, Nancy Futrell, Chung Y. Hsu, and David E Levy. Intravenous ancrod for treatment of acute ischemic stroke: The stat study: A randomized controlled trial. *Journal of the American Medical Association*, 283:2395–2403, 2000.

D.F. Signorini, Pater J.D. Andrews, P.A. Jones, J.M. Wardlaw, and J.D. Miller. Predicting survival using simple clinical variables: a case study in traumatic brain injury. *Journal of Neurology, Neurosurgery, and Psychiatry*, 66:20–25, 1999.

T. Sturmer, M. Joshi, R. Glynn, J. Avorn, K. Rothman, and S. Schneeweiss. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59:437–447, 2006.

C. Weimar, I.R. Konig, K. Kraywinkel, A. Ziegler, and H.C. Diener. Age and national institutes of health stroke scale score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia. *Stroke*, 35:158–162, 2004.

Christian Weimar, Tony W. Ho, Zaza Katsarava, and Hans-Christoph Diener. Improving patient selection for clinical acute stroke trials. *Cerebrovascular Diseases*, 21:386–392, 2006.

C. Weir and R. Walley. Statistical evaluation of biomarkers as surrogate endpoints: A literature review. *Statistics in Medicine*, 25:183–203, 2006.

David W. Wright, Arthur L. Kellerman, Vicki S. Hertzberg, Pamela L. Clark, Michael Frankel, Felicia C. Goldstein, Jeffrey P. Salomone, L. Leon Dent, Odette A. Harris, Douglas S. Ander, Douglas W. Lowery, Manish M. Patel, Donald D. Denson, Angelita B. Gordon, Marlena M. Wald, Sanjay Gupta, Stuart Hoffman, and Donald G. Stein. Protect: A randomized clinical trial of progesteron for acute traumatic brain injury. *Annals of Emergency Medicine*, 49:391–404, 2007.

Fiona B. Young, Kennedy R. Lees, and Christopher J. Weir. Stengthening acute stroke trials through optimal use of disability end points. *Stroke*, 34:2676–2680, 2003.

Fiona B. Young, Kennedy R. Lees, and Christopher J. Weir. Improving trial power through use of prognostic-adjusted end points. *Stroke*, 36:597–601, 2005.

Junni L. Zhang and Donald B. Rubin. Estimation of causal effects via principle stratification

when some outcomes are truncated by 'death'. *Journal of Educational and Behavioral Statistics*, 28:353–368, 2003.

Junni L. Zhang, Donald B. Rubin, and Fabrizia Mealli. *Evaluating the Effects of Job Training Programs on Wages through Principle Stratification*, volume 21. Amsterdam: Elsevier, 2006.

Zhong Zhao. Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *The Review of Economics and Statistics*, 86:91–107, 2004.

Appendices

Appendix A

Chapter 3 - Propensity Score

A.1 Theoretical Derivations

From Rosenbaum and Rubin (1984):

$$P(\mathbf{x}, z|e) = P(\mathbf{x}|e)P(z|e)$$

where $\mathbf{x}$ is a vector of observed covariates, z is the treatment indicator, and e is the propensity score, $e = e(\mathbf{x}) = P(z = 1|\mathbf{x})$. Proof:

$$P(\mathbf{x}, z|e) = P(\mathbf{x}|e)P(z|\mathbf{x}, e)$$

$P(z|\mathbf{x}, e) = P(z|\mathbf{x})$ since e is a function of x

$P(z = 1|\mathbf{x}) = e$by definition

$P(z = 1|e) = E(z|e) = E\{E(z|\mathbf{x})|e\} = E(e|e) = e$

$\Rightarrow P(z = 1|\mathbf{x}) = P(z = 1|e) \Rightarrow P(z|\mathbf{x}) = P(z|e)$

$\therefore P(\mathbf{x}, z|e) = P(\mathbf{x}|e)P(z|\mathbf{x}) = P(\mathbf{x}|e)P(z|e)$ as needed

A.2 Complete Pseudo-Simulation Results

Table A.1: Simulation results - Afibrillation (10% among black patients) - Comparing traditional and propensity score regression adjustment

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	0.04	(-0.02, 0.09)
10%	0.03	(-0.02, 0.08)
15%	-0.003	(-0.03, 0.04)
20%	-0.01	(-0.07, 0.07)
25%	-0.01	(-0.08, 0.09)
30%	0.02	(-0.05, 0.10)
35%	0.01	(-0.08, 0.11)
40%	0.02	(-0.09, 0.12)
45%	0.02	(-0.12, 0.15)
50%	-0.004	(-0.17, 0.18)
55%	0.02	(-0.17, 0.21)
60%	0.01	(-0.19, 0.22)
65%	-0.06	(-0.22, 0.08)
70%	-0.09	(-0.24, 0.05)
75%	-0.05	(-0.24, 0.10)
80%	0.004	(-0.21, 0.20)
85%	0.02	(-0.21, 0.23)
90%	0.006	(-0.21, 0.22)
95%	0.03	(-0.15, 0.32)

Table A.2: Simulation results - Hyperlipidemia (21% among black patients) - Comparing traditional and propensity score regression adjustment

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	-0.005	(-0.06, 0.06)
10%	-0.02	(-0.08, 0.06)
15%	-0.04	(-0.09, 0.04)
20%	-0.04	(-0.10, 0.05)
25%	0.004	(-0.05, 0.08)
30%	0.008	(-0.08, 0.09)
35%	0.01	(-0.06, 0.10)
40%	0.03	(-0.11, 0.16)
45%	0.009	(-0.14, 0.15)
50%	0.005	(-0.14, 0.13)
55%	0.02	(-0.12, 0.15)
60%	0.04	(-0.12, 0.19)
65%	0.03	(-0.12, 0.17)
70%	0.04	(-0.12, 0.18)
75%	0.07	(-0.11, 0.22)
80%	0.11	(-0.10, 0.30)
85%	0.13	(-0.006, 0.30)
90%	**0.11**	**(0.01, 0.26)**
95%	0.12	(-0.08, 0.26)

Table A.3: Simulation results - CAD (19% among black patients) - Comparing traditional and propensity score regression adjustment

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	-0.007	(-0.05, 0.07)
10%	-0.03	(-0.08, 0.09)
15%	-0.03	(-0.09, 0.04)
20%	-0.01	(-0.08, 0.07)
25%	0.03	(-0.03, 0.11)
30%	0.04	(-0.03, 0.14)
35%	0.03	(-0.07, 0.14)
40%	0.03	(-0.06, 0.15)
45%	0.05	(-0.07, 0.16)
50%	0.06	(-0.05, 0.18)
55%	0.04	(-0.10, 0.19)
60%	0.02	(-0.14, 0.17)
65%	-0.02	(-0.19, 0.13)
70%	0.01	(-0.16, 0.16)
75%	0.05	(-0.13, 0.22)
80%	0.07	(-0.12, 0.24)
85%	0.06	(-0.08, 0.23)
90%	0.06	(-0.08, 0.23)
95%	0.24	(-0.02, 0.08)

Table A.4: Simulation results - Afibrillation (10% among black patients) - Comparing traditional regression adjustment to stratification by propensity score

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	-0.32	(-0.72, 0.21)
10%	-0.88	(-0.47, 0.40)
15%	-0.37	(-1.32, 0.19)
20%	-0.31	(-1.33, 0.58)
25%	-0.31	(-0.93, 0.33)
30%	**-0.54**	**(-0.99, -0.09)**
35%	**-0.52**	**(-0.95, -0.06)**
40%	-0.34	(-1.18, 0.29)
45%	-0.46	(-0.98, 0.04)
50%	**-0.62**	**(-1.03, -0.09)**
55%	**-0.71**	**(-1.10, -0.17)**
60%	-0.4	(-1.18, 0.18)
65%	-0.56	(-1.12, 0.10)
70%	**-0.48**	**(-0.76, -0.29)**
75%	**-0.55**	**(-0.85, -0.32)**
80%	**-0.69**	**(-0.99, -0.49)**
85%	**-0.63**	**(-0.80, -0.64)**
90%	**-0.84**	**(-0.91, -0.81)**
95%	**-0.52**	**(-0.52, -0.42)**

Table A.5: Simulation results - Hyperlipidemia (21% among black patients) - Comparing traditional regression adjustment to stratification by propensity score

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	-0.22	(-0.30, 0.48)
10%	-0.26	(-0.68, 0.79)
15%	-0.21	(-0.65, 0.69)
20%	-0.17	(-0.41, 1.33)
25%	-0.16	(-0.53, 0.47)
30%	-0.19	(-0.54, 0.47)
35%	-0.29	(-0.92, 0.46)
40%	-0.44	(-0.99, 0.63)
45%	-0.38	(-0.55, 0.45)
50%	-0.44	(-0.90, 0.40)
55%	-0.3	(-0.82, 0.51)
60%	-0.39	(-0.94, 0.48)
65%	-0.56	(-0.68, 0.11)
70%	**-0.62**	**(-0.51, -0.05)**
75%	**-0.67**	**(-0.31, -0.14)**
80%	**-0.7**	**(-0.19, -0.17)**
85%	**-0.8**	**(-0.44, -0.40)**
90%	**-0.82**	**(-0.22, -0.63)**
95%	**-0.78**	**(-0.29, -0.43)**

Table A.6: Simulation results - CAD (19% among black patients) - Comparing traditional regression adjustment to stratification by propensity score

Simulated rate in white patients	Mean difference	95% empirical probability interval
5%	-0.19	(-0.48, 0.32)
10%	-0.15	(-0.59, 0.50)
15%	-0.52	(-1.73, 0.25)
20%	-0.27	(-1.52, 0.40)
25%	-0.18	(-1.06, 0.60)
30%	-0.17	(-1.03, 0.50)
35%	-0.29	(-1.04, 0.58)
40%	-0.35	(-0.95, 0.21)
45%	-0.23	(-0.81, 0.32)
50%	-0.33	(-0.86, 0.23)
55%	-0.48	(-1.06, 0.11)
60%	-0.35	(-1.06, 0.58)
65%	-0.56	(-1.06, 0.01)
70%	**-0.59**	**(-1.33, -0.14)**
75%	**-0.7**	**(-1.20, -0.23)**
80%	**-0.76**	**(-1.26, -0.34)**
85%	**-0.8**	**(-1.24, -0.36)**
90%	**-0.76**	**(-1.81, -0.15)**
95%	-0.58	(-2.42, 1.01)

A.3 Full Simulation Results - Marginal Mean

Table A.7: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	1.611	(0.401, 2.758)
'Correct' PS	1.557	(0.197, 2.951)
'Incorrect' Trad.	1.611	(0.409, 2.753)
'Correct' Trad.	1.555	(0.190, 2.926)
β_3	2	

Table A.8: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 3% hypertension rates among black patients vs. 82% among white patients ($\alpha_1 = 4.5$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	0.404	(-0.648, 1.451)
'Correct' PS	0.338	(-0.956, 1.572)
'Incorrect' Trad.	0.405	(-0.651, 1.441)
'Correct' Trad.	0.336	(-0.959, 1.584)
β_3	0	

Table A.9: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_1 = -4.0$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	2.432	(1.740, 3.069)
'Correct' PS	2.610	(1.932, 3.327)
'Incorrect' Trad.	2.432	(1.736, 3.083)
'Correct' Trad.	2.607	(1.922, 3.337)
β_3	2	

Table A.10: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_1 = -4.0$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	0.393	(-0.279, 1.093)
'Correct' PS	0.744	(0.149, 1.356)
'Incorrect' Trad.	0.394	(-0.286, 1.086)
'Correct' Trad.	0.743	(0.146, 1.358)
β_3	0	

Table A.11: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_1 = -4.0$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	2.622	(1.933, 3.333)
'Correct' PS	2.610	(1.932, 3.327)
'Incorrect' Trad.	2.623	(1.937, 3.336)
'Correct' Trad.	2.607	(1.922, 3.337)
β_3	2	

Table A.12: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 48% hyperlipidemia rates among black patients vs. 2% among white patients ($\alpha_1 = -4.0$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	0.753	(0.156, 1.373)
'Correct' PS	0.744	(0.149, 1.356)
'Incorrect' Trad.	0.753	(0.162, 1.377)
'Correct' Trad.	0.743	(0.146, 1.358)
β_3	0	

Table A.13: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 73% hypertension rate among black patients ($\alpha_1 = -0.2$), 69% among white, 27% hyperlipidemia rate among black patients, 24% among white ($\alpha_2 = -0.2$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	1.503	(1.279, 1.717)
'Correct' PS	1.514	(1.294, 1.732)
'Incorrect' Trad.	1.503	(1.281, 1.717)
'Correct' Trad.	1.513	(1.291, 1.728)
β_3	2	

Table A.14: Estimating β_3 as marginal mean difference between black and white patients estimated by 'correctly' and 'incorrectly' specified propensity score and traditional regression adjustment models; 73% hypertension rate among black patients ($\alpha_1 = -0.2$), 69% among white, 27% hyperlipidemia rate among black patients, 24% among white ($\alpha_2 = -0.2$)

Model	**Mean difference**	**95% Empirical Prob. Interval**
'Incorrect' PS	0.620	(0.389, 0.825)
'Correct' PS	0.631	(0.399, 0.841)
'Incorrect' Trad.	0.621	(0.395, 0.824)
'Correct' Trad.	0.631	(0.407, 0.842)
β_3	0	

Appendix B

Chapter 5 - Prognostic Scores and Sliding Dichotomy

B.1 Prognostic Score Calculations

Prognostic scores were calculated based on the logistic regression model 5.6

$$logit(P(fav_c = 1)) = \beta_0 + \beta_1 GCS + \beta_2 pupil + \beta_3 age$$

where the values listed in Table B.1 were estimated.

Table B.1: Parameter estimates from proportional odds model with GOS as outcome and using only control patients

Parameter	**Estimate**
β_0	-1.6362
$\beta 1$ - **GCS = 1**	0.9034
$\beta 1$ - **GCS = 2**	0.3735
$\beta 1$ - **GCS = 3**	0.4610
$\beta 1$ - **GCS = 4**	-1.5748
$\beta 2$ - **pupil = 1**	-0.6254
$\beta 2$ - **pupil = 2**	0.7769
β_3	0.0260

B.2 Computer Code to Generate Sample Size Comparisons for Traditional and Sliding Dichotomy Methods

```
#define probabilities of favorable outcomes under sliding dichotomy for worst and
best
#prognostic groups and treatment and control groups
p1c <- seq(0.1, 0.8, 0.025)
p1t <- seq(0.25, 0.95, 0.025)
p3c <- seq(0.1, 0.85, 0.025)
p3t <- seq(0.2,0.95,0.025)

#define distribution of patients into worst and best categories
w1 <- 0.35
w3 <- 0.65

x <- matrix(0,length(p1t),length(p3t))
p_bar <- matrix(0,length(p1t),length(p3t))
q_bar <- matrix(0,length(p1t),length(p3t))
pt <- matrix(0,length(p1t),length(p3t))
pc <- matrix(0,length(p1t),length(p3t))
qt <- matrix(0,length(p1t),length(p3t))
qc <- matrix(0,length(p1t),length(p3t))
es <- matrix(0,length(p1t),length(p3t))

for (i in 1:length(p1t))
{
for (j in 1:length(p3t))
{
p_bar[i,j] <- ((p1t[i]*w1)+(p3t[j]*w3)+(p1c[i]*w1)+(p3c[j]*w3))/2
q_bar[i,j] <- 1-p_bar[i,j]
pt[i,j] <- (p1t[i]*w1) + (p3t[j]*w3)
qt[i,j] <- 1-pt[i,j]
pc[i,j] <- (p1c[i]*w1) + (p3c[j]*w3)
```

```
qc[i,j] <- 1-pc[i,j]

es[i,j] <- pt[i,j]-pc[i,j]

x[i,j] <- (((sqrt(2*p_bar[i,j]*q_bar[i,j])*1.96) + (sqrt((pt[i,j]*qt[i,j]) +
(pc[i,j]*qc[i,j]))*1.64))/es[i,j])^2
}
}


#define difference in probability of favorable outcome in worst group (1) for
traditional definition vs. sliding (defined above)
#define difference in probability of favorable outcome in best group (3) for
traditional definition vs. sliding (defined above)

p1d <- -0.1
p3d <- 0.15
p1tt <- p1t + p1d
p1ct <- p1c + p1d
p3tt <- p3t + p3d
p3ct <- p3c + p3d


x_t <- matrix(0,length(p1tt),length(p3tt))
p_bart <- matrix(0,length(p1tt),length(p3tt))
q_bart <- matrix(0,length(p1tt),length(p3tt))
ptt <- matrix(0,length(p1tt),length(p3tt))
pct <- matrix(0,length(p1tt),length(p3tt))
qtt <- matrix(0,length(p1tt),length(p3tt))
qct <- matrix(0,length(p1tt),length(p3tt))
est <- matrix(0,length(p1tt),length(p3tt))

for (i in 1:length(p1tt))
```

```
{
for (j in 1:length(p3tt))
{
p_bart[i,j] <- ((p1tt[i]*w1)+(p3tt[j]*w3)+(p1ct[i]*w1)+(p3ct[j]*w3))/2
q_bart[i,j] <- 1-p_bart[i,j]
ptt[i,j] <- (p1tt[i]*w1) + (p3tt[j]*w3)
qtt[i,j] <- 1-ptt[i,j]
pct[i,j] <- (p1ct[i]*w1) + (p3ct[j]*w3)
qct[i,j] <- 1-pct[i,j]

est[i,j] <- ptt[i,j]-pct[i,j]

x_t[i,j] <- (((sqrt(2*p_bart[i,j]*q_bart[i,j])*1.96) + (sqrt((ptt[i,j]*qtt[i,j]) +
(pct[i,j]*qct[i,j]))*1.64))/est[i,j])^2
}
}



library(lattice)
#matrices x and x_t now contain required sample sizes for range of probabilities
temp <- x - x_t
contour(p1c,p3c,temp,xlab="P(fav=1) in prognostic group 1",
ylab="P(fav=1) in prognostic group 3", labcex=1,vfont = c("sans serif", "bold"))
```