CAUSAL EFFECTS IN RANDOMIZED TRIALS IN THE PRESENCE OF PARTIAL COMPLIANCE: BREASTFEEDING EFFECT ON INFANT GROWTH

TONG GUO

Department of Epidemiology and Biostatistics

MCGILL UNIVERSITY

MONTREAL QUEBEC, CANADA

July 24, 2009

A Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies

In Partial Fulfilment of the Requirements

for

The Degree of Doctor of Philosophy in Epidemiology and Biostatistics

DEDICATION

This document is dedicated to my daughter, Cathy Guo.

STATEMENT OF ORIGINALITY

All the work presented in this thesis is my original contribution. No any part of the thesis has been published or is in press elsewhere. The original methodological contributions include: (i) adapting principal stratification to partial compliance and defining compliance principal strata, (ii) development of the innovative dual propensity score framework to identify compliance principal strata, iii) proofs of sub-theorems related to the properties of the proposed estimators, (iv) adapting ordinal logistic regression to dual propensity score estimation, (v) development of weighting by stratum ranking and matching algorithm to identify compliance principal strata with dual propensity scores and to estimate principal effect.

Dr. Robert Platt contributed to the setting of the objectives and scope of this research, advising closely on development of the dual propensity score framework, helping with analytical derivations and interpretation of results, helping with design of the simulation study, and reviewing the dissertation multiple times throughout the development.

Dr. Michael Kramer contributed to motivating this research, the setting of the objectives, guiding on the concept of causal inference in

general and counterfactual framework in particular, advising on the development of the dual propensity score framework, providing analysis data, helping on interpretation of the application results, and reviewing Chapters 4 and 5.

Dr. Erica Moodie contributed to helping with analytical derivations and interpretation of results, advising on adapting ordinal logistic regression to dual propensity score estimation, and reviewing the dissertation.

Dr. Stanley Shapiro contributed to the setting of the objectives of this research, and reviewing the final draft of the dissertation.

ACKNOWLEDGMENTS

I am very glad to have this opportunity to express my gratitude to those persons who have helped me achieve this milestone in my life. I strongly feel that I could not have reached this goal without their dedicated support and guidance.

I must acknowledge my thesis advisor and good friend, Dr. Robert Platt, for his constant support (academically and morally) and patient guidance throughout this entire course of my research. I feel most fortunate to have worked with such a nice and gentle man.

I would like to extend heartfelt thanks to Dr. Michael Kramer for his wisdom, insight, and judgement. His inspiration motivated me to focus on this interesting and important topic, and it was a privilege to have been able to work for him. The data he has provided me offered great help for my dissertation.

I am deeply indebted to Dr. Erica Moodie. Her quick response to my spotty work inspired me to keep on working on this project at the most difficult time. Her experience and expertise benefited me in many different ways.

I also thank Dr. Stanley Shapiro, whose many helpful suggestions greatly improved the quality of the final draft of this dissertation. I appreciate the time he spent working with me in this regard.

During the past number of years I have received the care and support of many individuals, all of whom have contributed to making my time at McGill a positive and pleasant experience. I owe a great deal.

Lastly, I would like to dedicate my thesis to my family. Without them, this accomplishment could never have been achieved.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

There has been considerable growth in the statistics literature on methods for estimating causal effects from randomized controlled trials in which non-compliance occurs. However, the focus has been limited to all-or-none compliance. This thesis develops new methodology to estimate causal effects in a randomized trial setting in which non-compliance can be better classified as "full-partial-none" compliance and where subjects in both the experimental and control arm could receive experimental treatment to varying degrees regardless of treatment assignment. This new approach to address the problem is based on principal stratification theory. We define compliance stratification effects as a special case of principal stratification and use dual propensity scores (propensity scores estimated under both possible treatment assignments) to estimate compliance principal effects. We demonstrate that dual propensity scores have many of the attractive properties of the ordinary propensity score and that compliance stratification effects become estimable by adjusting for the estimated dual propensity scores using stratification, matching or regression. We apply our methodology to a breastfeeding promotion intervention trial and assess the causal effects of prolonged and exclusive breastfeeding on infant growth (weight or length) at one year of age.

ABRÉGÉ

La littérature statistique a connu un important essor en ce qui concerne les méthodes employées pour estimer les effets causaux à partir d'essais sur des échantillons aléatoires contrôlés en présence de la non-conformité. L'attention a toutefois été portée sur la présence ou l'absence totale de la conformité. Ce mémoire élabore une nouvelle méthodologie qui sert à estimer les effets causaux d'essais sur des échantillons aléatoires où la « non-conformité » est remplacée par une conformité « Totale-partielle-absente » et où les sujets, à la fois des côtés de l'expérimentation et du contrôle, pouvaient recevoir des traitements expérimentaux à différents degrés, indépendamment de l'application du traitement. Cette nouvelle façon d'aborder le problème se base sur la théorie de stratification principale. Nous définissons les effets de la stratification de la conformité comme étant un cas particulier de la stratification principale et utilisons des scores de propension duaux (estimés au-dessous des deux applications du traitement possibles) pour estimer les effets principaux de la conformité. Nous démontrons que les scores de propension duaux conservent beaucoup de propriétés intéressantes du score de propension normal et qu'ils peuvent servir à estimer les effets de stratification de la conformité. Nous appliquons notre méthodologie à l'allaitement naturel et évaluons les effets causaux d'un allaitement naturel exclusif et prolongé sur la croissance (le poids et la taille) du nourrisson à l'âge d'un an.

Chapter 1 INTRODUCTION

Non-compliance is not uncommon in randomized controlled trials (RCTs) in which study subjects are randomly assigned to different treatment groups. Non-compliance occurs when subjects fail to adhere to their assigned treatment, partially or completely. Often non-compliance does not occur randomly. It may be associated with prognostic factors and treatment response, and may represent an intrinsic characteristic of subjects. Historically, causal inference from randomized trials with non-compliance has been viewed as problematic (Goetghebeur and Shapiro, 1996). The problem arises when attempts are made to estimate the treatment effect that would have been reached if there had been perfect compliance. Criticism of these attempts includes shifting the focus from a pragmatic to an explanatory question, and biased selection of the treatment group based on post hoc features observed after randomization (Armitage, 1998).

The most popular analysis for RCTs with non-compliance is the intention-to-treat (ITT) approach (Sheiner and Rubin, 1995). This method ignores observed compliance information and compares those assigned to the treatment group to those assigned to the control group. This procedure provides a valid estimate of *the effect of treatment assignment* on the outcome. 'As-treated' (AT) and 'per protocol' (PP) are two other popular ways to analyze the data. AT analysis compares those who received the experimental treatment with those who received the control treatment, ignoring treatment assignment. PP analysis compares those who were assigned to and received the experimental treatment with those who were assigned to and received the control treatment, in other words, subjects who comply with the protocol. However, AT and PP

estimates generally do not estimate true causal effects because they compare groups of subjects who are fundamentally different - they are actually different mixtures of some unidentifiable subpopulations (Imbens and Rubin, 1997b). On the other hand, the ITT estimate compares two groups with the same expected mixture of some subpopulations; that is, the proportion of those subpopulations on average will be the same in the treatment and control groups.

Fundamentally, the ITT approach retains the benefit of randomization and therefore is considered to provide a true causal effect estimate. Critics of this approach point out that ITT focuses on the effect of assignment of treatment rather than the effect of receipt of treatment, and that comparison of treatment assignment is attenuated relative to the true causal effect of treatment received, because non-compliers will dilute whatever effect might have been revealed by compliers (Levis and Machin, 1993; Sheiner and Rubin, 1995). The latter argument is especially appealing when interest centers on biological efficacy and/or non-compliance is substantial, and ITT is likely (but not certainly) to underestimate the true treatment effect.

There has been considerable growth in the statistics literature on methods for estimating causal effects from RCTs in which non-compliance occurs. However, these methods have been limited to all-or-none compliance (see, for example, Angrist, Imbens and Rubin, 1996; Yau and Little, 2001; Frangakis, Rubin, and Zhou, 2002). This research project extends the methodology of estimating causal effects in the presence of all-or-none treatment compliance to a situation in which non-compliance is better classified as *full, partial,* or *no* treatment compliance. Compliance can be 'partial' in the sense that a fraction of an assigned treatment is taken. In this thesis, we use the propensity score

method to construct a situation where treatment received is unconfounded even though it is not randomly assigned. We define propensity scores under both possible treatment assignments (i.e., the treatment group and the control group) and call them the dual propensity score (DPS). We demonstrate that the DPS in an RCT has many of the attractive properties of the propensity score. We define compliance principal strata based on principal stratification theory and estimate the principal effects (the effects of treatment based on principal strata) using DPS stratification and matching.

This dissertation is organized as follows: Chapter 2 reviews the causal inference literature relative to non-compliance in RCTs. Chapter 3 introduces the Promotion of Breastfeeding Intervention Trial (PROBIT) and reviews results on breastfeeding on infant growth from three published papers based on PROBIT. Chapter 4 develops the DPS methodology and focuses on DPS stratification algorithms. Chapter 5 defines compliance principal stratification in the presence of partial compliance and focuses on DPS matching algorithms. Chapter 6 evaluates the newly developed methods with simulation studies. Finally, Chapter 7 reviews the contributions and limitations of the DPS methodology, and concludes this dissertation.

Chapter 2 LITERATURE REVIEW

This chapter provides a brief review of the literature on causal inference in the fields of statistics, epidemiology and clinical research, with a focus on methods based on the Rubin Causal Model (RCM) (Holland, 1986). We organize this chapter as follows. First, we introduce the concept of *causality* in clinical studies. Second, we define causal effects in clinical studies via potential outcomes and the counterfactual framework. Third, we review approaches that have been used to define and estimate causal effects in RCTs in the presence of non-compliance, including instrumental variable approach, complier average causal effect, principal stratification, G-estimation of structure nested models, as well as other approaches not formally based on the RCM framework. Fourth, we review propensity score methodology, since our new estimation approach (dual propensity score) will be based on propensity score methods. Last, we review the limitations of existing approaches and conclude this chapter with a brief summary.

2.1 Introduction of Causal Inference

The central aim of clinical studies is to establish a cause-effect relationship between an agent or treatment and an outcome. However, the appropriate methodology for extracting such relationships from data has been fiercely debated (Little and Rubin, 2000). In general, *causality* denotes a necessary relationship between one event (cause) and another event (effect), which is the direct consequence of the first. There are two fundamental questions of causality (Pearl, 2000): 1) what empirical evidence is required for legitimate inference of cause-effect relationship? And 2) given that we are willing to

accept causal information about a phenomenon, what inference can we draw from such information, and how? These questions have been without satisfactory answers for centuries. The modern analysis of causation can be traced back at least to philosophers in the eighteenth and nineteenth centuries (Rubin 1990; Sobel, 1995), but it was not until the 1920s that a formal statistical model for causal inference was proposed by Neyman (1923) for randomized experiments. Subsequently, the last half century witnessed a rapid increase in the use of formal methods for the analysis of causal effects (Greenland, 2000b).

Of the methods that appear in clinical studies and health sciences on causal inference, most can be identified with three approaches: counterfactual (potential outcomes) models, graphical models, and structural equation models (Greenland, 2000b). We focus our review and discussion on the counterfactual or potential outcomes framework, which has been a well-established and popular framework of causal inference in evaluation of treatments in health science. Little and Rubin (2000) argued that there were three formal statistical modes of causal inference in clinical and epidemiological studies via counterfactual models: one model-based and two randomization-based (Fisher's randomization-based inference and Neyman's randomization-based inference). They argued that Fisher's approach is the more direct conceptually and it is closely related to the mathematical idea of proof by contradiction. Neyman's approach, alternatively, can be viewed as drawing inferences by evaluating the expectations of statistics over the distribution induced by the assignment mechanism to calculate a confidence interval for the typical causal effect (Little and Rubin, 2000). Our focus is on Neyman's randomization-based inference approach.

2.2 Causal Inference in Clinical Studies via Potential Outcomes

The definition of *cause* is complex and challenging, but the idea of the causal effect of a treatment seems more straightforward and practically useful. A key concept of this view of causation is the so called *counterfactual* element: a certain event would not have occurred, if, contrary to fact, an earlier event had not occurred. Actually, this intuitive idea can be found in the empirical research literature in several fields, including economics (e.g., Heckman, 1996), epidemiology, (e.g., Robins, Hernan, and Brumback, 2000), the social and behavioural sciences (e.g., Sobel, 1990; Sobel, 1995), and statistics (e.g., Cox, 1992; Rubin, 1978). This intuitive definition of causal effects was first presented by Neyman (1923) for randomized experiments and randomization-based inference and later advocated by Rubin (1974; 1978) for nonrandomized observational studies, an approach also known as the Rubin Causal Model (Holland, 1986).

There are two essential parts of the RCM framework: one is potential outcomes, another is the assignment mechanism (Rubin, 2007b). Potential outcomes are all the outcomes that would be observed if each of the treatments could be applied to each of the units. The causal effects are then defined as comparisons of potential outcomes among a common set of units. The assignment mechanism describes how units were assigned the treatment they received. To describe this theory, suppose we want to study the effect of an experimental intervention S on a subsequent outcome Y on a population U. Let $Y(S = s)$ be the outcome of an experimental unit within the population U under an experimental condition $S = s$. $Y(S = s)$ represents the value that Y would take had S been s. Assume that there are two levels of the experimental intervention: $S = 1$ for experiment treatment

(E) and S = 0 for control treatment (C). In this case there are two 'potential versions' of Y; $Y(S = 1)$ and $Y(S = 0)$. $Y(S = 1)$ represents the value of Y that would have occurred had the individual received E, and $Y(S = 0)$ represents the value of Y that would have occurred had the individual received C. Under the RCM framework, the causal effect on a single individual is then defined as the comparison (e.g., difference) between $Y(S = 1)$ and $Y(S = 0)$. However, the fundamental problem of this theory is that it is impossible to observe the value of $Y(S = 1)$ and $Y(S = 0)$ on the same unit at the same time. The statistical solution for the fundamental problem is to use the population U to estimate the average causal effect (ACE) over U (Holland, 1986).

One key assumption of the RCM is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978). This assumption has two parts: 1) there is no interference between study units such that the treatment received status of one unit does not influence the outcome or treatment received status of another unit; and 2) only one of the potential outcomes will be observed for each unit, also known as the consistency assumption. Under SUTVA, individual-level causal effects can be defined without reference to other individuals in the study. Another assumption is the randomization assumption, which requires that treatments are randomly assigned to units and that all baseline variables are independent of treatment assignment. A weaker form of the randomization assumption requires that the potential outcomes are independent of treatment assignment given baseline covariates. For a randomized experiment, the treatment is randomly assigned and thus all measured and unmeasured confounders should be equally distributed between two treatment groups. For nonrandomized studies (or randomized studies with non-compliance), the treatment assignment (or receiving) generally are not independent

of all potential outcomes (Rubin, 2005), and thus a model is needed for the assignment mechanism so that all measured and unmeasured confounders are equally distributed between the two treatment groups given baseline covariates.

The RCM framework derives the cause and effect in simple terms of interventions and potential outcomes, rather than leaving them informal. However, because only one of the treatments can be administrated to a unit, for each unit only one potential outcome will be observable; the rest will remain counterfactual. Hence, it has been criticized by some authors for including structural elements that are unidentifiable by randomized experiments alone (Dawid, 2000). For example, the correlation among potential outcomes cannot be observed on one unit; therefore, nothing about the correlation of $Y(S = 1)$ and $Y(S = 0)$ can be inferred from observing interventions and outcomes alone. Nevertheless, the RCM framework provides conceptual clarification and highlights the limits of what statistical analyses can show without background theory about causal mechanisms (Greenland, 2000b).

2.3 Causal Inference in RCTs in the Presence of Non-compliance

2.3.1 Causal Inference in RCTs

The definition of causal effects via potential outcomes and the formal consideration of the assignment mechanism clarify the roles of two design features of clinical studies: the inclusion of a control group and the randomization of treatment assignment (Little and Rubin, 2000). This is the reason why RCTs (*randomized controlled* trials) with an intention-to-treat (ITT) analysis are considered to be the gold-standard, and causal conclusion can be drawn solely based on the distribution of statistics

induced by a randomized assignment without any additional assumptions. Under perfect treatment compliance, the standard ITT approach provides unbiased estimates of the effect of randomization (effectiveness), which is the same as the efficacy. Formally, efficacy can be defined as the etiological impact of actually receiving a treatment on an outcome of interest (Last et al., 2000), and it can be interpreted as the effect if everyone in the population were to take the treatment. It is also referred to as the explanatory approach in the sense that it seeks to provide evidence in choosing one treatment over another when both are administered under optimal conditions (Armitage, 1998). However, in the presence of treatment non-compliance, the effectiveness may differ from the efficacy. Therefore, new methods are needed for RCTs with non-compliance to allow researchers to compare treatments and to answer the question of efficacy while validly taking into account factors measured *after* randomization.

2.3.2 RCTs with Non-compliance

Non-compliance in RCTs occurs when subjects randomly assigned to treatment groups fail to adhere to their assigned treatment. For instance, subjects who are assigned to the control group receive the experimental treatment, or subjects who are assigned to the experimental group refuse the experimental treatment. In such cases, the analysis and interpretation for causal effects becomes complicated. A fundamental principle in RCTs in comparing treatment groups is that groups must be from the same population; so that they are alike in important aspects and differ only in the treatment received.

An approach that has been applied to estimate the causal effect of a treatment in RCTs with non-compliance is the instrumental variables (IV) approach, which treats

randomization as an instrumental variable (to be defined below). The IV technique has been known for decades and is widely used in econometrics (Greenland, 2000a). In a series of influential papers (Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Imbens and Rubin, 1997b; Hirano et al., 2000), Rubin and his colleagues reinterpreted the IV estimator in the RCM framework and laid out the assumptions under which this estimator has causal interpretation. The idea behind the IV approach is to estimate the effect of a variable correlated with the error term in a regression by using another variable that is correlated with the response, but that does not directly affect the response. The traditional definition qualifies a variable Z as instrumental (relative to the pair (X, Y)) if Z is associated with X but not associated with Y except through its association with X. Formally, suppose X and Y are the exposure and outcome of interest, and we can observe their relation to a third variable Z. Let V be the set of all variables that affect X and Y. The variable Z is called an instrument or instrumental variable relative to the total effect of X on Y if the following criteria hold: (i) Z is independent of V; (ii) Z is correlated with the error terms of X; and (iii) Z is independent of Y given X and V (Greenland, 2000a). Note the last assumption implies that Z has no direct effect on Y.

In a randomized trial with non-compliance, Z becomes treatment assignment, which is randomized and so fulfills assumption 1; X becomes treatment received (compliance), which is affected but not fully determined by assignment Z. Y is the outcome, which is affected by X but not directly affected by Z. The IV estimator corrects the ITT estimator for non-compliance and yields a direct estimate of treatment efficacy. Therefore, the potential outcome definition of causal effects together with IV technique

provide a useful basis for understanding non-compliance problems and assumptions implied by various estimation strategies (Little and Rubin, 2000).

2.3.3 Complier Average Causal Effect

Angrist, Imbens, and Rubin (1996) use an IV formula to estimate the causal effect in a randomized study with non-compliance, according to which the ITT measure should be divided by the fraction of subjects who comply with the treatment assigned to them in the experimental group. A simple version of this approach was first proposed by Sommer and Zeger (1991). Angrist, Imbens, and Rubin (1996) showed that the corrected formula is valid for the subpopulation of "responsive" subjects - subjects who would have changed treatment status if given a different assignment. They called this estimate the local average treatment effect (LATE) (Imbens and Angrist, 1994). LATE is the average treatment effect among *compliers* (those who both comply with their actual assignment and who would comply with the assignment not assigned). Unfortunately, this subpopulation cannot be identified. Subsequently, Imbens and Rubin (1997a; 1997b) extended this work, applying likelihood and Bayesian procedures to estimate LATE; they referred to LATE as complier average causal effect (CACE). Hirano et al. (2000) developed methods to allow for the presence of pre-treatment variables (covariates). When outcomes are continuous, Little and Yau (Little and Yau, 1998; Yau and Little, 2001) extended the method to a longitudinal study and estimated CACE by maximum likelihood. Frangakis and Rubin (Frangakis and Rubin, 2002; Frangakis, Rubin, and Zhou, 2002) addressed the issue of non-compliance by generalizing the instrumental

variables approach to principal stratification and using a full Bayesian approach to estimate the effect of treatment.

Robins and Greenland (1996) in their comment on the paper by Angrist, Imbens, and Rubin (1996) argued that three different treatment effects (TE) could be defined in RCTs with non-compliance: ATE (the global average treatment effect), which is the average effect of treatment in the entire study population; LATE or CACE, which is the ATE in the subpopulation of compliers; and ITT, which is the average effect of treatment assignment. The ATE is the difference between the mean outcome if all individuals had been assigned and complied with the treatment and the mean outcome if all individuals had been assigned and complied with the control treatment. Robins and Greenland (1996) showed that all three TE would equal zero under the sharp null hypothesis of no treatment effect. Further, they argued that under the alternative, the ATE can be of greater public health interest than the CACE or ITT. In the absence of covariates, the ATE and CACE are the same if the ATE is the same for compliers as for non-compliers if they had in fact complied. When this assumption does not hold, the ATE and CACE differ, but additional information is needed to estimate the difference (Little, Long, and Lin, 2008). Robins (1994) also introduced the class of structural nested mean models (see Section 2.3.5 for detailed review) for the average treatment effect on the treated (ATT) and showed that ATT is the IV estimand by assuming that the average treatment effect in the untreated equals that in the treated (Robins and Greenland, 1996). Imai et al. (2008) give a comprehensive review of terminology and definitions for causal effects.

In addition to the assumptions of SUTVA and randomization, use of the CACE requires the following assumptions (Angrist, Imbens, and Rubin, 1996): 1) Exclusion

restriction assumption implies that any effect of treatment assignment on the potential outcomes must be exclusively through the actual treatment received for the whole population; 2) Monotonicity assumption rules out the existence of subjects who take the opposite treatment to that assigned; 3) Nonzero denominator assumption implies that there exists at least one complier.

2.3.4 Principal Stratification

Based on the work presented in Angrist, Imbens, and Rubin (1996), Frangakis and Rubin (2002) introduced the principal stratification theory to estimate the causal effects *within* separate partially observed subpopulations. The primary goal is to compare the effects of treatments, adjusting for post-treatment characteristics such that the adjusted estimands are causal effects. The principal stratification model stratified the population into latent classes or principal strata based on the potential values of a post-treatment variable S^{obs}, under the randomized treatment assignment. The post-treatment variable S^{obs} is considered to encode characteristics of the unit as well as of the treatment. Because these principal strata are based on potential outcomes for S^{obs} under different randomized treatment conditions for each individual, treatment effects on outcome within each principal stratum can be interpreted causally (Frangakis and Rubin, 2002; Frangakis, Rubin, and Zhou, 2002; Jin and Rubin, 2008). The key property of principal strata is that they are not affected by treatment assignment and therefore can be used just as any pre-treatment covariate, such as age category. Adjusting for the post-treatment variable within principal strata always generates causal effects because it always compares potential outcomes for a common set of people.

In RCTs in the presence of all-or-none treatment non-compliance, principal stratification models have been used to estimate the effects of the randomized treatment assignment within principal strata based on each subject's prospective compliance behaviour under each treatment assignment (Frangakis and Rubin, 1999; Frangakis). Such compliance principal strata are not affected by actual treatment assignment. When the control group does not have access to the experimental treatment (e.g., Frangakis and Rubin, 1999; Frangakis and Rubin, 2002), there are two compliance principal strata: compliers and never-takers. Compliers are those subjects who would take the experimental treatment only when assigned to it. Never-takers are those who would refuse the experimental treatment regardless of treatment assignment. When the control group does have access to the experimental treatment, there are two more compliance principal strata: always-takers and defiers. Always-takers are those who would take the experimental treatment regardless of treatment assignment. And defiers are those would do the opposite of what they are assigned.

When only compliers and never-takers exist, both Bayesian (e.g., Hirano et al., 2000; Imbens, Rubin, and Zhou, 2000) and likelihood approaches (e.g., Little and Yau, 1998; O'Malley and Normand, 2005; Yau and Little, 2001) have been used. Peng et al. (2004) compared Bayesian methods to likelihood methods and concluded that both methods yield similar results. When there exist three or four compliance principal strata, to our knowledge only the Bayesian approach (e.g., Barnard et al., 2003) has been used.

However, the fundamental problem is that the principal stratum to which a subject belongs cannot be observed directly. Inference about principal effects requires prediction of the subject's missing membership in the principal strata (Frangakis and Rubin, 2002;

Frangakis, Rubin, and Zhou, 2002; Jin and Rubin, 2008). Under the likelihood approach and assuming there are no always-takers and defiers, the EM algorithm is used by treating the unobserved compliance strata in the control group as missing (Little and Yau, 1998). Under the Bayesian approach, the Markov chain Monte Carlo (MCMC) technique has been use to implement the mixture distribution estimation with the specification of prior distributions (Hirano et al., 2000; Ten Have et al., 2004). All approaches classify non-compliance as either all or none. In the cases of partial compliance, only the Bayesian approach with full likelihood has been developed very recently (Jin and Rubin, 2008).

The principal stratification approach requires the same set of assumptions as in the CACE. These assumptions are as follows. 1) SUTVA - potential outcomes do not depend on the treatment status of other individuals. 2) Randomization - the treatment assignment is randomized so that principal effects can be expressed as the comparison between two treatment groups. 3) Exclusion restriction (ER) assumption of treatment assignment given treatment received - the assigned intervention cannot operate through other means apart from the treatment receiving, therefore, the causal effects are zero for compliance principal strata (i.e., always-takers and never-takers) in which subjects receive (or do not receive) the experimental treatment regardless of their treatment assignment. 4) Monotonicity assumption of treatment assignment and treatment received - there are no defiers when controls have access to the experimental treatment. 5) Nonzero denominator assumption - the population includes some compliers.

2.3.5 Other Approaches

The compliance as an explanatory variable approach is a model-based approach to estimating efficacy in RCTs with non-compliance (Efron and Feldman, 1990). Efron and Feldman (1990) brought this new idea to model the causal effect of an experimental treatment in a placebo-controlled trial. They treated compliance as a continuous explanatory variable so that mean effect is modeled as a linear function of the percentage of assigned experimental treatment that is actually taken. They considered compliance with assigned treatment is an attribute of the subject, and compares subjects who received a treatment dose on the treatment group with comparable subjects on the placebo group who took an equal amount of placebo. The key assumptions of this approach include 1) no contamination - subjects randomized to the control group have no access to the experimental treatment; 2) perfect blinding - the placebo presents an identical challenge for compliance does as the experimental treatment. The second assumption states that compliance subgroups observed in both treatment groups are comparable when they have the same position on the compliance distributions within each randomized group. The assumption is usually inconsistent with data and was criticized by Albert and Demets (1994), who showed how sensitive the estimator is to this assumption.

Structural Nested Mean Models (SNMMs) have been proposed to analyze randomized trials with continuous outcomes and non-compliance (Robins, 1994; Fischer-Lapp and Goetghebeur, 1999; Moodie et al., 2008; Greenland, 2009) in longitudinal settings. The estimators avoid the assumption of comparability between compliance on

the placebo and the treatment, and analyze the causal effects as a function of subject characteristics (see, e.g., Goetghebeur and Shapiro, 1996; Zeger, 1998). SNMMs incorporate post-baseline information into the modeling of the association between the exposure variable and compliance. In the placebo-controlled setting, SNMMs introduce parameters expressing the causal effect of exposure in compliance selected subgroups, and model the causal effect in function of variables observed on the experimental group (Fischer-Lapp and Goetghebeur, 2004). The average treatment effects are estimated for subpopulations characterized by baseline covariates as well as the experimental treatment actually received and the treatment-free response that they would experience on placebo (Fischer and Goetghebeur, 2004). The basis of this model is G-estimation, which assumes that the potential outcome variable is independent of the assignment when treatment is not taken (Robins, 1997).

SNMMs have been used to analyze data from longitudinal studies with time-varying treatment regimes. The difficulty arises because a time-varying regime may not only be influenced by antecedent causes of the outcome but may also influence later causes, which in turn may influence the treatment regime. Under the assumption of no unmeasured confounders of compliance behaviour, these semi-parametric models specify a functional form for the difference between the mean responses under the different treatments and define a causal contrast at interval as a conditional expected difference between two counterfactual outcomes, given history. More precisely, SNMMs describe the effect on response of a particular treatment (versus none, or some standard treatment) at a particular treatment interval before following a particular treatment regimen in all

subsequent intervals, conditioning on variables measured up to that treatment interval (Moodie et al., 2008).

Marginal Structure Models (MSMs) are an alternative to SNMMs for estimating the causal effect that observed exposure levels could have on the entire study population. These methods use inverse-probability-of-treatment-weighted (IPTW) estimation (Robins, Hernan, and Brumback, 2000; Hernan, Brumback, and Robins, 2001), i.e. weighting observations by inverse probability of treatment received, and allowing causal inferences under much less restrictive independence assumptions than those required by standard methods (Greenland, 2000b). MSMs can provide valid effect estimation even when treatment compliance and confounders vary over time and treatment affects the confounders. MSMs approximate RCT estimate by re-weighting observations of a non-randomized study based on the observed covariates history of the subjects. Application of an MSM requires specifying a nuisance model for the probability of treatment assigned conditional on past confounder history. Then, an unadjusted weighted model for the effect of treatments on the outcome is fitted, where the weights are defined as the products of inverse probabilities of treatment. If the treatment models are appropriate, i.e., describe correctly the true conditional probability of being assigned a treatment given the confounders, then the estimates of the marginal effects of treatments can be interpreted causally.

Doubly Robust (DR) estimation builds on the propensity score approach and the IPTW approach of Robins and his colleagues (Robins, 1998a; 1998b; 1999a; 1999b; and

Robins, Hernan, and Brumback, 2000). DR estimation combines inverse probability weighting by a propensity score with regression modeling of the relationship between covariates and outcome in such a way that as long as *either* the propensity score model *or* the regression model is correctly specified, the effect of the exposure on the outcome will be correctly estimated, assuming that there are no unmeasured confounders (Robins, Rotnitzky, and Zhao, 1994; Robins, 2000; Bang and Robins, 2005). Specifically, one estimates the probability that a particular subject receives a given treatment as a function of that individual's covariates (propensity score). Each individual observation is then given a weight equal to the inverse of this propensity score to create two pseudo-populations of exposed and unexposed subjects that now represent what would have happened to the entire population under two treatment conditions. Maximum likelihood regression is conducted within these pseudo-populations with adjustment for confounders and risk factors. Results from extensive simulations by Lunceford and Davidian (2004) as well as Bang and Robins (2005) confirm the theoretical properties of this estimator. While DR estimators have been shown to be powerful tools for modeling, they are not in common usage yet, in part because they are difficult to implement. The DR estimator procedure runs two sets of models: one for the probability of receiving a dichotomous treatment or exposure, and another to predict either the probability of the outcome (for a dichotomous outcome) or its mean value (for a continuous outcome) within strata of the exposure. Often the causal effect of interest is the difference in means if everyone in the population received the experimental treatment versus everyone receiving no treatment. The calculation of the estimator and its standard error can be found in Lunceford and Davidian (2004).

There are still many other approaches have been proposed in the recent years to estimate causal effects in general, and to address non-compliance in RCT settings in particular, including the compliance scores approach (Follmann, 2000; Joffe et al., 2003), likelihood methods (e.g., O'Malley and Normand, 2005), multiple imputation methods (e.g., Taylor and Zhou, 2008), structural equation models (e.g., Robins, Greenland, and Hu, 1999), non-parametric bounds models (e.g., Balker and Pearl, 1997), and nested compliance class models to model time-varying non-compliance (e.g., Lin et al., 2008). We will not review these methods in this thesis.

2.4 Propensity Scores

Propensity score methods have been used in many fields to reduce the bias inherent in nonrandomized observational studies. A propensity score is the conditional probability of exposure to treatment, rather than control, given the observed covariates (D'Agostino, 1998). The propensity score was originally proposed as a method for balancing many covariates between two groups (Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1984; Rubin 1997). The propensity score provides an unconfounded mechanism whereby subgroups of units with the same distribution of the covariates involved in the assignment mechanism can be fairly compared to estimate the effect on outcomes. This is because the two subgroups appear to have been created by randomization, conditionally on measured covariates. This method can potentially balance a very large number of covariates by estimating the probability (propensity) of assignment given those covariates. The basic idea is to replace the collection of covariates with one function of these covariates. For observed covariates, theory assures

that, given any value of the propensity score, the subgroups of treated and untreated have the same joint distribution in all the covariates that were used to estimate that propensity score (Rosenbaum and Rubin, 1984). This is one of the main advantages of propensity score methods over multiple linear regression that include all the covariates contributing to the estimation of propensity score. It allows a straightforward check for whether the adjustment has made the subgroups comparable with respect to the observed covariates. Another advantage is that when covariate balance is achieved and no further regression adjustment is necessary, the propensity score method does not rely on the correct specification of the function form of the relationship (e.g., linearity or log linearity) between the outcome and the covariates (Dehejia and Wahba, 1999; Rubin, 1997). Furthermore, propensity score methods can be objective in the sense that propensity score modeling and sub-classification can be completed without ever looking at the outcomes. The propensity score is often used for matching (e.g., Heckman et al., 1996), stratification (e.g., Rosenbaum and Rubin, 1984), regression (e.g., D'Agostino, 1998), or weighting adjustment (e.g., D'Agostino, 1998). In an empirical examination, Austin and Mamdani (2006) compared the estimated treatment effect using these different propensity score methodologies. In the following two sections, we briefly review matching and stratification methods on propensity scores.

2.4.1 Matching on the Propensity Score

Propensity score matching methods have been widely used in the statistical literature but are relatively new to the settings of RCTs (Austin, 2008). This approach attempts to create pairs of treated and untreated subjects that closely resemble each other, with respect to the distributions of their observable characteristics related to the treatment exposure prior to the treatment. Then, the outcomes between two groups are compared to estimate the average treatment effect for the treatment. Most propensity score matching is based on 1-1 matching without replacement (Austin, 2008). In 1-1 matching, pairs of treated and untreated subjects are formed with a similar propensity score. In matching without replacement, an untreated subject who has been matched with a treated subject is no longer available for consideration as a potential match for other treated subjects. There are different types of matching algorithms for pair matching, including greedy matching (Austin, 2009), optimal matching (Rosenbaum, 1989), and 5→1 digit matching (Parsons, 2001). With greedy matching, a treated subject is selected, and then a nearest untreated subject is selected for matching to this treated subject. The alternative to greedy matching is optimal matching (Rosenbaum, 1995). With optimal matching, pairs of treated and untreated subjects are formed so as to minimize the total within-pair difference in the propensity score. For computation reasons, optimal matching can be difficult to implement in medium to large datasets (Austin, 2009). Most methods for propensity score matching used in the medical literature are based on greedy nearest-neighbour matching within fixed caliper widths, which attempt to match each treated subject to the nearest untreated subject within a specific caliper width. A competing method of matching that is common in the medical literature is 5→1 digit matching

(Parsons, 2003). Using this approach, treated subjects are first matched to untreated subjects on the first five digits of the propensity score. For those treated subjects that remain unmatched, matches with untreated subjects are then attempted by matching with remaining untreated subjects on the first four digits of the propensity score. This process proceeds until unmatched treated subjects are matched to untreated subjects on the first digit of the propensity score. Treated and untreated subjects that remain unmatched are then discarded. Austin (2008) conducted a systematic review of propensity score matching that was employed in the medical literature and found eight different methods that accounted for the methods that were used in 83 percent of the studies. Austin (2009) then compared the performances of different methods of propensity score matching through simulation and found that eight different matching methods resulted in a similar number of matched pairs and that qualitatively similar balance in measured baseline variables between treated and untreated subjects was observed in the different sample. Austin (2009) also provided recommendation on how to choose caliper width under different scenarios.

2.4.2 Stratification on the Propensity Score

The idea of stratification is to partition propensity scores into a set of intervals (strata). This method is also known as interval matching, blocking, and sub-classification (Rosenbaum and Rubin, 1983). Imbens (2000) suggests that under normality, the use of five strata for propensity score removes most of the bias associated with all covariates, since all bias under unconfoundedness is associated with the propensity score.

2.5 Discussion

We by no means provide a comprehensive review of the vast literature on the causal inference. For example, we have not reviewed the extensive literature on structure equation modeling (e.g., Duncan, 1974; Sobel, 1990), and modeling through causal diagrams (e.g., Greenland, Pearl, and Robins, 1999), although these subjects are closely related (Pearl, 2000). For deep and intensive review on these two fields, Greenland (2000b) recommends Sobel's (1995) discussion of the connections among causal concepts in philosophy, statistics, and social sciences, and Pearl's (2000) unified approach to counterfactual, graphical and structural equations models.

Although the RCM framework has been resisted by some authors (e.g., Dawid, 2000), we believe it derives the cause and effect in simple terms, and provides conceptual clarity and a set of tools for estimating causal effects from RCTs in the presence of non-compliance. Many methods have been developed for addressing the issue in recent years, starting with work by Robins and Greenland (1994), Angrist, Imbens and Rubin (1996), Imbens and Rubin (1997a; 1997b), and Frangakis and Rubin (2002). However, the focus of these contributions has been limited to situations where subjects in the treatment group can either take or not take the treatment, and subjects in the control group have no access to the experimental treatment. The objective of this thesis is to develop new methodology to address more complex non-compliance issues in RCT settings.

Chapter 3 BREASTFEEDING AND INFANT GROWTH: THE PROBIT STUDY

The Promotion of Breastfeeding Intervention Trial (PROBIT) is a cluster-randomized trial conducted in the Republic of Belarus (Kramer et al., 2001). A cluster-randomized trial is one in which clusters of individuals, rather than individuals themselves, are randomized to different experimental groups (Donner and Klar, 2000). Cluster randomization is preferred over individual randomization in this case because of the nature of the intervention. This trial originally randomized 34 maternity hospitals and one each of their affiliated polyclinics (i.e., the outpatient clinics in which children are followed for routine health care) either to receive a breastfeeding promotion intervention (experimental group) or to continue the maternity hospital and polyclinic practices (control group). One of 34 study sites was removed from the trial because of documented falsification of outcome data in the first wave of the study, and two sites were unwilling to participate after they learned of their allocation (one experimental and one control). This left a sample size of 17,046 randomized mother-infant pairs. The experimental intervention was modeled on the WHO/UNICEF Baby-Friendly Hospital Initiative (BFHI), which comprises ten steps that maternity hospitals must implement to become certified as "baby-friendly." The control intervention consisted of the continued current maternity hospital and polyclinic practices that existed at the time of randomization, which were characterized by delayed onset of breastfeeding, routine separation of mother and infant, scheduled feeding, frequent supplementation with formula and other liquids, and early introduction of solid foods (Kramer et al., 2001).

Healthy breastfed newborn infants weighing at least 2,500 grams at birth were enrolled during their postpartum hospital stay. Follow-up data on infant feeding, infections, and growth were collected at polyclinic visits at 1, 2, 3, 6, 9, and 12 months. A total of 17,046 infants were recruited from the 31 randomized sites, with only 555 (3.3 %) lost to follow-up prior to 12 months. Study sites were stratified by region (West [Brest and Grodno regions] versus East [all other regions]) and urban versus rural location, because women in the West and in rural areas have traditionally breastfed for longer duration and more exclusively than those in the East and those in urban regions. However, no attempts were made to standardize the measurements on weight, length, and head circumference across the study sites.

As reported in the main initial report of the trial (Kramer et al., 2001), the experimental intervention was successful in prolonging the duration of any breastfeeding. Infants were classified as exclusively breastfed at 3 months if the cross-sectional feeding information obtained at 1, 2, and 3 months indicated that no liquid or solid foods other than breast milk were being administered to the infant. An infant was considered to be exclusively breastfed at 6 months if, in addition to the above criteria, he/she was not receiving any other liquid or solid foods at the 6-month visit. The proportions of women still breastfeeding (to any degree) in the experimental group versus control group were 70 % versus 60 % at 3 months and 48 % versus 36 % at 6 months. The intervention was particularly effective in increasing the degree of breastfeeding as well. The proportion of women who were exclusively breastfeeding was seven-fold higher in the experimental group at 3 months (44.7 % versus 6.4 %) and even higher at 6 months (6.7 % versus 0.7 %).

Four papers based on PROBIT were published during 2002-2004 (Kramer, Guo, Platt et al., 2002, 2003, 2004a, and 2004b), in which I conducted all of the statistical analyses and am the second author. Among those four papers, three focused on breastfeeding and infant growth. These three papers motivated this research project and the dissertation, and are summarized in the following three subsections of this chapter.

3.1 Breastfeeding and Infant Growth: Biology or Bias?

3.1.1 Introduction

Available evidence suggested that prolonged and exclusive breastfeeding is associated with lower infant weight and length by 6 to12 months of age (see, for example, Dewey et al., 1992). That evidence was based on small observational studies with considerable potential for bias, including confounding, reverse causality and selection bias. With respect to confounding, breastfeeding mothers in developed countries differ considerably from formula-feeding mothers. In particular, they tend to be of higher socioeconomic status and are probably more "nutrition-conscious." As a result, they may be less likely to over-feed their infants independently of the choice of feeding mode. Reverse causality can create a bias in the opposite direction: slow-growing infants who are "falling off" their growth curve trajectories may be deliberately supplemented or weaned in an effort to reverse those trends. Selection bias is another concern. Breastfeeding is a "one-way street;" once breast-fed infants are weaned, they seldom if ever return to breastfeeding. Fast-growing infants may outstrip their mothers' milk supply; their hunger may then lead to crying and poor sleeping, which may subsequently lead to supplementation. Once supplementation begins, it is difficult to discontinued and

the mother cannot return to exclusive breastfeeding for some infants, leading to reduced suckling, a reduced milk supply, and the hastening of weaning. Thus, infants who continue breastfeeding may be a select subgroup whose modest growth does not tax their mothers' milk supply.

3.1.2 Methods

The statistical analysis of PROBIT used an ITT approach with adjustment for both cluster-level and individual-level covariates. To assess the potential for bias in observational studies of breastfeeding, analyses were also carried out as if an observational study had been conducted. Infants who were weaned in the first month (n = 1378) were used to approximate a formula-fed cohort and compared with those breastfed (to some degree) for the full 12 months of follow-up with either at least 3 months (n=1271) or at least 6 months (n=251) of exclusive breastfeeding. At the time of the trial, this roughly corresponded to the feeding recommendation of WHO and UNICEF respectively.

3.1.3 Results for Weight and Length

The intervention resulted in higher infant weight and length gain in the first 3 months but no discernible differences by 12 months of age. The observational analyses suggested that prolonged and exclusive breastfeeding led to slower weight and length gains between 3 and 12 months. The weight in the experimental group was significantly higher than that of the control group at 1 month (61 g), and the difference increased through 3 months (88 g at 2 months and 106 g at 3 months), declined somewhat thereafter (89 g at 6 months and 58 g at 9 months, and then disappeared by 12 months (-7 g at 12 months). Length followed a similar pattern (0.16, 0.32, 0.50, 0.46, 0.31 and 0.18

cm at 1, 2, 3, 6, 9, 12 months, respectively). In the observational analyses, infants weaned in the first month were slightly lighter and shorter at birth and their weight and length declined by 1 month, but they caught up to both the experimental and other observational groups by 6 months, and were heavier and longer by 12 months. Among infants in the two prolonged and exclusive breastfeeding groups, weight fell slightly between 3 and 12 months; length fell below the reference by 6 months with catch-up to the reference by 12 months.

3.1.4 Conclusions

The ITT results suggest that prolonged and exclusive breastfeeding may actually accelerate weight and length gain in the first few months, with no deficit detectable by 12 months of age. The early difference between the two groups may reflect the seven-fold higher proportion of experimental versus control infants who were exclusively breastfed at 3 months and the acceleration in growth from birth to 3 months among exclusively breastfed infants. The observational data showing faster weight and length gains with early weaning and slower gains with prolonged and exclusive breastfeeding may reflect selection bias, unmeasured confounding differences or a true biological effect of formula feeding.

3.2 Three versus Six Months of Exclusive Breastfeeding: Does It Matter?

3.2.1 Introduction

Although the health benefits of breastfeeding are widely acknowledged, opinions and recommendations have been strongly divided on the optimal duration of exclusive breastfeeding. Until recently, the WHO recommended exclusive breastfeeding for 4 to 6

months with the introduction of complementary foods thereafter, whereas UNICEF recommended exclusive breastfeeding for about 6 months. Few published studies, however, have directly compared the infant and maternal health consequences of these two feeding policies.

3.2.2 Methods

Of the entire randomized cohort, 2,862 infants exclusively breastfed for 3 months with continued mixed breastfeeding through at least 6 months were compared with 621 infants who were exclusively breastfed for 6 months. These two sub-cohorts comprised the subjects studied in this observational analysis and are referred to as the 3-month and 6-month groups, respectively.

3.2.3 Results for Weight and Length

Monthly weight gain from 3 to 6 months was slightly greater in the 3-month exclusively breastfed group [difference = 29 g per month (95 % confidence interval = 13 to 45 g per month)], as was length gain during the same period [difference = 1.1 mm per month (95 % confidence interval = 0.5 to 1.6 mm per month)]. No significant differences were observed in weight or length gain from 6 to 9 months, but the 6-month group had a faster length gain from 9 to 12 months [difference = –0.9 mm per month (95 % confidence interval = –1.5 to –0.3 mm per month)].

3.2.4 Conclusions

Complementary feeding between 3 and 6 months led to increases in both weight gain and length gain during that period, suggesting a “dilution” of the earlier effect by the end of the first year and/or compensatory “catch-up” or “catch-down” growth.

3.3 Feeding Effects on Growth during Infancy

3.3.1 Introduction

Previous studies have consistently reported higher weight and length gains in infants fed formula and/or other milks compared with infants following WHO/UNICEF recommendations for prolonged and exclusive breastfeeding. Despite the general consistency of the studies, many questions remain unanswered concerning the effects of specific aspects of infant feeding on growth during infancy and beyond. Considerable recent interest has also been generated by the suggestion of growth effects attributable to differences in protein content between breast milk and most infant formulas, and the even higher protein concentrations of whole cow's milk and other milks.

3.3.2 Methods

We conducted an observational cohort study nested within PROBIT. Infant growth was compared during the intervals 1 to 3, 3 to 6, 6 to 9, and 9 to 12 months, using hierarchical multivariate regression to control for size at the beginning of each interval, maternal education, geographic region, and urban versus rural location.

3.3.3 Results

Mixed breastfeeding and formula/other milk (versus breast milk only) were associated with significantly higher length gain from 1 to 3 months. In the 3- to 6-month interval, mixed breastfeeding and formula/other led to significantly higher weight and length, whereas cereal intake was associated with large and highly significant reductions in both measures. Mixed breastfeeding and formula/other milk continued to have positive albeit smaller associations with weight and length gains in the 6- to 9-month and 9- to 12-month intervals.

3.3.4 Conclusions

The results confirm the growth-accelerating effects of formula and other milks (versus breast milk) on weight and length gain throughout infancy, with a dose-response gradient and largest association observed at 3 to 6 months. However, this analysis remains an observational (cohort) analysis and therefore does not benefit from the randomized trial design.

3.4 Discussion

Previous observational studies of breastfed versus formula-fed infants have reported reduced weight gain and length gain in infants who receive exclusive and prolonged breastfeeding (see, for example, Dewey et al., 1992; Nielsen et al., 1998). Those studies have been limited by several important methodologic problems common in observational studies of infant feeding and growth, including inadequate control for socioeconomic status, regression to the mean (smaller infants tend to "catch up," whereas larger infants tend to "catch down"), and reverse causality (the feeding given is dependent on growth up to the time of the feeding decision, so that feeding can be a consequence of growth, as well as growth being a consequence of feeding).

In contrast, ITT results of breastfeeding on infant growth, as shown in Section 3.1, offered no support to the prevailing premise that prolonged and exclusive breastfeeding inexorably leads to deficits in weight and length during the first year of life. Instead, the results show that infants in the experimental group grew more rapidly for the first 3 months in both weight and length than did those in the control group, but that the differences disappeared by 12 months of age. However, the observational results in

infants weaned within the first month suggest that these infants were selected by virtue of their falling growth trajectories; these infants grew faster in weight and length even beyond the time of "catch-up," suggesting either intentional over-feeding by their mothers to promote maximal growth or a true biological effect of formula feeding (with supplementation by solids) in accelerating growth trajectories in the first 12 months of life. Similarly, those infants with prolonged and exclusive breastfeeding showed growth patterns similar to those reported in previous observational studies with a rise through 3 months and a fall thereafter. Again, these data cannot distinguish whether the observed growth trajectories represent a true biological effect of prolonged breastfeeding or reflect a selection bias as confounding differences in preference for thinner infants among mothers who practice exclusive and prolonged breastfeeding.

ITT attempts to overcome the methodologic problems discussed above. However, ITT could not compare breastfeeding versus not breastfeeding, nor more prolonged and exclusive breastfeeding versus shorter and less exclusive breastfeeding. Rather, it assessed the effect of an intervention to promote longer and more exclusive breastfeeding. The experimental intervention increased infants average breastfeeding duration and degree. More prolonged and exclusive breastfeeding must have resulted in a larger effect than the average effect observed for the entire experimental group when analyzed by ITT. Therefore, more in-depth analyses are needed to estimate the efficacy of treatment itself, rather than the effectiveness of the treatment assignment based on ITT.

Chapter 4 CAUSAL EFFECTS IN RCTS WITH ALL-OR-NONE COMPLIANCE

4.1. Introduction

Since it was introduced by Rosenbaum and Rubin in 1983, the propensity score technique has been used in many fields to adjust for nonrandom treatment assignment and to make causal inferences. However, it has not been popularly used in randomized controlled trials (RCTs) for an obvious reason -- randomization has achieved what propensity scores intend to achieve: specifically, to balance observed and unobserved covariates and thus to reduce confounding bias. In RCTs with imperfect compliance, however, treatment received is not necessarily the same as treatment assigned; therefore, treatment receipt can no longer be considered as independent of covariates, and estimation of the causal effect of treatment receipt on outcome is subject to bias. In a recently published paper (Rubin, 2007a), Rubin advocated the position that observational studies can and should be designed to approximate randomized experiments as closely as possible. He also promoted the use of propensity score methods to objectively create the subgroups of similar treated and untreated units, which are balanced with respect to covariates.

In this chapter, we examine an extension to the propensity score method in an RCT setting in the presence of all-or-none treatment compliance. We attempt to reconstruct a situation where treatment received is free of confounding bias even though it is not randomly assigned. In a way, this is similar to the role that propensity scores have played in observational studies and enable us to model the relationship between

observed exposure and observed covariates in order to solve complex problems of non-compliance.

Following the paper by Rosenbaum and Rubin (1983) and most of the literature on propensity scores, we make an assumption of unconfoundedness. Then, we define propensity scores under both possible treatment assignments (i.e., treatment and control) and call them the *dual propensity score* (DPS). We demonstrate that the DPS has many of the attractive properties of the propensity score in that, under unconfoundedness, adjusting for these two-dimensional scalar functions of the covariates removes all biases associated with differences in the measured or observed covariates. We estimate the complier average causal effect (CACE) by adjusting for the estimated dual propensity scores using two standard approaches: stratification and regression. We also show results from a naive approach: we classify subjects in both treatment and control arms as compliers based on their observed exposure under actual assignment and their estimated counterfactual propensity scores under alternative assignment. Then, we estimate the CACE based on the subset of newly identified compliers. Finally, we apply this methodology to PROBIT (see Chapter 3) and assess the causal effect of prolonged and exclusive breastfeeding on infant growth.

4.2. The Basic Framework

4.2.1 Notation and Assumptions

Consider a randomized controlled trial with two arms -- experimental (E) and control (C). Following the notation of classical causal inference and potential outcomes (for example, Holland, 1986), for each subject i, $i = 1, \ldots, n$, let Z_i ($z_i = 0,1$) be an

indicator of treatment assignment, which is random; let $S_i(z)$ be an indicator of treatment received for z = 0,1; and let $Y_i(Z_i, S_i(z))$ be the outcome, which we assume to be continuous for simplicity. Further, we assume that compliance to treatment assignment is either all or none so that $S_i(z)$ can only be 1 or 0, with 1 denoting treatment receipt and 0 no treatment receipt.

For each subject *i*, $i = 1, \ldots, n$, we observe: the treatment assignment Z_i (1 for assigned to E and 0 to C), the actual treatment received S_i^* (1 if experimental treatment received and 0 otherwise), the observed outcome Y_i^* and a p-dimensional vector of pretreatment covariates $\mathbf{X}_i$. $\mathbf{X}_i$ includes baseline covariates and demographic characteristics, and the elements of $\mathbf{X}_i$ are assumed to have been measured prior to receipt of treatment.

It is important to distinguish between observed versus potential outcomes. Potential outcomes are all the outcomes that would be observed if each of the treatments could be applied to each of the units. Then, a comparison of the potential outcomes of the same group of units under the two treatment conditions can be interpreted as a causal comparison. Let $S_i(z)$ and $Y_i(Z_i, S_i(z))$ denote potential outcomes, which are hypothetical and cannot be observed at the same time (counterfactual) for both z = 1 and z = 0. For instance, if $S_i(1)$ is observed then $S_i(0)$ becomes counterfactual, and if $S_i(1)$ is observed then $S_i(0)$ becomes counterfactual. We use the 'star' superscript (e.g., S_i^*) to denote observed, and we use the 'pound' superscript (e.g., $S_i^{\#}$) to denote not observed. These can be formally be defined as $S_i^* = S_i(z)$ and $S_i^{\#} = S_i(1 - z)$ for z = 0 and 1. And $S_i^* = S_i(1)$ if z = 1, which is a form of the consistency assumption due to Rubin.

We assume that groups E and C are exchangeable owing to randomization. We further assume that the population consists of different subpopulations that determine their receipt of treatment under both E and C assignments. Each subject's membership to the subpopulation is an inherent characteristic and determines his/her exposure to experimental treatment under both assignments. Although only one assignment and corresponding exposure can be observed and subpopulations cannot be identified, we assume that all covariates related to exposures are observed and the probability of exposure to treatment under alternative assignment can therefore be predicted. These assumptions are analogous to and consistent with the theory of principal stratification laid out by Frangakis and Rubin (2002).

4.2.2 Complier Average Causal Effect

We start from the assumption of 'no interference between units' - the Stable Unit Treatment Value Assumption (SUTVA) of Angrist, Imbens, and Rubin (1996). The idea is that the causal effect of treatment for a particular individual does not depend on assignment of treatment to other individuals. Under SUTVA, we can define the causal effect as the difference between two potential outcomes:

DEFINITION 1: CAUSAL EFFECTS OF Z ON *S* AND Z ON *Y*. *The causal effect for individual i of* Z *on S is* $S_i(1) - S_i(0)$ *and the causal effect of* Z *on Y is* $Y_i(1, S_i(1)) - Y_i(0, S_i(0))$.

Assuming further that the exclusion restriction holds, that is, $Y_i(1, s) = Y_i(0, s) = Y_i(s)$ for $s = 0, 1$ (Angrist, Imbens, and Rubin, 1996), we define the causal effect of treatment receipt S on outcome Y (efficacy) as follows:

DEFINITION 2: CAUSAL EFFECT OF S ON Y. *The causal effect of S on Y for individual i is* $Y_i(1) - Y_i(0)$.

With assumptions of SUTVA and exclusion restriction, Angrist, Imbens, and Rubin (1996) defined the average causal effect of treatment assignment (effectiveness) as

$$\Delta_{ITT} = E[Y_i(1, S_i(1)) - Y_i(0, S_i(0))]. \quad [4.2.1]$$

where ITT denotes intention-to-treat.

Further, Angrist, Imbens, and Rubin (1996) defined the *complier average causal effect* (CACE) as the average causal effect of S on Y for a group of subjects called *compliers* and showed that

$$\Delta_{ITT} = E[(Y_i(1,1) - Y_i(0,0)) \cdot (S_i(1) - S_i(0))]$$

$$= \Delta_{CACE} \cdot \Pr[S_i(1) - S_i(0) = 1] \quad [4.2.2]$$

where $$\Delta_{CACE} = E[(Y_i(1,1) - Y_i(0,0)) \mid S_i(1) - S_i(0) = 1] \quad [4.2.3]$$

DEFINITION 3: COMPLIER. *A subject is a complier if and only if* $S_i(z) = z$ for $z = 0, 1$, or the exposure vector $\boldsymbol{S_i} = [S_i(1), S_i(0)] = [1,0]$.

A subject is a complier if and only if he/she takes the treatment he/she was assigned to take and he/she would have complied with the assignment had he/she been assigned to the other group. In other words, compliers are subjects who are induced to take the treatment because they are assigned to that treatment. Randomization of assigned treatment therefore is equivalent to randomization of treatment received. Consequently, CACE is an unbiased estimate of the average treatment effect among compliers and has a causal interpretation.

DEFINITION 4: CACE. *The compliers average causal effect is* $E[(Y_i(1,1) - Y_i(0,0)) \mid \boldsymbol{S_i} = [1,0]]$.

In an RCT with perfect compliance, $S_i(z) = z$ holds for each and every subject, and $\Pr[S_i(1) - S_i(0) = 1] = 1$. Therefore, $\Delta_{CACE} = \frac{\Delta_{ITT}}{\Pr[S_i(1) - S_i(0) = 1]} = \Delta_{ITT.}$ In an RCT with all-or-none compliance, the following table (Table 4-1) defines four subpopulations, or four types of compliance strata: never-taker (*n*), always-taker (*a*), defier (*d*) and complier (*c*).

Table 4-1 Compliance Stratification

Compliance stratum	For z = 0,1, if	$\boldsymbol{S_i} = [S_i(1), S_i(0)]$
Never-takers (*n*)	$S_i(z) = 0$	[0,0]
Always-takers (*a*)	$S_i(z) = 1$	[1,1]
Defiers (*d*)	$S_i(z) = 1 - z$	[0,1]
Compliers (*c*)	$S_i(z) = z$	[1,0]

Never-takers do not receive the treatment even if they are assigned to it; always-takers receive the treatment regardless of their assignment; defiers always do the opposite of their assignment; and compliers are fully co-operative, i.e., they do what they are assigned to do. As did Angrist, Imbens, and Rubin (1996), we assume there are no defiers. Since $S_i(1)$ and $S_i(0)$ can never be jointly observed for the same subject at the same time, we cannot directly observe a subject's compliance type. Therefore, at least some compliers are unidentifiable without assumptions.

As shown in the Table 4-2, for each combination of observed z_i and S_i, there can be as many as two different potential compliance strata, even assuming there are no defiers ($\boldsymbol{S_i} = [0,1]$).

Table 4-2 Compliance Strata by Observed Exposure

Compliance stratum	z_i	S^*_i	$S^\#_i$	$[S^*, S^\#]$
c or *n*	0	0	0 or 1	[0,?]
a	0	1	1	[1,1]
n	1	0	0	[0,1]
c or *a*	1	1	1 or 0	[1,?]

However, we notice in the table that always-takers (*a*) under z = 0 and never-takers (*n*) under z = 1 are both identifiable. If we can build two predictive models $\Pr[S^*_i = 1|\mathbf{X}, z = 1]$ and $\Pr[S^*_i = 0|\mathbf{X}, z = 0]$ in which these two subpopulations are identified through the models under observed assignment, we can then resolve the rest of the identification problem by classifying *a* under z = 1 and *n* under z = 0 using the predicted values from $\Pr[S^\#_i = 1|\mathbf{X}, z = 1]$ and $\Pr[S^\#_i = 0|\mathbf{X}, z = 0]$. Notice that these two models $\Pr[S^*_i = 1|\mathbf{X}, z = 1]$ and $\Pr[S^*_i = 0|\mathbf{X}, z = 0]$ are exactly the same as those used in

propensity score methodology. Thus, we use propensity scores to estimate the causal effect in RCTs with all-or-none treatment compliance.

4.2.3 Dual Propensity and Counterfactual Propensity Scores

The propensity score is the conditional probability of exposure to treatment, rather than control, given the observed covariates (Rosenbaum and Rubin, 1983). The propensity score was originally proposed as a method for producing balance of many covariates between two groups (Rosenbaum and Rubin, 1983, 1984) . This method can potentially balance a very large number of covariates by estimating the probability (propensity) of assignment given those covariates. The basic idea of the propensity score method is to replace the collection of covariates with one function of these covariates. As a scalar summary of multidimensional covariates, the propensity score is often used for matching, stratification, or weighting adjustments.

In this section, we define separate propensity scores for each individual *i* under two different treatment assignments. At this stage, we assume that individual *i* is potentially assigned to both groups at the same time, and that the two propensity scores are known. To simplify the notation, we will drop *i* where it is obvious.

We start from a strong assumption that outcome and treatment receipt are independent given the observed covariates and treatment assignment.

DEFINITION 5: WEAK UNCONFOUNDEDNESS. *For all z*, *treatment receipt S*(z) *is weakly unconfounded given covariates* **X** *and treatment assignment* z *if*

$$Y(Z, S(z)) \perp S(z) \mid \mathbf{X} \quad \textit{for } z \in \{0, 1\} \qquad [4.2.4]$$

where $\perp$ *denotes statistical independence.*

In other words, the covariates **X** strongly predict who will receive treatment under z = 1 and who will not receive the treatment under z = 0. The definition of weak unconfoundedness follows from the definition in Imbens (2000) and is similar to the assumption of 'no unmeasured confounders' made by Robins (1994). Note that the same set of covariates **X** has been used to predict the probability of exposure under both treatment assignments.

DEFINITION 6: DUAL PROPENSITY SCORES. *Let* $r^Z(\mathbf{X})$ *be the conditional probability of receiving the assigned treatment given the covariates* **X** *and treatment assignment* Z:

$$r^Z(\mathbf{X}) = \Pr[S(z) = z \mid \mathbf{X}] \qquad \textit{for } z \in \{0, 1\} \qquad [4.2.5]$$

The dual propensity scores (DPS) ***R*** are *pair of scores* $[r^1(\mathbf{X}), r^0(\mathbf{X})]$ *where*

$$r^1(\mathbf{X}) = \Pr[S(1) = 1 \mid \mathbf{X}, Z = 1] \text{ and } r^0(\mathbf{X}) = \Pr[S(0) = 0 \mid \mathbf{X}, Z = 0] \qquad [4.2.6]$$

This definition is also in line with the definition of the generalized propensity score of Hirano and Imbens (2004) and Imbens (2000). Recall that $S(1)$ and $S(0)$ cannot both be observed at once in the same individual. We observe exposure under the actual assignment, $S^* = S(z)$. We do not observe exposure under the alternative assignment, $S^\# = S(1 - z)$. Consequently, only one of the dual propensity scores is observable (observable in the sense that all covariates that generate the propensity score are observable). We call the observed one the 'actual' propensity score, and the unobservable one the 'counterfactual' propensity score.

DEFINITION 7: ACTUAL AND COUNTERFACTUAL PROPENSITY SCORE. *The actual propensity score is* $r^{*} = r^{1}(\mathbf{X})Z + r^{0}(\mathbf{X})(1\text{-}Z)$, *and the counterfactual propensity score is* $r^{\#} = r^{1}(\mathbf{X})(1\text{-}Z) + r^{0}(\mathbf{X})Z$.

The actual propensity score is the propensity score under actual treatment assignment and is observable, while the counterfactual propensity score is the propensity score under the alternative treatment assignment and is unobservable. The dual propensity scores can be written as the pair $(r^{*}, r^{\#})$ and will be denoted as a two-dimensional vector **r**.

Next, we show that if treatments received with two assignments are weakly unconfounded given observed covariates, then they are weakly unconfounded given the dual propensity scores. The proof follows the method for the generalized propensity score of Hirano and Imbens (2004) and Imbens (2000).

THEOREM 1: WEAK UNCONFOUNDEDNESS GIVEN DUAL PROPENSITY SCORES. *Suppose that treatment receipt is weakly unconfounded given covariates* **X**, then $Y(Z, S(z)) \perp S(z) \mid r^{Z}(\mathbf{X})$ for $z \in \{0, 1\}$

Proof: We need to show that $\Pr[S(z) = z \mid Y(Z, S(z)), r^{Z}(\mathbf{X})] = \Pr[S(z) = z \mid r^{Z}(\mathbf{X})]$

From Definition 6 [4.2.5], we know

$\Pr[S(z) = z \mid \mathbf{X}, r^{Z}(\mathbf{X})]$

$= \Pr[S(z) = z \mid \mathbf{X}]$

$= r^{Z}(\mathbf{X})$.

Then, $\Pr[S(z) = z \mid r^{Z}(\mathbf{X})]$

$= E_x[\Pr[S(z) = z \mid r^Z(\mathbf{X})] \mid \mathbf{X}]]$

$= E_x[\, r^Z(\mathbf{X})]$

$= r^Z(\mathbf{X})$.

By the weak unconfoundedness assumption [4.2.4],

$\Pr[S(z) = z \mid Y(Z, S(z)), r^Z(\mathbf{X}), \mathbf{X}]$

$= \Pr[S(z) = z \mid Y(Z, S(z)), \mathbf{X}]$

$= \Pr[S(z) = z \mid \mathbf{X}]$

$= r^Z(\mathbf{X})$,

so $\Pr[S(z) = z \mid Y(Z, S(z)), r^Z(\mathbf{X})]$

$= E_x[\Pr[S(z) = z \mid Y(Z, S(z)), r^Z(\mathbf{X}), \mathbf{X}]]$

$= E_x[r^Z(\mathbf{X})]$

$= r^Z(\mathbf{X})$.

Hence, $\Pr[S(z) = z \mid Y(Z, S(z)), r^Z(\mathbf{X})] = \Pr[S(z) = z \mid r^Z(\mathbf{X})]$.

Theorem 1 shows that if the outcome and the receipt of treatment are independent given the observed covariates, then it is independent of outcomes given the individual's DPS. Within strata with the same pair of values of $(r^1(\mathbf{X}), r^0(\mathbf{X}))$, the probability of being a complier (i.e., $\boldsymbol{S_i} = [0,1]$) is independent of potential outcomes given the observed covariates; i.e., the treated and untreated groups are balanced with respect to covariates. In other words, that treatment receipt is unconfounded given the DPS.

It is easy to show that treatment receipt under the alternative assignment is weakly unconfounded given the counterfactual propensity score (CPS). For a subject actually assigned to the experimental treatment E (z = 1), his/her outcome Y given treatment

receipt S and actual propensity score is observed, so we show that the observed treatment receipt S^* is independent of the outcome given the actual propensity score. Because of randomization, we assume the same would have been true had the subjects assigned to the control treatment C (z = 0) been assigned to E (z = 1). The reverse argument holds true for subjects actually assigned to C.

Formally, under z = 1,

$Y(1, S(1)) \perp S(1) \mid r^1(\mathbf{X}) \Rightarrow Y^* \perp S^* \mid r^*$ and $Y^\# \perp S^\# \mid r^\#$ (due to randomization).

Under z = 0,

$Y(0, S(0)) \perp S(0) \mid r^0(\mathbf{X}) \Rightarrow Y^* \perp S^* \mid r^*$ and $Y^\# \perp S^\# \mid r^\#$ (due to randomization).

In summary, $(Y^*, Y^\#) \perp (S^*, S^\#) \mid (r^*, r^\#)$; that is to say, observed and unobserved treatment receipt is independent of the outcome given the dual propensity scores.

In an observational study, treatment exposure is usually self-selected and thus S may not be independent of the potential outcome. Indeed, the same characteristics that lead an individual to be exposed to a treatment may also be associated with his/her potential outcome (confounding by indication). The causal effect of S_i on Y_i cannot be estimated without bias unless we can assume no unmeasured confounders. In contrast, Rosenbaum and Rubin (1983) showed that **X** is independent of S given any value of the propensity score $r(\mathbf{X})$, so individuals from either treatment group with the same propensity score are balanced in the sense that the distribution of **X** is the same regardless of treatment status.

As a one-dimensional summary of multidimensional covariates, the propensity score is often used for matching (e.g., Heckman et al., 1996), stratification (e.g.,

Rosenbaum and Rubin, 1984), regression (e.g., D'Agostino, 1998), or weighting adjustment (e.g., D'Agostino, 1998). We intend to use the two-dimensional DPS to estimate the CACE in the same way that the propensity score has been used in observational studies. By adjusting for the DPS, we can estimate the efficacy of treatment itself, rather than the conventional 'effectiveness' of the treatment assignment based on ITT. We focus primarily on stratification, where individuals are stratified based on estimated dual propensity scores and the difference is estimated as the average of within-stratum effects; and on regression, where the DPS are used as regressors in the model.

THEOREM 2: BIAS REMOVAL WITH DUAL PROPENSITY SCORES. *Suppose that treatment receipt is weakly unconfounded given covariates* **X** and treatment assignment z. Then,

(*i*) $\mu(z, s) = E[Y(Z, S(z))| Z = z, S(z) = s] = E[Y^* | Z = z, S^* = s]$ and

(*ii*) $\Delta_{CACE} = E[\mu(1,1) - \mu(0,0) | \mathbf{r}]$ where **r** is the dual propensity score.

Proof: To prove (*i*), we show the case for z = 1, *s* = 1. The other cases follow.

Under z =1, we have $S^* = S(1)$ and $Y^* = Y(1, S(1))$. Hence,

$\mu(1,1) = E[Y(1, S(1))| Z = 1, S(1) = 1]$

$= E[Y^* | Z = 1, S^* = 1]$

which proves part (*i*).

To prove part (*ii)*, we start from the definition of the CACE [Definition 4],

$\Delta_{CACE} = E[Y(1,1) - Y(0,0) | \boldsymbol{S_i} = [1,0]]$

$= E[Y(1,1) | \boldsymbol{S_i} = [1,0]] - E[Y(0,0) | \boldsymbol{S_i} = [1,0]]$,

where

$E[Y(1,1) \mid \boldsymbol{S_i} = [1,0]]$

$= E[Y(1,1) \mid Z = 1, S(1) = 1, S(0) = 0]$

$= E[Y^* \mid Z = 1, S^* = 1, S^{\#} = 0]$

$= E_r[E[Y^* \mid Z = 1, S^* = 1]|\mathbf{r}]$ (due to Theorem 1)

$= E_r[\mu(1,1) \mid \mathbf{r}]$ (part (*i*))

Similarly, we show

$E[Y(0,0) \mid \boldsymbol{S_i} = [1,0]$

$= E[Y(0,0) \mid Z = 0, S(1) = 1 \text{ and } S(0) = 0]$

$= E_r[\mu(0,0) |\mathbf{r}]$.

Hence,

$\Delta_{CACE} = E[Y(1,1) - Y(0,0) \mid \boldsymbol{S_i} = [1,0]]$

$= E_r[\mu(1,1) \mid \mathbf{r}] - E_r[\mu(0,0) \mid \mathbf{r}]$

$= E_r[\mu(1,1) - \mu(0,0) \mid \mathbf{r}]$.

4.3. Models

4.3.1 Estimating Dual Propensity Scores

Estimation of the DPS is straightforward using the Logit model, among other techniques. Since propensity scores are usually unknown, they are typically estimated from the observed data (z, S^*, **X**) by assuming that $r^z(\mathbf{X})$ follow parametric models, e.g., a logistic regression model.

Recall that the dual propensity scores **R** are $[r^1(\mathbf{X}), r^0(\mathbf{X})]$ with $r^1(\mathbf{X}) = \Pr[S(1) = 1 \mid \mathbf{X}, z = 1]$ and $r^0(\mathbf{X}) = \Pr[S(0) = 0 \mid \mathbf{X}, z = 0]$.

Consider $r^1(\mathbf{X})$, and let $r^1(X, \beta_{z=1}) = \dfrac{1}{1+\exp(-X^T\beta_{z=1})}$, where $\beta_{z=1}$ is $(p \times 1)$ vector. From the observed data (z = 1, S^*, **X**), $\beta_{z=1}$ can be estimated by the maximum likelihood estimator $\hat{\beta}_{z=1}$ assuming the model is correctly specified

$$\sum_{i=1}^{n_1} \varphi_\beta(S_i, X_i, \beta_{z=1}) = \sum_{i=1}^{n_1} \frac{S_i - r(X_i, \beta_{z=1})}{r(X_i, \beta_{z=1})\{1 - r(X_i, \beta_{z=1})\}} \partial / \partial \beta_{z=1} \{r(X_i, \beta_{z=1})\} = 0 \qquad [4.3.1]$$

Then, the estimated propensity score is $\hat{r}^*(X_i, \hat{\beta}_{z=1}, z=1) = \dfrac{1}{1+\exp(-X_i^T\hat{\beta}_{z=1})}$, which is the probability of receiving the experimental treatment if assigned to the experimental group among subjects who were actually assigned to that group.

Applying the same model to the subjects with z = 0, we can estimate the counterfactual propensity score as $\hat{r}^{\#}(X_i, \hat{\beta}_{z=1}, z=0) = \dfrac{1}{1+\exp(-X_i^T\hat{\beta}_{z=1})}$, which is the probability of receiving the experimental treatment if assigned to the experimental group among subjects who were assigned to the control group.

Similarly, let $r^0(X_i, \beta_{z=0}) = \dfrac{1}{1+\exp(-X^T\beta_{z=0})}$, and $\beta_{z=0}$ can be estimated the same way as in [4.3.1], assuming the model is correctly specified:

$$\sum_{i=1}^{n_0} \varphi_\beta(S_i, X_i, \beta_{z=0}) = \sum_{i=1}^{n_0} \frac{S_i - r(X_i, \beta_{z=0})}{r(X_i, \beta_{z=0})\{1 - r(X_i, \beta_{z=0})\}} \partial / \partial \beta_{z=0} \{r(X_i, \beta_{z=0})\} = 0 \qquad [4.3.2]$$

Then, we estimate the actual propensity score as

$\hat{r}^*(X_i, \hat{\beta}_{z=0}, z=0) = \dfrac{1}{1+\exp(-X_i^T\hat{\beta}_{z=0})}$, which is the probability of not receiving the experimental treatment if assigned to the control group among subjects who were actually assigned to the control group.

Applying the same model to the subjects with z = 1, we estimate the counterfactual propensity score as $\hat{r}^{\#}(X_i, \hat{\beta}_{z=0}, Z=1) = \frac{1}{1+\exp(-X_i{}^T \hat{\beta}_{z=0})}$, which is the probability of not receiving the experimental treatment if assigned to the control group among subjects who were actually assigned to the experimental group. We now have an estimate of dual propensity score $\hat{r} = [\hat{r}^*, \hat{r}^{\#}]$.

4.3.2 Estimating CACE Using a Naive 'Plug-in' Approach

Assume that models exist under both assignments z =1 and z = 0, respectively, where the actual propensity scores for subjects who comply with their assignment are all higher than the scores for subjects who fail to comply under both treatment assignments. In this case, compliers become identifiable. In other words, there are two 'cut-off' points for r^* and $r^{\#}$, respectively, such that compliance is more likely to occur when $r^* \geq r^0_{\text{cut-off}}$ (or $r^1_{\text{cut-off}}$) and at the same time $r^{\#} \geq r^1_{\text{cut-off}}$ (or $r^0_{\text{cut-off}}$), while non-compliers (i.e., always takers and never takers) tend to have the opposite, i.e., $r^* < r^0_{\text{cut-off}}$ (or $r^1_{\text{cut-off}}$) and at the same time $r^{\#} < r^1_{\text{cut-off}}$ (or $r^0_{\text{cut-off}}$)$r^{\#}$.

Formally, there exist functions that

$$r^z_{i \in \{S^* = z\})} > r^z_{j \in \{S^* = 1 - z\})} \qquad \textit{for } z \in [0, 1]$$

Then, subject *i* is a complier if $S^* = z$ and $r^{\#} > \min(r^{1-z}_{i \in \{S^* = z\})})$ for z = 0, 1.

We can show that

$\text{CACE} = E[Y(1,1) - Y(0,0) \mid S(1) - S(0) = 1]$

$= E[Y(1,1) \mid S(1) - S(0) = 1] - E[Y(0,0) \mid S(1) - S(0) = 1]$

$= E[Y^* \mid z = 1, S^* = 1, S^{\#} = 0] - E[Y \mid z = 0, S^* = 0, S^{\#} = 1]$

$= E[Y^* \mid z = 1, S^* = 1, I\{r^{\#} > \min(r^0)\}] - E[Y^* \mid z = 0, S^* = 0, I\{r^{\#} > \min(r^1)\}]$

where $I\{\}$ is the index function, and $\min(r^0)$ and $\min(r^1)$ are two cut-off points based on the predictive models.

In reality, a set of clear cut-offs may not exist. However, is it reasonable to believe that certain degrees of misclassification may still yield estimates of the CACE that are only mildly biased. These questions are evaluated in a simulation to assess the robustness of CACE estimates with respect to the degrees of compliers misclassification in Chapter 6.

In order to identify compliers, we follow steps shown as below:

1. Estimate $\beta_{z=1}$ as in (4.3.1) and $\beta_{z=0}$ as in (4.3.2).
2. Calculate estimated dual propensity scores $\hat{r}^1(X_i, \hat{\beta}_{z=1})$ and $\hat{r}^0(X_i, \hat{\beta}_{z=0})$ for all subjects.
3. Sort all observations on $\hat{r}^1(X_i, \hat{\beta}_{z=1})$ for S* = 1.
4. Define a cut-off point as $\min(r^1)$.
5. Identify subjects with predicted $r^{\#} > \min(r^1)$ (assigned to z =0).
6. Keep only subjects with S* = 0 and $r^{\#} > \min(r^1)$ under z = 0.
7. Repeat steps 3-6 and keep only subjects with S* = 1 and $r^{\#} > \min(r^0)$ under z = 1.
8. Compute the differences between the two treatment groups.

Formally, the estimate is

$$\hat{\Delta}_{CACE} = \left\{ \frac{\sum_{i=1}^{n} z_i s_i Y_i I\{\hat{r}_i^0 \geq \min(\hat{r}^0)\}}{\sum_{i=1}^{n} z_i s_i I\{\hat{r}_i^0 \geq \min(\hat{r}^0)\}} - \frac{\sum_{i=1}^{n} (1-z_i)(1-s_i) Y_i I\{\hat{r}_i^1 \geq \min(\hat{r}^1)\}}{\sum_{i=1}^{n} (1-z_i)(1-s_i) I\{\hat{r}_i^1 \geq \min(\hat{r}^1)\}} \right\} \quad [4.3.3]$$

Sensitivity analyses are conducted by using different cut-offs to evaluate how sensitive the estimate is to the choice of cut-off.

4.3.3 Estimating CACE Using DPS Stratification

Stratification classifies subjects into strata determined by observed background characteristics, or by a scalar score of all the observed background covariates (propensity score). One outstanding question is how many propensity-score strata should be used in empirical analysis. Cochrane (1968) has shown that five subclasses are often enough to remove 95 percent of the bias associated with a single covariate. Imbens (2000) suggests that under normality the use of five strata for propensity score removes most of the bias associated with all covariates, since all bias under unconfoundedness is associated with the propensity score. Following the same argument, we classify subjects into 25 (5x5) 'compliance' strata (5 strata for each dual propensity score) based on DPS. Recall that the DPS are the probability of receiving the experimental treatment in the experimental group and the probability of not receiving the experimental treatment in the control group. So the higher the scores, the higher the rank (from one to five) of the stratum, and the more likely the subject is a complier. Once the strata are defined, the treatment effect is evaluated by comparing subjects directly between the two treatment groups within each stratum. Then, the mean of the differences across strata is summarized using different weighting schemes, as shown below.

The technique used for determining strata is straightforward and consists of the following steps:

1. Estimate $\beta_{Z=1}$ as in (4.3.1) and $\beta_{Z=0}$ as in (4.3.2).
2. Calculate estimated dual propensity scores $\hat{r}^1(X_i, \hat{\beta}_{z=1})$ and $\hat{r}^0(X_i, \hat{\beta}_{z=0})$ for all subjects.
3. Sort all observations on $\hat{r}^1(X_i, \hat{\beta}_{z=1})$.

4. Create *m* (e.g., 5) strata according to the sample quantiles of the $\hat{r}^1$, where the *j*th sample quantile $\hat{q}_j^1, j = 1,...,m$ is such that the proportion of $\hat{r}^1 \le \hat{q}_j^1$ is roughly *j*/*m* with $\hat{q}_0^1 = 0$ and $\hat{q}_m^1 = 1$.

5. Sort all observations on $\hat{r}^0(X_i, \hat{\beta}_{z=1})$.

6. Create *m* (e.g., 5) strata according to the sample quantiles of the $\hat{r}^0$, where the *k*th sample quantile $\hat{q}_k^0, k = 1,...,m$ is such that the proportion of $\hat{r}^0 \le \hat{q}_k^0$ is roughly *k*/*m* with $\hat{q}_0^0 = 0$ and $\hat{q}_m^0 = 1$.

7. Create *m* × *m* (e.g., 25) relatively homogeneous strata $\begin{bmatrix} \hat{q}_j \\ \hat{q}_k \end{bmatrix}, j,k = 1,...,m$.

8. Calculate the difference of the sample means of E[Y*| z =1] and E[Y*| z =0] within each of the *m* × *m* strata.

9. Compute a weighted average of the differences across strata, using three different weighting strategies; 1) equal weight, 2) the proportion of observations falling in its stratum, and 3) the compliance strata ranks.

   Formally, the estimator is

$$\hat{\Delta}_{CACE} = \sum_{j=1}^{m}\sum_{k=1}^{m} w_{jk} \left\{ \frac{\sum_{i=1}^{n} z_i Y_i I\left\{ \begin{bmatrix} \hat{r}_i^1 \\ \hat{r}_i^0 \end{bmatrix} \in \begin{bmatrix} \hat{Q}_j \\ \hat{Q}_k \end{bmatrix} \right\}}{n_{jk}^1} - \frac{\sum_{i=1}^{n} (1-z_i) Y_i I\left\{ \begin{bmatrix} \hat{r}_i^1 \\ \hat{r}_i^0 \end{bmatrix} \in \begin{bmatrix} \hat{Q}_j \\ \hat{Q}_k \end{bmatrix} \right\}}{n_{jk} - n_{jk}^1} \right\} \qquad [4.3.4]$$

where $\begin{bmatrix}\hat{Q}_j \\ \hat{Q}_k\end{bmatrix} = \begin{bmatrix}(\hat{q}_{j-1,}\hat{q}_j\,] \\ (\hat{q}_{k-1,}\hat{q}_k\,]\end{bmatrix}$, $n_{jk} = \sum_{i=1}^{n} I\left\{\begin{bmatrix}\hat{r}_i^1 \\ \hat{r}_i^0\end{bmatrix} \in \begin{bmatrix}\hat{Q}_j \\ \hat{Q}_k\end{bmatrix}\right\}$ is the number of subjects in stratum, $\begin{bmatrix}\hat{Q}_j \\ \hat{Q}_k\end{bmatrix}$, $n_{jk}^1 = \sum_{i=1}^{n} z_i I\left\{\begin{bmatrix}\hat{r}_i^1 \\ \hat{r}_i^0\end{bmatrix} \in \begin{bmatrix}\hat{Q}_j \\ \hat{Q}_k\end{bmatrix}\right\}$ is the number of subjects assigned to z = 1, and w_{jk} is the weight for each stratum with $w_{jk} = \frac{1}{m \times m}$, or $w_{jk} = \frac{n_{jk}}{n}$, or $w_{jk} = \frac{j \times k}{m \times m}$.

4.3.4 Estimating CACE Using DPS Regression

Propensity scores can also be used in regression (covariate adjustment), in which the propensity score is included into the regression model as a regressor to adjust the final estimate of the treatment effect. D'Agostino (1998) argued that if one stratifies and then uses regression adjustment within the strata, then the estimated treatment effect is a more efficient estimator than one based on matching. Estimating procedures based on matching will be presented in the next chapter.

The approach consists of the following steps:

1. Estimate $\boldsymbol{\beta}_{\mathbf{z=1}}$ as in (4.3.1) and $\boldsymbol{\beta}_{\mathbf{z=0}}$ as in (4.3.2).
2. Calculate the estimated dual propensity scores $\hat{r}^1(X_i, \hat{\beta}_{z=1})$ and $\hat{r}^0(X_i, \hat{\beta}_{z=0})$ for all subjects.
3. Include the two propensity scores and their interaction term into the regression of response on treatment received.

In practice, the resulting estimate is often similar to the one for DPS stratification.

4.4. Application: PROBIT

4.4.1 PROBIT

The Breastfeeding Promotion Intervention Trial (PROBIT) was conducted in the Republic of Belarus. Details of the study methods are outlined in Kramer et al. (2001), and related analyses (summarized in Chapter 3) have been reported in Kramer, Guo, Platt et al. (2002, 2003 and 2004) and elsewhere. The experimental intervention of PROBIT was successful in prolonging the duration and exclusivity of breastfeeding (BF). As shown in Kramer et al. (2001), however, a key feature of the trial is the substantial overlap in breastfeeding duration and exclusivity in the two randomized groups. As expected, many women in the experimental group did not exclusively breastfeed for 3 months or continue breastfeeding for 6 months, while some women in the control group did. Hence, the standard ITT analysis substantially underestimates the differences in outcome due to prolonged, exclusive breastfeeding versus a shorter duration and/or lesser degree of breastfeeding.

We classified infants as exclusively breastfed at 3 months if the cross-sectional feeding information obtained at 1, 2, and 3 months indicated that no liquid or solid foods other than breast milk were being administered to the infant. The proportions of infants still breastfeeding (to any degree) in the experimental groups as compared to those in the control groups were 70 versus 60 % at 3 months and 48 versus 36 % at 6 months, respectively. The proportion of infants still breastfeeding exclusively in the experimental versus control groups were 43.7 % versus 6.4 % at 3 months, 6.7 % versus 0.7 % at 6 months, respectively. Mothers' breastfeeding behaviour can also be described as all or none: mothers who exclusively breastfed for the first 3 months and continued

breastfeeding for at least 6 months were considered as having had prolonged and exclusive breastfeeding, or fully compliant, and otherwise as non-compliant. There are a total of 3,072 (34.7 %) prolonged and exclusive breastfeeders in the experimental group versus 408 (5.0 %) in the control group.

4.4.2 Methods

Two propensity scores are estimated: one under the experimental arm and one under the control arm, using logistic regression models. Covariates with a large impact on the exposure were selected, as well as the interactions between the selected covariates and the interactions between the selected covariates and other covariates (Austin, 2007). Covariates selected include region (West versus East and urban versus rural), maternal age (<20, 20-34, and >=35 years), maternal education (incomplete secondary, complete secondary, partial university, and complete university), prior history of having breastfed an infant for >=3 months (yes/no), caesarean delivery (yes/no), maternal smoking during pregnancy (yes/no), other children living in the household (0,1, >=2), gender (male versus female), gestational age (completed week), birth weight (g), birth length (cm) and birth head circumference (cm). Initially, all covariates were included in the models as main effects, then the interaction terms between variables showing significant effect (p-value ≤ .05, confirmed by the stepwise model selection procedure) were included with the rest of variables as well.

Table 4-3 shows the frequency distribution, means and standard deviations of the estimated DPS by observed breastfeeding behaviour under actual treatment assignment. The last row shows the overall means and standard deviations of the estimated DPS and the predicted counterfactual DPS. These statistics are almost identical (i.e.,

$\hat{r}^1(z=1) = \hat{r}^1(z=0) = 0.35$, $\hat{r}^0(z=0) = 0.40$ and $\hat{r}^0(z=1) = 0.41$), confirming our belief that the subjects in two groups are exchangeable in terms of DPS because of randomization.

Table 4-3 Frequency of DPS and Summary Statistics

Frequency	r^1 in Experimental Group (n = 8865)		r^0 in Control Group (n = 8181)	
	$S^*=0$ (5793)[2]	$S^*=1$ (3072)	$S^*=0$ (7773)	$S^*=1$ (408)
$0.0 \le \hat{r} < 0.1$	64	4	0	0
$0.1 \le \hat{r} < 0.2$	596	114	0	0
$0.2 \le \hat{r} < 0.3$	1830	644	0	0
$0.3 \le \hat{r} < 0.4$	1976	1054	0	0
$0.4 \le \hat{r} < 0.5$	907	733	0	0
$0.5 \le \hat{r} < 0.6$	357	417	0	1
$0.6 \le \hat{r} < 0.7$	53	92	3	0
$0.7 \le \hat{r} < 0.8$	8	12	57	12
$0.8 \le \hat{r} < 0.9$	0	1	936	132
$0.9 \le \hat{r} < 1.0$	0	1	6777	263
Mean $\hat{r}$ (std)[1]	0.33 (.109)	0.38 (.114)	0.95 (0.046)	0.90 (.049)
Overall Mean $\hat{r}$ (std)[1]	0.35 (.114) r^1(z = 0) : 0.36 (.114)		0.95 (.0.047) r^0(z = 1) : 0.95 (.0.047)	

[1] std = standard deviation [2] 2 missing values of r^1

4.4.3 Results

Table 4-4 shows the baseline comparison of the 'compliers' identified using the naive plug-in approach (cut-offs: min(r^1) = 0.25 and min(r^0) = 0.92). The groups were relatively comparable. However, the differences between the two distributions of Hospital Region appear to be substantial (e.g., 50.8 % vs. 33.5 % in Eastern urban region). Similarly, there are meaningful differences in Maternal Education (e.g., 44.2 % vs. 54.3 % in advanced secondary or partial university) and, to a lesser extent, in

Breastfed History (26.0 % vs. 31.2 %). Given those observed differences, we still consider the two groups to be comparable.

Table 4-4 Baseline Characteristics for Compliers

Variable	Experimental N = 2151	Control N = 5338
Hospital region		
East Belarus (urban)	1092 (50.8)	1786 (33.5)
East Belarus (rural)	360 (16.7)	1150 (21.5)
West Belarus (urban)	239 (11.1)	668 (12.5)
West Belarus (rural)	460 (21.4)	1734 (32.5)
Maternal age (yr)		
<20	293 (13.6)	639 (12.0)
20-34	1730 (80.4)	4472 (83.8)
≥35	128 (6.0)	227 (4.3)
Maternal education		
Incomplete secondary	49 (2.3)	146 (2.7)
Complete secondary	819 (38.1)	1559 (29.2)
Advanced secondary or Partial university	951 (44.2)	2899 (54.3)
Complete university	332 (15.4)	734 (13.6)
Breastfed history	560 (26.0)	1664 (31.2)
Caesarean	286 (13.3)	523 (9.8)
Maternal smoking during pregnancy	50 (2.3)	73 (1.4)
Number of other children in household		
0	1273 (59.2)	3016 (56.5)
1	714 (33.2)	1839 (34.5)
≥2	164 (7.6)	483 (9.0)
Male sex	1114 (51.8)	2762 (51.7)
Gestational age (wk)	39.5	39.3
Birth weight (g)	3478	3431
Birth length (cm)	52.4	52.1
Birth head circumference (cm)	35.2	34.9

Tables 4-5 and 4-6 present the CACE estimates of prolonged and exclusive breastfeeding on infant weight and length gains, respectively using the cut-offs of min(r^1) = 0.25 and min(r^0) = 0.92) and the sensitivity analyses using three different sets of cut-offs (min(r^1) = 0.18, 0.20, 0.30; and min(r^0) = 0.90, 0.91, 0.94, respectively) for dual propensity scores. The results show that the estimates are not very sensitive to the different cut-offs. One explanation is that, given only 5 % of subjects with S^* =1 ('non-compliance') under the control arm, the predictive model may not be very powerful to identify non-compliers.

Table 4-5 Effect of BF on Weight Gain (g) through 12 Months -- Naive Plug-in Approach

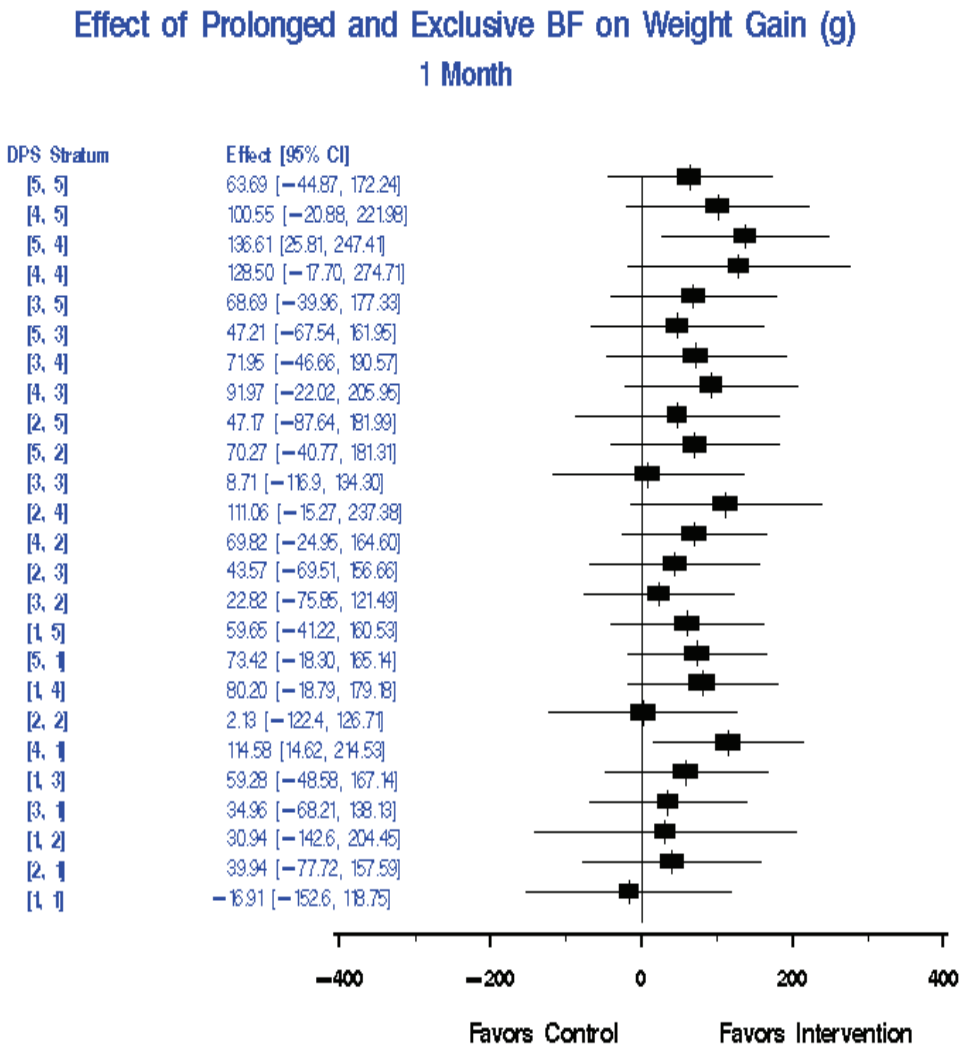
Time	Main (SE) (0.25,0.92)	Sensitivity Analyses (SE)		
		Cut-off (0.18,0.90)	Cut-off (0.20,0.91)	Cut-off (0.30,0.94)
1 m	126 (22.37)	111 (23.17)	111 (22.87)	130 (25.85)
2 m	183 (22.39)	165 (23.18)	165 (22.88)	190 (25.86)
3 m	186 (22.36)	164 (23.15)	166 (22.86)	198 (25.84)
6 m	96 (22.38)	67 (23.17)	75 (22.88)	106 (25.86)
9 m	1 (22.42)	-14 (23.21)	-12 (22.91)	7 (25.91)
12 m	-110 (22.40)	-112 (23.18)	-114 (22.89)	-106 (25.87)

Table 4-6 Effect BF on Length Gain (cm) through 12 Months -- Naive Plug-in Approach

Time	Main (SE) (0.25,0.92)	Sensitivity Analyses (SE)		
		Cut-off (0.18,0.90)	Cut-off (0.20,0.91)	Cut-off (0.30,0.94)
1 m	0.24 (0.151)	0.19 (0.137)	0.18 (0.141)	0.20 (0.149)
2 m	0.36 (0.151)	0.29 (0.137)	0.27 (0.141)	0.29 (0.149)
3 m	0.43 (0.151)	0.35 (0.137)	0.33 (0.141)	0.35 (0.149)
6 m	0.21 (0.151)	0.11 (0.137)	0.11 (0.141)	0.16 (0.149)
9 m	-0.02 (0.151)	-0.08 (0.137)	-0.10 (0.141)	-0.09 (0.150)
12 m	-0.21 (0.151)	-0.24 (0.137)	-0.27 (0.141)	-0.31 (0.150)

Figure 4-1 shows the stratified effects and their 95 % confidence intervals for prolonged and exclusive breastfeeding on weight gain at 1 month using forest plots, which have been popularly used in meta-analysis. The DPS strata are listed as $[q^1, q^0]$ and are partially sorted by compliance rank from top to the bottom; e.g., stratum [5, 5] comprises subjects with the highest DPS who are more likely to be compliers. The forest plots of the stratified effects on weight and length gains through 12 months can be found in Appendix 1. The plots are quite consistent and no strong patterns were observed.

Figure 4-1 Effect of Prolonged and Exclusive BF on Weight Gain (g) at 1 month



Tables 4-7 and 4-8 show the effects of prolonged and exclusive breastfeeding on weight and length gains through the first 12 months of life. The first column is the intent-to-treat causal effect estimate of the experimental intervention, while the second column is the CACE based on the naive plug-in approach. The third column is the estimated causal effect of prolonged and exclusive breastfeeding adjusted with dual propensity scores as regressors in the models. The last three columns show the causal effects of

prolonged and exclusive breastfeeding stratified by dual propensity scores using three different weighting schemes. Since variance estimates for the effect of treatment received may be inaccurate (because the dual propensity scores are estimated from the same set of data), bootstrap standard errors are provided. We used a nonparametric bootstrap (Efron and Tibshirani, 1998) in which we sampled subjects with replacement and used the bootstrap samples to recalculate the treatment effects adjusted for the dual propensity scores.

Table 4-7 Effect of Prolonged and Exclusive BF on Weight Gain (g) through 12 Months

Time	ITT	Naive Plug-in[1]	Regression Δ (SE)	Stratification Δ (SE) Weight by		
				Equal Weight	Proportion	Strata
1 m	61	126	62 (5.8)	62 (7.2)	63 (7.2)	75 (8.8)
2 m	88	183	89 (7.0)	89 (8.0)	89 (8.0)	98 (10.2)
3 m	106	186	108 (8.8)	108 (9.8)	108 (9.8)	110 (12.3)
6 m	89	96	91 (12.1)	92 (13.0)	91 (12.7)	85 (16.4)
9 m	58	1	60 (13.5)	63 (14.5)	62 (14.2)	56 (18.2)
12 m	-7	-110	-5 (15.1)	2 (15.8)	1 (15.5)	3 (20.3)

[1] cut-off $\min(r^1) = 0.25$ and $\min(r^0) = 0.92$

Table 4-8 Effect of Prolonged and Exclusive BF on Length Gain (cm) through 12 Months

Time	ITT	Naive Plug-in[1]	Regression Δ (SE)	Stratification Δ (SE) Weight by		
				Equal Weight	Proportion	Strata
1 m	0.16	0.24	0.16 (0.02)	0.13 (0.03)	0.13 (0.03)	0.07 (0.04)
2 m	0.32	0.36	0.31 (0.03)	0.28 (0.03)	0.27 (0.03)	0.18 (0.04)
3 m	0.50	0.43	0.49 (0.03)	0.44 (0.04)	0.44 (0.04)	0.32 (0.05)
6 m	0.46	0.21	0.45 (0.04)	0.40 (0.04)	0.40 (0.04)	0.29 (0.05)
9 m	0.31	-0.02	0.30 (0.04)	0.26 (0.04)	0.25 (0.04)	0.17 (0.05)
12 m	0.18	-0.21	0.17 (0.04)	0.14 (0.05)	0.13 (0.05)	0.08 (0.06)

[1] cut-off $\min(r^1) = 0.25$ and $\min(r^0) = 0.92$

4.4.4 Discussion of PROBIT Results

As shown in Tables 4-7 and 4-8, the regression approach produced very similar results to those of the ITT analysis, while the naive plug-in approach yielded the most 'inflated' results compared to the ITT analysis. Nevertheless, the direction of the effects was the same. Because the naive approach excluded observed 'noncompliant' subjects from the analysis plus the 'predicted' always-takers and never-takers, it is, in a way, close to a 'per-protocol' approach. The approaches based on regression and stratification with weighting by proportion also yielded similar results. Using different weighting schemes, especially the scheme with stratum ranking, we estimated CACE by overweighting the differences over the subset of 'compliers' and underweighting the differences for the rest of the strata. As will be shown later in simulations, the estimates from weighting methods will be considered less-reliable. The detailed discussion will be presented in Chapter 6 Section 6.3.

Several limitations of our analysis require discussion. One such limitation is our simplified classification of prolonged and exclusive breastfeeding as either all or none. The 'none' group in fact comprises a mixture of different breastfeeding behaviour, including mothers who stopped breastfeeding during the first month, mothers who breastfed for at least 6 months but with lesser exclusivity, and mothers who breastfed exclusively for at least 3 months but stopped breastfeeding by 6 months. Only 5 % of mothers in the control arm met the criteria of prolonged and exclusive breastfeeding, which led to low power in the predictive models. This issue will be addressed in the next chapter, where breastfeeding is classified as prolonged and exclusive, mixed, or none. Another limitation is the assumption of weak unconfoundedness, or 'no unmeasured confounding.' This is an untestable assumption, and it is not known whether all relevant covariates that could confound the effect of breastfeeding on infant growth were collected. In fact, it is likely that breastfeeding is a dynamic process and that baseline characteristics alone are not sufficient to fully capture compliance. Potential unobserved confounding (e.g., maternal depression) may result in biases in either direction. The measurement error for infant weight and length and for covariates may also have impacted on our analysis. Finally, measurements of weight and length were not standardized among the study sites. However, measurement errors are likely to be nondifferential with respect to prolonged and exclusive breastfeeding and therefore should bias the results toward the null.

4.5. Discussion

Non-compliance and causal inference for efficacy in RCTs have received considerable recent attention from researchers. However, from Sommer and Zeger (1991) to Angrist, Imbens and Rubin (1996) to Yau and Little (2001), most have assumed that subjects in the control group have no access, and hence no exposure, to the experimental treatment. The implications of this assumption are that 1) there are no always-takers and that subpopulations in the experimental group are observable (i.e., exposed = compliers, unexposed = never-takers), 2) the 'complier' indicator is missing in the control group, and 3) a single causal effect is defined for compliers and can be easily estimated as the ratio between the ITT estimate and the proportion of compliers in that population, i.e., as an instrumental variables estimate (Yau and Little, 2001).

Without denying the existence of always-takers, we addressed non-compliance utilizing propensity score methodology to correct the effect attenuation due to non-compliance. We developed a dual propensity scores approach, which can be used as a tool to identify compliance strata and subpopulations simultaneously, and to estimate the CACE using different weighting strategies without attempting to definitively and accurately identify the compliers. However, we have to acknowledge that in most placebo-controlled trials of drugs or other new treatments, the subjects in the control group have no access to the experimental treatment. Moreover, in active-controlled trials of two different drugs, often neither group will have access to the 'other' drugs. That means there are no always-takers in many drugs trials, which apparently will limit the

range of potential applications of the proposed methods (See Chapter 7 for more discussion).

Our approach was to obtain the potential treatment-free outcome through stratification, and to estimate CACE by weighting based on each subject's likelihood of compliance (i.e., by stratum ranking). Weighting is key to producing a LATE-type estimate without explicitly identifying the principal compliance strata, and it is very close to the inverse-probability-of-treatment weighting (IPTW) used in observational studies. We can either weight the strata as shown in this chapter, or match using DPS as shown in the next chapter.

In summary, we believe the dual propensity scores approach is an innovative and useful tool to estimate the causal effect in RCTs with all-or-none compliance.

Chapter 5 CAUSAL EFFECTS IN RCTS IN THE PRESENCE OF PARTIAL COMPLIANCE

5.1. Introduction

There has been considerable growth in the statistics literature on methods for estimating causal effects from RCTs in which non-compliance occurs, however, the focus of these contributions has been limited to all-or-none compliance. Our research project extends the methodology of estimating causal effects to a situation in which non-compliance is better classified as full-partial-none treatment compliance. Compliance can be 'partial' in the sense that a fraction of an assigned treatment is taken. We make use of statistical techniques developed in the previous chapter to address the issue of nonrandom treatment receipt. Specifically, we implement the dual propensity score (DPS) matching method to estimate the *compliance stratification effects* based on principal stratification. DPS matching, or matching by the predicted conditional probabilities of treatment exposure under both assignments, allows the identification of the compliance principal stratification so the principal effects become estimable.

In an RCT setting with partial non-compliance, our approach attempts to obtain information on the counterfactual outcome for each compliant subject in the experimental group (or in the control group) has been by creating a comparison group from the control group based on their DPS and classifying them into predefined compliance principal stratifications. A control subject with the closest estimated DPS is selected for each complier. We focus on one-to-one matching without replacement, using caliper matching and a variation of nearest-neighbour matching. In case of ties (i.e., more than one match), one control subject is chosen randomly from the set of possible matches.

This chapter is organized as follows. In Section 2, we introduce the theory of principal stratification and, based on this theory, we define compliance principal stratification and compliance stratification effects in an RCT setting with partial compliance. Furthermore, we show that compliance stratification effects are principal effects. In Section 3, we focus on the statistical models to estimate the newly-defined compliance principal effects by introducing the DPS and its matching strategies. In Section 4, we focus on the implementation steps of DPS matching, using an ordinal logistic regression model, and provide details on estimating procedures. In Section 5, we apply the methodology to PROBIT and assess the causal effects of prolonged and exclusive breastfeeding on infant growth. Finally, Section 6 summarizes and concludes the chapter.

5.2. Compliance Principal Stratification

5.2.1 Principal Stratification

Consider a group of units $i = 1, ..., n$ where each can be potentially assigned either an experimental treatment (z = 1) or a control treatment (z = 0). Let $S_i(z)$ be a post-treatment variable measured in addition to outcome Y and let $S_i(z) = s$ for z = 0, 1. Let $Y_i(z, s_i(z))$ be the outcome if unit i is assigned treatment z with post-treatment value of $s_i(z)$. Formally, principal stratification has been defined by Frangakis and Rubin (2002) as follows:

DEFINITION 1: BASIC PRINCIPAL STRATIFICATION. *The basic principal stratification P_0 with respect to post-treatment variable S is the partition of units i* = 1, 2, …, *n such that within any set of P_0, all units have the same vector* $[S_i(1), S_i(0)]$.

DEFINITION 2: PRINCIPAL STRATIFICATION. *A principal stratification P with respect to post-treatment variable S is a partition of the units whose sets are unions of sets in the basic principal stratification P_0.*

DEFINITION 3: PRINCIPAL EFFECT. *Let P be a principal stratification with respect to the post-treatment variable S and let S_i^P indicate the stratum of P to which unit i belongs. Then a principal effect with respect to that principal stratification is defined as a comparison of potential outcomes under a control versus experimental condition within a principal stratum k in P, i.e., a comparison between the ordered sets* $\{Y_i(1,1) : S_i^P = k\}$ and $\{Y_i(0,0) : S_i^P = k\}$.

A principal effect has the following properties: 1) the stratum S_i^P to which unit i belongs is unaffected by treatment for any principal stratification *P*; and 2) any principal effect, as defined in Definition 3, is a causal effect within the principal stratum.

5.2.2 Compliance Principal Stratification with All-or-None Compliance

Imbens and Rubin (1997b) describe the compliance behaviour by one of the four mutually exclusive compliance strata: *compliers* (*c*); *never-takers* (*n*); *always-takers* (*a*); and *defiers* (*d*). In other words, a complier takes his/her assigned treatment no matter what that assignment is; an always-taker always takes the experimental treatment; a never-taker always takes the control treatment; and a defier always does the opposite of what he/she is told. We assume there are no defiers (*d*), as did Angrist, Imbens and Rubin (1996).

THEOREM 1: COMPLIANCE STRATIFICATION *P* IS A BASIC PRINCIPAL STRATIFICATION.

Proof: Let the treatment compliance indicator $S_i(z)$ be a post-treatment variable where $S_i(z) = 0$ or 1, and $i = 1, 2, \ldots, n$, and $z = 0, 1$. The compliance stratification *P* is defined as: if $[S_i(1), S_i(0)] = [0,0]$, then $P_i = n$; if $[S_i(1), S_i(0)] = [1,1]$, then $P_i = a$; if $[S_i(1), S_i(0)] = [1,0]$, then $P_i = c$; if $[S_i(1), S_i(0)] = [0,1]$, then $P_i = d$. Each subject *i* belongs to one and only one compliance stratm $P_i = k$ for $k \in \{a, n, c, d\}$. Within each stratum *k*, all units have the same vector $[S_i(1), S_i(0)]$. Therefore, compliance stratification *P* is a basic principal stratification.

THEOREM 2: CACE IS A PRINCIPAL EFFECT.

Proof: CACE is defined as the difference of two potential outcomes within the principal stratum $P_i = c$ or the vector $[S_i(1), S_i(0)] = [1,0]$. Therefore, by definition, CACE is the principal effect with respect to treatment receipt $S_i(z)$, for $z = 1, 0$.

5.2.3 Compliance Principal Stratification with Full-Partial-None Compliance

We extend the description of compliance behaviour from a dichotomous classification (all-or-none treatment compliance) to a trichotomous classification (full-partial-none treatment compliance) by considering partial compliance. It is often the case that subjects are exposed to different levels of treatment, even though the treatment assignment is binary. An example would be full-dose, partial-dose or no-dose compliance with a test drug in a placebo-controlled trial.

Extending the work of Imbens and Rubin (1997b), we define compliance behaviour by the values of the vector $S_i(z)$ with $S_i(z) = 0, ½, 1$, for $z = 0, 1$. The trichotomous treatment compliance indicator $S_i(z)$ denotes the receipt of treatment given assignment z; $S_i(1) = 1$ if subject *i* receives a full dose, $S_i(1) = ½$ if subject *i* receives a partial dose and $S_i(1) = 0$ if subject *i* does not receive any dose, given *i* is assigned to receive the treatment; and $S_i(0) = 1, ½$ or 0 if *i* receives a full/partial/no dose given subject *i* is assigned not to receive the treatment. In contrast to the random assignment of Z_i, subject *i* chooses whether or not to comply with the treatment assigned. This self-selection is a nonrandom process and may introduce bias between treatment received and response to treatment.

In particular, subject *i* will be in one of the nine (3^2) mutually exclusive compliance strata *P*: *partial-none-complier* (denoted by *pnc*); *full-none-complier* (*fnc*); *full-partial-complier* (*fpc*); *none-partial-defier* (*npd*); *none-full-defier* (*nfd*); *partial-full-defier* (*pfd*); *partial-taker* (*p*); *never-taker* (*n*); and *always-taker* (*a*). The naming convention is based on the amount of treatment taken if assigned to treatment (the first letter) and amount of treatment taken if assigned to the control group (the second letter) if

the amounts of treatment taken are assumed to differ under the two assignments. In other words, an always-taker, a never-taker and a partial-taker always receive the same amount (all, none, or partial, respectively) of experimental treatment regardless of whether or not treatment has been assigned. A defier (a *nfd*, a *pfd*, or a *npd*) always does the opposite, to various degrees, of what he/she is assigned to do. A complier (a *pnc*, a *fnc*, or a *fpc*) increases his/her exposure to experimental treatment to various degrees if he/she is assigned to the experimental group relative to the control group. To simplify our discussion, we define all three types of defiers (*npd*, *nfd* and *pfd*) as defiers (*d*); later, we assume that there are no defiers. The compliance strata can be summarized in Table 5-1:

Table 5-1 Compliance Stratum by Observed Exposure

Compliance stratification P	$[S_i(1), S_i(0)] =$	If	For
$P_i = n$	[0,0]	$S_i(z) = 0$	z = 1,0
$P_i = p$	[½,½]	$S_i(z) = ½$	z = 1,0
$P_i = a$	[1,1]	$S_i(z) = 1$	z = 1,0
$P_i = d$	[0,1], [½,1], [0, ½]	$S_i(1) < S_i(0)$	
$P_i = pnc$	[½,0]	$S_i(z) = z/2$	z = 1,0
$P_i = fnc$	[1,0]	$S_i(z) = z$	z = 1,0
$P_i = fpc$	[1,½]	$S_i(z) = ½ + z/2$	z = 1,0
$^{*}P_i = c$	[1,0], [1,½], [½, 0]	$S_i(1) > S_i(0)$	

$^{*}c$ is the combination of *pnc*, *fnc* and *fpc*.

For example, a *pnc* complies fully when he/she is assigned to the control group and takes only partial treatment when assigned to the experimental group; a *fnc* complies fully under both assignments; a *fpc* complies fully when he/she is assigned to the experimental group but takes partial treatment when assigned to the control. Finally, we

combine all three types of compliers (*pnc*, *fnc* and *fpc*) into one group and call them 'compliers (*c*).'

THEOREM 3: COMPLIANCE STRATUM *P* IS A PRINCIPAL STRATIFICATION.

Proof: Let treatment compliance indicator $S_i(z)$ be the post-treatment variable and $S_i(z)$ = 0, ½, and 1, where $i = 1, 2, \ldots, n$, and z = 0, 1. For each subject *i*, the value of the ordered pair $[S_i(1), S_i(0)]$ is fixed; that is to say, there is one and only one compliance stratification $P_i = k$ to which subject *i* belongs. And for each compliance stratification *k*, all units either have the same vector of $[S_i(1), S_i(0)]$ (for $k \in \{a, n, p, pnc, fnc, fpc\}$) or have unions of the same vector of $[S_i(1), S_i(0)]$ (for $k \in \{d, c\}$, e.g., $c = pnc \cup fnc \cup fpc$ or $[S_i(1), S_i(0)] = [1,0] \cup [1,½] \cup [½, 0]$).

DEFINITION 4**:** The full complier average causal effect (FCACE), the partial complier average causal effect (PCACE) and the full-partial-complier average causal effect (FPCACE). *The full-complier average causal effect is* $E[(Y_i(1,1) - Y_i(0,0)) \mid P_i = fnc]$, *the partial-complier average causal effect is* $E[(Y_i(1,½) - Y_i(0,0)) \mid P_i = pnc]$, and *the full-partial-complier average causal effect is* $E[(Y_i(1,1) - Y_i(0,½)) \mid P_i = fpc]$.

FCACE is the average difference in expected outcome between the two treatment groups in the subpopulation of full-none-compliers (we refer to them as 'full compliers'). Similarly, PCACE is the average difference in outcome between the two treatment groups in the subpopulation of partial-none-compliers. Lastly, FPCACE is the average of difference in outcome between the two treatment groups in the subpopulation of full-

partial-compliers. We refer to the three effects as compliance stratification effects. Since all three compliance stratification effects are defined as comparisons of potential outcomes under different treatment levels within a compliance stratum that are unaffected by treatment assignment, they are unconfounded treatment effects and may be interpreted causally.

THEOREM 4: COMPLIANCE STRATIFICATION EFFECTS ARE PRINCIPAL EFFECTS.

Proof: FCACE, PCACE and FPCACE are defined as the difference in two potential outcomes within the principal stratum P_i = *fnc*, *pnc*, and *fpc* respectively, or equivalently, with the same vector $[S_i(1), S_i(0)] = [1,0]$, [½, 0] and [1, ½]. Therefore, by definition, FCACE, PCACE and FPCACE are the principal effects with respect to treatment receipt $S_i(z)$, for z = 0, 1.

DEFINITION 5**:** REVISED CACE (RCACE under partial compliance). *The complier average causal effect is* $E[(Y_i(S_i(1),1) - Y_i(S_i(0), 0)) \mid S_i(1) > S_i(0)]$.

RCACE is the average difference in outcome between two treatment groups in the subpopulation of compliers: subjects who would have increased their exposure to the experimental treatment at least to a pre-defined degree had they been assigned to the experimental group. As noted in Table 5-1, there are three groups, i.e., $[S_i(1), S_i(0)] = [1,0]$, [1,½], or [½, 0], with $S_i(1) > S_i(0)$ or $S_i(1) - S_i(0) \geq ½$. These are three groups of

subjects who comply with assigned treatment to various degrees, and we refer to them as 'complier.'

THEOREM 5: REVISED CACE IS A PRINCIPAL EFFECT.

Proof: The RCACE is defined as the difference in two potential outcomes within the principal stratum $P_i = c$. Therefore, by definition, RCACE is the principal effect with respect to treatment receipt $S_i(z)$, for z = 1 and 0.

5.3. Statistical Models

As we have shown, adjustment for treatment non-compliance based on principal stratification methodology always generates causal effects, because it compares potential outcomes for a set of subjects with common compliance behaviour. However, the fundamental problem is that the principal stratum P_i to which a subject belongs cannot be observed directly, since for each subject *i*, only one of the pair $[S_i(1), S_i(0)]$ can be observed. As shown in Table 5-2, for every observed z_i and S_i combination, there can be two or three different potential values for P_i, assuming there are no defiers.

Table 5-2 Potential Compliance Strata Based on Observed Exposure

z_i	S_i^{obs}	Potential values of P_i
0	0	*pnc*, *fnc*, *n*
0	½	*fpc*, *p*
0	1	*a*
1	0	*n*
1	½	*p*, *pnc*
1	1	*a*, *fnc*, *fpc*

Inference about compliance stratification effects requires prediction of the subject's missing membership within the compliance strata, as determined by S_i^{mis}. To simplify the notation, we will drop *i* where it is obvious.

5.3.1 Dual Propensity and Counterfactual Propensity Scores

Since we cannot identify the compliance strata, or more generally, principal strata, we use the newly developed dual propensity scores to address the issue of identification problems. We define the dual propensity score in an RCT setting with two treatment assignments and in the presence of all-or-none treatment compliance. Two propensity scores are calculated under both possible assignments for all subjects. For detailed theory of dual propensity score methods, please refer to Chapter 4.

Recall that $S(1)$ and $S(0)$ cannot both be observed at once. We observe treatment under the actual assignment, $S^* = S(z)$. We do not observe treatment under the alternative assignment, $S^\# = S(1 - z)$. Consequently, only one of the dual propensity scores is observable. We call the observed one the 'actual' propensity score, and the unobservable one the 'counterfactual' propensity score.

5.3.2 Compliance Stratification Effects through Matching

There are three commonly used propensity score methods: covariate adjustment, stratification or subclassification and matching (D'Agostino, 1998). Stratification and covariate adjustment have been discussed in the previous chapter; in this chapter we focus our discussion on matching.

For each subject in the experimental group, we find a matched subject in the control group who is likely to belong to the same compliance stratum, based on the

estimated DPS $\hat{r} = [\hat{r}^1, \hat{r}^0]$. The matched pairs are then classified into the compliance strata defined in Section 3 based on their observed exposure levels under both the experimental and control conditions. The advantage is that instead of classifying each individual into compliance strata, we classify each matched pair of subjects into compliance strata, with potential outcomes available under both conditions. Compliance stratification effects are estimated based on these newly identified strata.

Assuming that the propensity score is known for each subject, and for each subject *i* under z =1, there exists *i*' as its matched pair under z = 0. From Definition 4,

$$\begin{aligned}\text{FCACE} &= \text{E}[(Y_i(1,1) - Y_i(0,0)) \mid P_i = fnc] \\ &= \text{E}[(Y_i(1,1) - Y_i(0,0)) \mid [S_i(1), S_i(0)] = [1, 0]] \\ &= \text{E}[(\text{Y}_i^*(1) - \text{Y}_{i'}^{\#}(0)) \mid [\text{S}_i^*, \text{S}_i^{\#}] = [1, 0]] \\ &= \text{E}[(\text{Y}_i^*(1) - \text{Y}_{i'}(0)) \mid [\text{S}_i^*, \text{S}_{i'}] = [1, 0]]\end{aligned}$$

where $\text{Y}_i^*(1)$ is the observed outcome when $\text{S}_i(1) = 1$, $\text{Y}_i^{\#}(0)$ is the missing outcome when $\text{S}_i(0) = 0$ (which is not observable under z = 0), and $\text{Y}_{i'}(0)$ is the outcome from the matched subject under z = 0 with $\text{S}_{i'} = 0$.

Similarly, from Definition 4,

$$\begin{aligned}\text{PCACE} &= \text{E}[(Y_i(½,1) - Y_i(0,0)) \mid P_i = pnc] \\ &= \text{E}[(Y_i(½,1) - Y_i(0,0)) \mid [S_i(1), S_i(0)] = [½, 0]] \\ &= \text{E}[(\text{Y}_i^*(½) - \text{Y}_{i'}^{\#}(0)) \mid [\text{S}_i^*, \text{S}_i^{\#}] = [½, 0]] \\ &= \text{E}[(\text{Y}_i^*(½) - \text{Y}_{i'}(0)) \mid [\text{S}_i^*, \text{S}_{i'}] = [½, 0]]\end{aligned}$$

where $\text{Y}_i^*(½)$ is the observed outcome when $\text{S}_i(1) = ½$, $\text{Y}_i^{\#}(0)$ is the missing outcome when $\text{S}_i(0) = 0$ (which is not observable under z = 0), and $\text{Y}_{i'}(0)$ is the outcome from the matched subject under z = 0 with $\text{S}_{i'} = 0$.

$$\text{FPCACE} = E[(Y_i(1,1) - Y_i(½,0)) \mid P_i = fpc]$$

$$= E[(Y_i(1,1) - Y_i(½,0)) \mid [S_i(1), S_i(0)] = [1, ½]]$$

$$= E[(Y_i^*(1) - Y_i^{\#}(½)) \mid [S_i^*, S_i^{\#}] = [1, ½]]$$

$$= E[(Y_i^*(1) - Y_{i'}(½)) \mid [S_i^*, S_{i'}] = [1, ½]]$$

where $Y_i^*(1)$ is the observed outcome when $S_i(1) = 1$, $Y_i^{\#}(½)$ is the missing outcome when $S_i(0) = ½$ (which is not observable under z = 0), and $Y_{i'}(½)$ is the outcome from the matched subject under z = 0 with $S_{i'} = ½$.

Lastly, from Definition 5,

$$\text{RCACE} = E[(Y_i(S_i(1),1) - Y_i(S_i(0),0)) \mid P_i = c]$$

$$= E[(Y_i(S_i(1),1) - Y_i(S_i(1),0)) \mid S_i(1) > S_i(0)]]$$

$$= E[(Y_i^*(S_i^*) - Y_i^{\#}(S_i^{\#})) \mid S_i^* > S_i^{\#}]$$

$$= E[(Y_i^*(S_i^*) - Y_{i'}(S_{i'})) \mid S_i^* > S_{i'}]$$

where $Y_i^*(S_i^*)$ is the observed outcome when $S_i(1) = S_i^*$ ($S_i^* = 1$ or ½), $Y_i^{\#}(S_i^{\#})$ is the missing outcome when $S_i(0) = S_i^{\#}$($S_i^{\#} = ½$ or 0) (which is not observable under z = 0), and $Y_{i'}(S_{i'})$ is the outcome from the matched subject under z = 0 with $S_i^* > S_{i'}$.

It is easy to show that RCACE is actually a weighted average of FCACE, PCACE and FPCACE, since the three conditions with $S_i^* > S_{i'}$ are $[S_i^*, S_{i'}] = [1,0] \cup [½, 0] \cup [1,½]$, which are partitions for FCACE, PCACE and FPCACE, respectively.

$$\text{RCACE} = E[(Y_i^*(S_i^*) - Y_{i'}(S_{i'})) \mid S_i^* > S_{i'}]$$

$$= E[(Y_i^*(S_i^*) - Y_{i'}(S_{i'})) \mid [S_i^*, S_{i'}] = [1,0] \cup [½, 0] \cup [1,½]]$$

$$= E\{E[(Y_i^*(1) - Y_{i'}(0)) \mid [S_i^*, S_{i'}] = [1, 0]]$$

$$+ E[(Y_i^*(½) - Y_{i'}(0)) \mid [S_i^*, S_{i'}] = [½, 0]]$$

$$+ E[(Y_i^*(1) - Y_{i'}(½)) \mid [S_i^*, S_{i'}] = [1, ½]]\}$$

$$= E\{FCACE + PCACE + FPCACE\}$$

5.3.3 Dual Propensity Score Matching Algorithm

In the next three sections, we assume that the dual propensity scores are known. For the dual propensity score matching and counterfactual propensity score matching, we focus on one-to-one matching without replacement to match each 'compliant' (e.g., $S_i(1) = 1$ or ½) subject in the experimental group to a single subject in the control group. If more than one subject has been identified as a match with the same distance, only one is selected at random. The subjects left unmatched are dropped from the analysis.

Many different matching algorithms have been published, including nearest-neighbour (NN) matching, caliper and radius matching, stratification and interval matching, and weighting. The most straightforward matching estimator is NN matching where an 'untreated' subject is chosen as a matching partner for each and every 'treated' subject who is closest in terms of matching variables.

In our case of DPS matching, for each subject i in the experimental group, NN searches for the subject i' in the control group with the closest distance $D_{ii'}$. The $D_{ii'}$ can be defined in multiple ways. Common choices include the Euclidean distance and the weighted sum of the absolute differences between the DPS. The two-dimensional Euclidean distance can be defined as the following expression:

$$D_{ii'} = \left(i' = j \mid \min_{j \in J} \sqrt{(r_i^1 - r_j^1)^2 + (r_i^0 - r_j^0)^2} \right) \qquad [5.3.1]$$

where *i'* the subject matched with subject *i* from a set of controls *J*. When an exact match has been found for *i*, then $D_{ii'} = 0$.

The $D_{ii'}$ can be calculated by matching with or without replacement. In matching with replacement, a subject in the control group can be used more than once as a match, whereas in the matching without replacement, a subject is considered only once; that is, if a subject *j* is chosen as a match it will be removed from the set *J*. Matching with or without replacement involves a trade-off between bias and variance. It is believed that matching with replacement improves the average quality of matching while increasing the variance of the estimator; on the other hand, matching without replacement may lead to more poor matches but increase the number of distinct subjects in the control group used to construct the counterfactual outcome (Smith and Todd, 2005). One way to improve the quality of matches while using matching without replacement is to impose a tolerance level on the maximum propensity score distance (caliper). Propensity score calipers are discussed in Rosenbaum and Rubin (1985b) and Rosenbaum (1989). Imposing a caliper works similarly as allowing for replacement; bad matches are avoided and hence the matching quality rises. Formally, the caliper matching selects the nearest neighbour within a caliper of width δ and can be stated as follows:

$$D_{ii'} = \left(i' = j \mid \min_{j \in J} \sqrt{(r_i^1 - r_j^1)^2 + (r_i^0 - r_j^0)^2} \;\&\; \sqrt{(r_i^1 - r_j^1)^2 + (r_i^0 - r_j^0)^2} < \delta \right) \qquad [5.3.2]$$

As Smith and Todd (2005) note, a possible drawback of caliper matching is that it is difficult to know a priori what choice of the tolerance level (δ) is reasonable. Another drawback is the loss of many unmatched subjects from analysis.

5.3.4 Counterfactual Propensity Score Matching Algorithm

As a sensitivity analysis, we use the one-to-one NN matching without replacement to seek the nearest match based on the absolute differences $D_{ii'}$ between the counterfactual propensity score:

$$D_{ii'} = \left(i' = j \mid \min_{j \in J} (\mid r_i^{\#} - r_j^{*} \mid) \right) \qquad [5.3.3]$$

We implement a matching algorithm as the following: for each subject *i* in the experimental (or control) group, the algorithm searches for the exact match from the control (or experimental) group with the same CPS and then randomly selects one match. Once a match is made, the match is not reconsidered. The algorithm starts with the first five digits of the CPS. If an appropriate control cannot be selected, then a four-digit match on the CPS is attempted. If an appropriate match cannot be formed on the first four digits, then a three-digit match is attempted. This process is repeated until matches are attempted on the first two digits. If subject *i* cannot be matched to any control subject, then the subject is left unmatched and dropped from the analysis.

5.3.5 Stratification Algorithm

The idea of stratification matching is to partition the DPS into a set of intervals (strata). This method is also known as interval matching, blocking, and subclassification (Rosenbaum and Rubin, 1983). The detailed algorithm can be found in Section 4.3.3.

5.4. Estimation

5.4.1 Ordinal Logistic Regression

We propose using the ordinal logistic regression model to estimate the propensity to receive the ordered exposure of treatment. The ordinal logistical regression model is a natural extension of binary logistic regression; it models the cumulative logit of the probability (Agresti 2002) and uses maximum likelihood methods to estimate summary odds ratios. There are two popular forms of the ordinal logistic regression model: the proportional odds (PO) model and the continuation ratio (CR) model. Both the PO and CR ordinal regression models are linear and additive on the logit scale, and both are estimated using maximum likelihood methods. The PO model is sometimes referred to as the 'ordinal logistic' model; it is also referred to as a 'cumulative odds' model because it is defined by the log odds of the cumulative probabilities.

We model an ordinal exposure (3 levels) using the proportional odds (PO) form of an ordinal logistic model to predict the exposure of treatment using baseline covariates. Using the same notation in Section 5.3 and 5.4, let S^* be the observed ordinal exposure or doses with three possible categories (0, ½, 1), which correspond to three exposure levels (none, partial and full). For such an ordinal logistic model, there are two cumulative logits in descending order:

$$\log\left[\frac{P(S_i = 1)}{1 - P(S_i = 1)}\right] = \log\left[\frac{P(S_i = 1)}{P(S_i = 0,½)}\right] \qquad [5.4.1]$$

$$\log\left[\frac{P(S_i \geq ½)}{1 - P(S_i \geq ½)}\right] = \log\left[\frac{P(S_i = ½,1)}{P(S_i = 0)}\right] \qquad [5.4.2]$$

Or in the reverse order:

$$\log\left[\frac{P(S_i = 0)}{1 - P(S_i = 0)}\right] = \log\left[\frac{P(S_i = 0)}{P(S_i = ½,1)}\right] \qquad [5.4.3]$$

$$\log\left[\frac{P(S_i \leq ½)}{1 - P(S_i \leq ½)}\right] = \log\left[\frac{P(S_i = 0,½)}{P(S_i = 1)}\right] \qquad [5.4.4]$$

The model is formulated under z = 1 as

$$\log\left[\frac{P(S_i \geq j)}{1 - P(S_i \geq j)}\right] = X^T \beta_{z=1} \qquad \text{for } j = 1 \text{ and } ½ \qquad [5.4.5]$$

where **X** is a vector of predictors.

The model is formulated under z = 0 as

$$\log\left[\frac{P(S_i \leq j)}{1 - P(S_i \leq j)}\right] = X^T \beta_{z=0} \qquad \text{for } j = 0 \text{ and } ½ \qquad [5.4.6]$$

where **X** is a vector of predictors.

There is an implicit assumption that the regression coefficients β are independent of *j*, the cut-off level for S. A homogeneity of effect across 'cut-points' is assumed, with a single odds ratio summarizing the effect of interest over all cut-points. The assumption may be informally tested by fitting separate logistic models to see whether the coefficients look similar or very different. It is common practice in a 'single' propensity score approach to include all available covariates that might affect the exposure as predictors, with the objective of capturing the exposure propensity precisely (Austin, 2007; D'Agostino, 1998). Others argue that one should instead concentrate only on those variables with a large impact on both the exposure and the outcome under scrutiny (Rubin, 1997; Brookhart et al., 2006). However, consistent estimation of the propensity score might require including variables that affect treatment receipt but have little if any

affect on the outcome. In this chapter, covariates with a large impact on the exposure were selected, as well as some of their interactions (Austin 2007).

5.4.2 Estimating Dual Propensity Scores Using the Ordinal Logistic Model

Recall that dual propensity scores are $r^1(\mathbf{X}) = \Pr[S(1) = 1 \mid \mathbf{X}, z = 1]$ and $r^0(\mathbf{X}) = \Pr[S(0) = 0 \mid \mathbf{X}, z = 0]$. As proposed in the previous section, we use the ordinal logistic regression model to predict the propensity for ordinal treatment categories. The parameter estimates from the models are used to calculate the DPS for each subject. Specifically, we model $S \geq j$ under $z = 1$ and $S \leq j$ under $z = 0$ to predict the exposure under both treatment assignments using baseline covariates. Using the fitted models, we predict binary events $S = 1$ under $z = 1$ and $S = 0$ under $z = 0$ with the corresponding predicted probabilities

$$\hat{r}_i^1 = \frac{1}{1 + \exp[-X_i{}^T \hat{\beta}_{z=1}]} \qquad [5.4.7]$$

and

$$\hat{r}_i^0 = \frac{1}{1 + \exp[-X_i{}^T \hat{\beta}_{z=0}]}. \qquad [5.4.8]$$

5.4.3 Estimating Principal Compliance Effects Based on DPS Matching

We implement a caliper matching based on the Euclidean distance, with a caliper with a width of 0.6 of the larger standard deviation of the dual propensity scores, i.e., $\delta = 0.6*(\max(\text{std}(r^1), \text{std}(r^1))$ (Austin, 2009). If no member of the control group falls within the caliper for subject *i*, then the subject is left unmatched and dropped from the analysis. We follow the steps shown below:

1. Estimate $\beta_{z=1}$ as in (5.4.5) and $\beta_{z=0}$ as in (5.4.6).
2. Estimate the dual propensity scores $\hat{r}^1(X_i, \hat{\beta}_{z=1})$ and $\hat{r}^0(X_i, \hat{\beta}_{z=0})$ for all subjects.

3. For subject *i*, calculate the D_{ij} as defined in 3.2 for all subjects in the control group *J*.
4. Find the match *i*' with smallest D_{ij}, then drop the subject *i*' from the set of controls *J*. If more than one match is found, randomly select one.
5. Repeat steps 3-4 for each and every subject with S* = 1 (or ≥ ½) under z =1.
6. Classify the matched pairs into compliance strata as defined in Table 5-2.
7. Compute differences between the two treatment groups conditional on the newly identified subgroups, i.e., $[S_i^*, S_{i'}] = [1, 0]$.

5.4.4 Estimating Principal Compliance Effects Based on CPS Matching

We implement a matching algorithm (from 5 digits to 2 digits, see Section 5.3.4) to seek the nearest match based on the absolute differences of the counterfactual propensity score. We start with the first 5 digits of the CPS, then the first 4 digits of the CPS, then the first 3 digits of the CPS, and finally the first 2 digits of the CPS. Subject *i* from the experimental group is matched with a subject from the control group based on $\hat{r}^0(X_i, \hat{\beta}_{z=0})$. We implement the same algorithm to match subjects from the control group to subjects from the experimental group based on $\hat{r}^1(X_i, \hat{\beta}_{z=0})$. The steps are exactly the same. We show only steps for the CPS based on $\hat{r}^0(X_i, \hat{\beta}_{z=0})$ as below:

1. Estimate $\beta_{z=1}$ as in (5.4.7) and $\beta_{z=0}$ as in (5.4.8).
2. Estimate the counterfactual propensity score $\hat{r}^0(X_i, \hat{\beta}_{z=0})$ for all subjects.
3. Sort subjects with S* = 1 (or ≥ ½) in the experimental group by $\hat{r}^0(X_i, \hat{\beta}_{z=0})$.
4. Sort all subjects in the control group by $\hat{r}^0(X_i, \hat{\beta}_{z=0})$.
5. For subject *i*, search the exact match from the control set *J* with the same CPS.

6. Find the exact match *i*' and drop the subject *i*' from the set of controls *J*. If more than one match is found, select one randomly.
7. Repeat steps 5-6 for each subject with S* = 1 (or ≥ ½) under Z =1.
8. Classify the matched pairs into compliance strata as defined in Table 5-2
9. Compute differences between the two treatment groups conditional on the newly identified subgroups, i.e., $[S_i^*, S_{i'}] = [1, 0]$

5.4.5 Estimating Principal Compliance Effects Based on DPS Stratification

The detailed estimation procedure can be found in Section 4.3.4. Subjects are classified into compliance strata based on their estimated DPS. Once the strata are defined, the treatment effect is evaluated by comparing subjects directly between the two treatment groups within each stratum. Then, the mean of the differences across the strata is summarized using three different weighting schemes.

5.4.6 Estimating Standard Errors

Computing the standard error for the estimator of the causal effect is not straightforward. The estimated variance of the compliance stratification effects should also include the variability due to the estimation of the propensity scores. These estimation steps add variation beyond normal sampling variation (Heckman, Ichimura and Todd, 1998). Bootstrapping can be a useful technique for estimating standard errors where analytical estimates are biased or unavailable (Efron and Tibshirani, 1998). The basic strategy follows these steps: 1) sample with replacement *N* records from the identified subpopulation, where *N* is the number of units in the analysis data set; 2) calculate and save the statistic of interest, say $\hat{\Delta}$, in this sample; and 3) repeat steps 1 and

2 many times (usually 1000 times, we used 500 repetitions due to the large sample size). This produces an empirical distribution for $\hat{\Delta}$, which approximates the sampling distribution and thus standard error of the population mean of the parameter of interest.

5.5. Application: PROBIT

The Breastfeeding Promotion Intervention Trial (PROBIT) was conducted in the Republic of Belarus. Details of the study methods are outlined in Kramer et al. (2001), and related analyses (summarized in Chapter 3) have been reported in Kramer, Guo, Platt et al. (2002, 2003, and 2004) and elsewhere. The experimental intervention was successful in prolonging the duration of any breastfeeding and in increasing the degrees of breastfeeding. Infants were classified as exclusively breastfed at 3 or 6 months if the cross-sectional feeding information obtained at the first 3 or 6 months indicated that no liquid or solid foods other than breast milk were being administered to the infant.

Regardless of the assignment to the experimental group or control group, mothers' breastfeeding behaviour can be classified into three mutually exclusive and clinically distinct categories: 1) Early weaners (EW): mothers who stopped breastfeeding during the first 3 months; 2) Partial/mixed breastfeeders (MBF): mothers who breastfed for ≥ 3 months but either failed to exclusively breastfeed during the first 3 months or failed to continued breastfeeding for at least 6 months; 3) Full breastfeeders (FBF): mothers who exclusively breastfed for the first 3 months and continued breastfeeding for the first 6 months. We refer to this group as prolonged and exclusive breastfeeders.

Table 5-3 Frequency of Breastfeeding Behaviour

Number of Subjects n (%)	Experimental (n = 8865)	Control (n = 8181)
Early weaner	2632 (29.7)	3313 (40.5)
Mixed breastfeeder	3161 (35.7)	4460 (54.5)
Full breastfeeder	3072 (34.7)	408 (5.0)

We consider exclusive breastfeeding for at least 3 months and continued breastfeeding for at least 6 months to be the measure of 'compliance' in the experimental arm, and weaning during the first 3 months as the measure of 'compliance' in the control arm. As the subjects with high probability of doing both are more likely to be 'compliers,' we try to identify these individuals.

5.5.1 Results

Two propensity scores are estimated: one under the experimental arm and one under the control arm, using ordinal logistic regression models. Covariates with a large impact on treatment receipt were selected, as well as some of their interactions. Covariates selected include region (West versus East and urban versus rural), maternal age (<20, 20-34, and >=35 years), maternal education (incomplete secondary, complete secondary, partial university, and complete university), prior history of having breastfed an infant for >=3 months (yes/no), caesarean delivery (yes/no), maternal smoking during pregnancy (yes/no), other children living in the household (0,1, >=2), gender (male versus female), gestational age completed (weeks), birth weight (g), birth length (cm) and birth head circumference (cm). Covariates were initially included into the models as main effects, and then interaction terms between variables having shown a significant effect (p-value < .05, confirmed by the stepwise model selection procedure) and the remaining variables were included as well.

5.5.1.1 FCACE of Prolonged and Exclusive BF on Infant Growth

In this section, DPS have been estimated as predicted probabilities of binary events S =1 under z =1 and S = 0 under z = 0. Correspondingly, FCACE is estimated using the different approaches described in the previous sections. Table 5-4 shows the frequency distribution and means and standard deviations of the estimated DPS by observed breastfeeding behaviour under actual treatment assignment. The most frequent scores fall between 0.2 and 0.5, with a trend that the higher the estimated score percentile, the higher the proportion of the observed 'compliers' (i.e., $S^* = 1$ under the experimental group and $S^* = 0$ under the control group). The last row shows the overall

means and standard deviations of the estimated DPS and the predicted counterfactual DPS. These statistics are almost identical (i.e., $\hat{r}^1(z=1) = \hat{r}^1(z=0) = 0.35$, $\hat{r}^0(z=0) = 0.40$ and $\hat{r}^0(z=1) = 0.41$), confirming our belief that, owing to randomization, subjects in two groups are exchangeable in terms of DPS.

Table 5-4 Frequency Distribution of DPS Percentile and Summary Statistics

Frequency	r^1 in Experimental Group (n = 8865)			r^0 in Control Group (n = 8181)		
	$S^*=0$ (2630)[2]	$S^*=½$ (3161)	$S^*=1$ (3072)	$S^*=0$ (3313)	$S^*=½$ (4460)	$S^*=1$ (408)
$0.0 \le \hat{r} < 0.1$	64	14	6	1	8	3
$0.1 \le \hat{r} < 0.2$	385	228	113	92	389	88
$0.2 \le \hat{r} < 0.3$	852	848	625	407	989	112
$0.3 \le \hat{r} < 0.4$	904	1178	1087	724	1244	122
$0.4 \le \hat{r} < 0.5$	321	591	756	862	1003	61
$0.5 \le \hat{r} < 0.6$	89	242	377	717	571	16
$0.6 \le \hat{r} < 0.7$	14	51	95	362	204	6
$0.7 \le \hat{r} < 0.8$	1	8	12	107	39	0
$0.8 \le \hat{r} < 0.9$	0	1	1	38	13	0
$0.9 \le \hat{r} < 1.0$	0	0	0	3	0	0
Mean $\hat{r}$ (std)[1]	0.30 (.107)	0.34 (.109)	0.38 (.113)	0.45 (0.142)	0.38 (.134)	0.30 (.120)
Overall Mean $\hat{r}$ (std)[1]	0.35 (.115) $r^1(z=0)$: 0.35 (.114)			0.40 (.143) $r^0(z=1)$: 0.41 (.138)		

[1] std = standard deviation [2] 2 missing values of r^1

Figures 5-1 and 5-2 show the scatter plots of $\hat{r}^1$ versus $\hat{r}^0$ by treatment groups.

Figure 5-1 Distribution of Dual Propensity Scores - Experimental Group

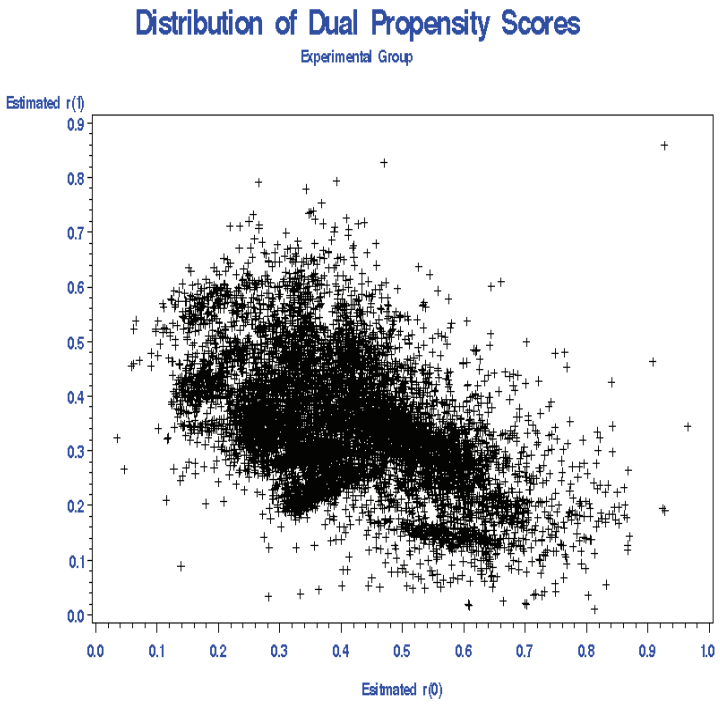
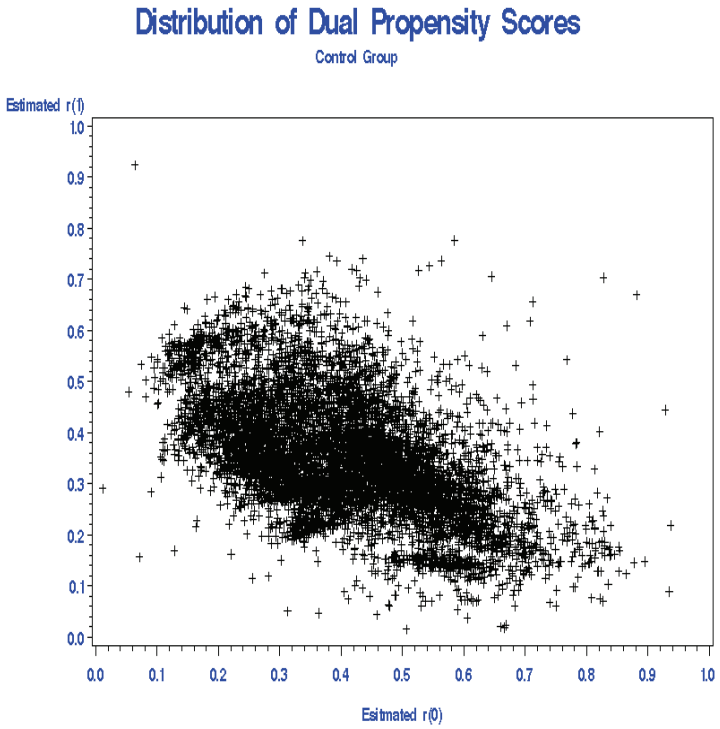


Figure 5-2 Distribution of Dual Propensity Scores - Control Group



DPS caliper matching was performed using Euclidean distance, as described in Section 5.3.3, with a caliper with a width of $\delta = 0.86$ (calculated as 0.6*(max {0.115, 0.143})). Each of the 3,072 full breastfeeders in the experimental group was matched with one subject in the control group with the smallest $D_{ii'}$. Only one pair was dropped from the analysis because their calculated distance $D_{ii'}$ was greater than the caliper of width δ (0.86). Among 3,071 matched controls, 1,190 (38.8 %) stopped breastfeeding prior to 3 months ('compliers' in the control group). Thus, these 1,190 matched-pairs were identified as the full compliers - the subpopulation on which FCACE is defined. Table 5-5 shows the baseline comparison of these full compliers. The two matched groups are comparable in terms of baseline characteristics. The last row also shows the summary statistics of the calculated Euclidean distance $D_{ii'}$. The mean is 0.0035 and the median is 0.0026.

Table 5-5 Baseline Characteristics for the Full Compliers

Variable	Experimental N = 1190	Control N = 1190
Hospital region		
East Belarus (urban)	492 (41.3)	448 (37.7)
East Belarus (rural)	211 (17.7)	260 (21.9)
West Belarus (urban)	268 (22.5)	196 (16.5)
West Belarus (rural)	219 (18.4)	286 (24.0)
Maternal age (yr)		
<20	179 (15.0)	182 (15.3)
20-34	950 (79.8)	952 (80.0)
≥35	61 (5.1)	56 (4.7)
Maternal education		
Incomplete secondary	53 (4.5)	34 (2.9)
Complete secondary	421 (35.4)	371 (31.2)
Advanced secondary or Partial university	535 (45.0)	625 (52.5)
Complete university	181 (15.2)	160 (13.5)
Prior breastfeeding history	297 (25.0)	264 (22.2)
Caesarean	161 (13.5)	123 (10.3)
Maternal smoking during pregnancy	24 (2.0)	22 (1.9)
Number of other children In household		
0	709 (59.6)	704 (59.2)
1	356 (29.9)	370 (31.1)
≥2	125 (10.5)	116 (9.8)
Male sex	610 (51.3)	615 (51.7)
Gestational age (wk)	39.4	39.3
Birth weight (g)	3405	3414
Birth length (cm)	51.9	52.1
Birth head circumference (cm)	35.2	34.9
Summary statistics for $D_{ii'}$	Mean:0.0035; STD: 0.0037; Median: 0.0026; Max: 0.0377; Min: 0.0	

CPS matching was also performed using the algorithm described in Section 5.3.4. The matching procedure was performed in two ways: matching each full breastfeeder in the experimental group to one subject in the control group (CPS Matching E), or matching each earlier weaner in the control group to one subject in the experimental

group (CPS Matching C). A total of 1,146 matched pairs resulted from CPS Matching E (Two subjects were unmatched and dropped from the analysis), and 1,072 matched pairs from CPS Matching C (Twenty subjects were unmatched and dropped from the analysis).

Table 5-6 and Table 5-7 show the effects of prolonged and exclusive breastfeeding compared to early weaning on infant weight and length gains respectively, through the first 12 months of life. The first column is ITT effect of the experimental intervention, while the next three columns are the FCACEs based on the full compliers identified following the matching procedures described in Sections 5.4.3 and 5.4.4.

Table 5-6 Effect of Prolonged and Exclusive BF on Weight Gain (g) through 12 Months

Time	ITT	FCACE Δ (SE*)		
		DPS Caliper Matching	CPS Matching E	CPS Matching C
1 m	61	153 (14.8)	172 (15.6)	132 (11.0)
2 m	88	209 (18.0)	218 (18.5)	186 (12.9)
3 m	106	161 (22.8)	194 (25.0)	168 (15.5)
6 m	89	15 (31.7)	5 (30.6)	10 (22.0)
9 m	58	-86 (35.3)	-92 (34.6)	-46 (24.8)
12 m	-7	-164 (38.5)	-145 (37.4)	-104 (25.9)

*SE: bootstrap standard errors.

Table 5-7 Effect of Prolonged and Exclusive BF on Length Gain (cm) through 12 Months

Time	ITT	FCACE Δ (SE*)		
		DPS Caliper Matching	CPS Matching E	CPS Matching C
1 m	0.16	0.22 (0.07)	0.28 (0.06)	0.21(0.04)
2 m	0.32	0.31 (0.08)	0.41 (0.07)	0.21 (0.05)
3 m	0.50	0.26 (0.09)	0.40 (0.09)	0.29 (0.06)
6 m	0.46	0.03 (0.10)	0.10 (0.11)	-0.08 (0.08)
9 m	0.31	-0.23 (0.11)	-0.16 (0.11)	-0.31 (0.07)
12 m	0.18	-0.18 (0.11)	-0.23 (0.11)	-0.26 (0.07)

*SE: bootstrap standard errors.

5.5.1.2 RCACE of Prolonged and Exclusive BF on Infant Growth

In this section, the DPS have been estimated as predicted probabilities of binary events $S \geq ½$ under $z = 1$ and $S = 0$ under $z = 0$. Since more than 95 % of subjects have $S \leq ½$ under $z = 0$, we decided to model $S = 0$ instead. Correspondingly, RCACE is estimated in the same way that the FCACE is estimated. Table 5-8 shows the frequency distribution and means and standard deviations of the estimated DPS by observed breastfeeding behaviour under actual treatment assignment. The last row shows the overall means and standard deviations of the estimated DPS and the predicted counterfactual DPS. Notice that the numbers for r^0 in the control group are the same as in Table 5-4.

Table 5-8 Frequency of DPS Percentile and Summary Statistics

Frequency	r^1 in Experimental Group (n = 8865)			r^0 in Control Group (n = 8181)		
	$S^* = 0$ (2630) [2]	$S^* = ½$ (3161)	$S^* = 1$ (3072)	$S^* = 0$ (3313)	$S^* = ½$ (4460)	$S^* = 1$ (408)
$0.0 \le \hat{r} < 0.1$	4	1	0	1	8	3
$0.1 \le \hat{r} < 0.2$	10	1	0	92	389	88
$0.2 \le \hat{r} < 0.3$	28	4	4	407	989	112
$0.3 \le \hat{r} < 0.4$	48	27	5	724	1244	122
$0.4 \le \hat{r} < 0.5$	234	103	64	862	1003	61
$0.5 \le \hat{r} < 0.6$	345	275	166	717	571	16
$0.6 \le \hat{r} < 0.7$	858	954	729	362	204	6
$0.7 \le \hat{r} < 0.8$	867	1241	1267	107	39	0
$0.8 \le \hat{r} < 0.9$	233	532	792	38	13	0
$0.9 \le \hat{r} < 1.0$	3	23	45	3	0	0
Mean $\hat{r}$ (std)[1]	0.66 (.125)	0.71 (.102)	0.74 (.096)	0.45 (0.142)	0.38 (.134)	0.30 (.120)
Overall Mean $\hat{r}$ (std)[1]	0.71 (.108) $r^1(z = 0)$: 0.70 (.112)			0.40 (.143) $r^0(z = 1)$: 0.41 (.138)		

[1] std = standard deviation [2] 2 missing values of r^1

Figures 5-3 and 5-4 show the scatter plots of $\hat{r}^1$ versus $\hat{r}^0$ by treatment groups.

Figure 5-3 Distribution of Dual Propensity Scores - Experimental Group

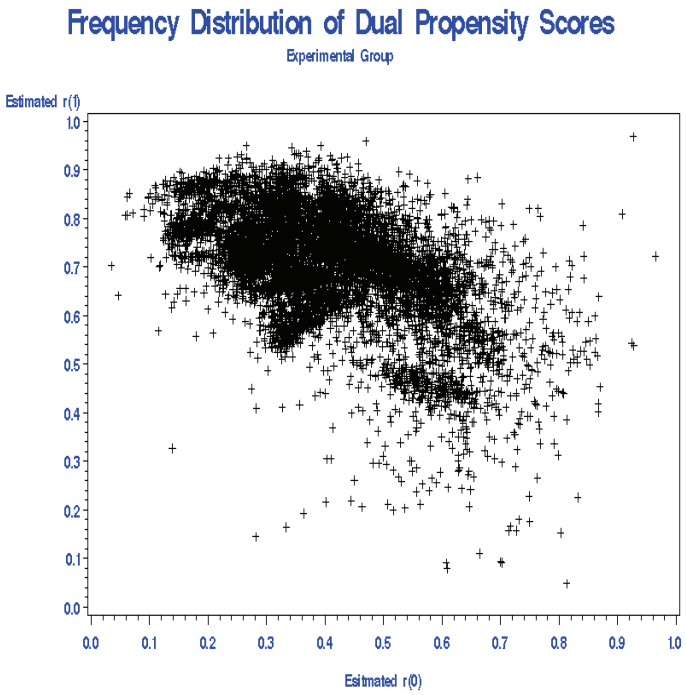
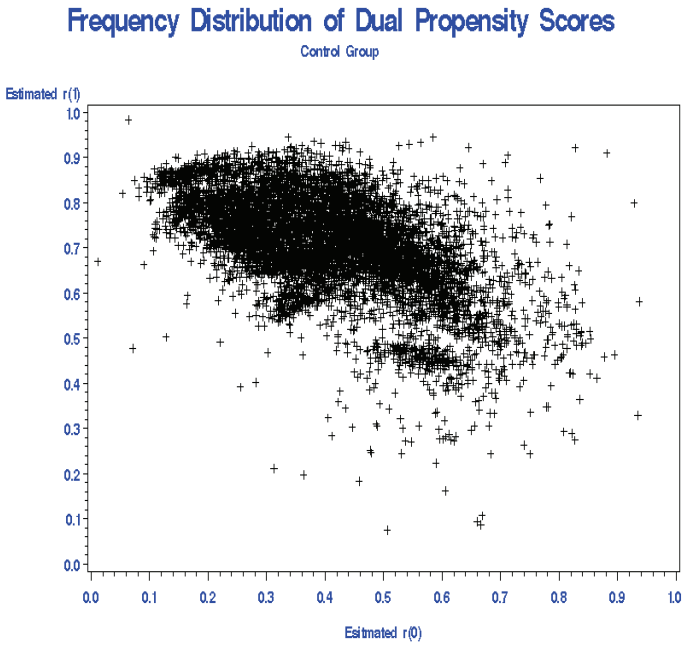


Figure 5-4 Distribution of Dual Propensity Scores - Experimental Group



DPS caliper matching was performed using a tolerance level of $\delta = 0.86$. Each of the 6,233 full and mixed breastfeeders in the experimental group was matched with one subject in the control group with the smallest $D_{ii'}$. Two pairs were dropped because their calculated distance $D_{ii'}$ was greater than the tolerance level of δ (0.86). Among the 6,233 matched controls, 2,491 (40.0 %) discontinued breastfeeding prior to 3 months ('compliers' in the control group). Thus, these 2,491 matched pairs were identified as the compliers: subjects who would have been early weaners had they been assigned to the control group and would have been full or mixed breastfeeders had they been assigned to the experimental group, i.e., the combination of the two subpopulations *fnc* and *pnc*. Table 5-9 shows the baseline similarity of these compliers.

Table 5-9 Baseline Characteristics for the Full Compliers

Variable	Experimental N = 2491	Control N = 2491
Hospital region		
East Belarus (urban)	1053 (42.3)	914 (36.7)
East Belarus (rural)	365 (14.7)	519 (20.8)
West Belarus (urban)	652 (26.2)	469 (18.8)
West Belarus (rural)	421 (16.9)	589 (23.7)
Maternal age (yr)		
<20	367 (14.7)	394 (15.8)
20-34	2000 (80.3)	1977 (79.4)
≥35	124 (5.0)	120 (4.8)
Maternal education		
Incomplete secondary	112 (4.5)	86 (3.5)
Complete secondary	859 (34.5)	810 (32.5)
Advanced secondary or Partial university	1169 (46.9)	1257 (50.5)
Complete university	351 (14.1)	338 (13.6)
Prior breastfeeding history	520 (20.9)	485 (19.5)
Cesarean	321 (12.9)	278 (11.2)
Maternal smoking during pregnancy	76 (3.1)	48 (1.9)
Number of other children In household		
0	1524 (61.2)	1457 (58.5)
1	746 (30.0)	796 (32.0)
≥2	221 (8.9)	238 (9.6)
Male sex	1305 (52.4)	1289 (51.8)
Gestational age (wk)	39.4	39.3
Birth weight (g)	3415	3412
Birth length (cm)	51.8	52.0
Birth head circumference (cm)	35.2	34.8
Summary statistics for $D_{ii'}$	Mean:0.0050; STD: 0.0069; Median: 0.0031; Max: 0.0846; Min: 0.0	

Tables 5-10 and 5-11 show the effects of breastfeeding (full and mixed) on infant weight and length gains respectively, compared to early weaning through the first 12 months of life. The first column is the intent-to-treat causal effect of the experimental intervention. The next three columns are the RCACEs based on the compliers identified

using the previously described matching procedures. A total of 2,456 and 2,243 matched pairs are identified as compliers from CPS Matching E and C respectively. The last three columns are the RCACEs based on stratification.

Table 5-10 Effect of BF on Weight Gain (g) through 12 Months

Time	ITT	RCACE Δ (SE) (by Matching and Stratification)					
		DPS Caliper	CPS Matching E	CPS Matching C	Equal Weight	Weight by Proportion	Weight by Strata
1 m	61	124 (11.0)	135 (10.9)	126 (11.3)	68 (10.3)	68 (7.0)	88 (12.7)
2 m	88	158 (12.9)	171 (13.3)	173 (14.1)	94 (11.6)	93 (7.9)	112 (15.5)
3 m	106	133 (15.5)	154 (17.1)	157 (16.8)	114 (13.5)	112 (9.5)	118 (18.1)
6 m	89	16 (22.0)	35 (23.1)	40 (22.9)	116 (17.0)	93 (12.1)	135 (23.6)
9 m	58	-39 (24.8)	-54 (25.6)	-14 (25.3)	75 (20.2)	62 (13.8)	91 (27.2)
12 m	-7	-100 (25.9)	-102 (27.4)	-54 (27.3)	5 (21.3)	-1 (14.8)	33 (29.5)

Table 5-11 Effect of BF on Length Gain (cm) through 12 Months

Time	ITT	RCACE Δ (SE) (by Matching and Stratification)					
		DPS Caliper	CPS Matching E	CPS Matching C	Equal Weight	Weight by Proportion	Weight by Strata
1 m	0.16	0.24 (0.04)	0.27 (0.04)	0.23 (0.05)	0.10 (0.04)	0.14 (0.03)	-0.004 (0.05)
2 m	0.32	0.29 (0.05)	0.37 (0.05)	0.31 (0.06)	0.26 (0.04)	0.29 (0.03)	0.12 (0.06)
3 m	0.50	0.32 (0.06)	0.44 (0.06)	0.37 (0.06)	0.46 (0.05)	0.45 (0.04)	0.32 (0.06)
6 m	0.46	0.23 (0.08)	0.33 (0.07)	0.19 (0.08)	0.45 (0.06)	0.41 (0.04)	0.38 (0.09)
9 m	0.31	0.02 (0.07)	0.11 (0.07)	-0.03 (0.08)	0.28 (0.06)	0.26 (0.04)	0.27 (0.09)
12 m	0.18	0.03 (0.07)	0.03 (0.07)	0.003 (0.08)	0.15 (0.06)	0.13 (0.05)	0.24 (0.09)

5.5.2 Discussion of PROBIT Results

The FCACE estimates were larger than the CACE estimates but in the same direction. For weight gain, the matching approach yielded results that were approximately double the magnitude of results from the ITT analysis at first 2 months, comparable at 3 month, and much smaller at 9 and 12 months. In contrast, the stratifications approach yielded results very close to the effects from ITT analysis throughout the one year period. The same trend is observed for length gain as well. Overall, infant growth varied significantly in early infancy but the differences were not evident at one year of age.

One limitation of our analysis is that measurement error for infant weight and length and for covariates may impact our findings. Because the primary hypotheses of PROBIT were not related to infant growth outcomes, measurements of weight and length were not standardized among the study sites. However, the impact should have been non-differential with respect to treatment assignment and actual breastfeeding behaviour, and therefore should have biased effects on infant growth toward the null. The negative correlation between the dual propensity score is observed in Figures 5-1 to 5-4. This reflects mothers' breastfeeding behaviour under both assignments (see Table 5.8), the definition of the dual propensity score, and the underline distributions of the compliance strata.

5.6. Discussion

In an RCT with partial treatment compliance, compliant subjects usually differ systematically from subjects who are not compliant in terms of their observed and

unobserved covariates. In this chapter, we define compliance stratification effects based on principal stratification theory and present theoretically justified strategies to estimate compliance stratification effects. In particular, DPS and CPS matching strategies allow the identification of pre-defined compliance stratification and lead to estimation of the corresponding compliance stratification effects. Using data from the PROBIT study, we apply this theory in practice.

An important feature of matching approaches is that the goal is to construct pre-defined principal strata. After subjects are matched, their membership in compliance strata is identified. A drawback is that the unmatched subjects were discarded. Those subjects are not directly used in estimating the pre-defined compliance stratification effects.

Several other limitations of our analysis require discussion. One such limitation is the assumption of 'no unmeasured confounders.' This is a strong and untestable assumption, and it is not known whether all relevant covariates that could confound the effect of breastfeeding on infant growth were collected. In fact, it is likely that breastfeeding is a dynamic process and that baseline characteristics alone are not sufficient to fully capture compliance. DPS methods (matching or stratification) can lead to valid principal stratification only when the assignment mechanism is truly unconfounded given the observed covariates.

Chapter 6 SIMULATION STUDIES

Monte Carlo simulation studies are conducted to evaluate the performance of the proposed estimators in Chapter 4 and 5, including intention-to-treat (ITT), naive plug-in (NPI), regression (REG), stratification with equal weight (SEW), proportion weight (SPW) and strata rank weight (SRW), and matching by dual propensity score (MDPS) and by counterfactual propensity score (MCPS). The primary objective is to estimate CACE (as discussed in Chapter 4) and FCACE (as discussed in Chapter 5) over randomly generated samples from simulated populations, which vary dual propensity score distributions and the proportions in the principal compliance strata, to examine performance of the proposed estimators and to identify estimators that perform well across populations.

6.1 CACE in RCTs with All-or-none Compliance

6.1.1 Simulation Specifications

In this section we describe the design of Monte Carlo simulations. A Monte Carlo simulation study involves random sampling techniques to generate a series of random samples from distributions that represent the study population of interest (Burton et al. 2006). For each generated random sample, different algorithms are applied and summary statistics are calculated. Then, empirical estimates of characteristics of the sampling distribution are obtained and compared to the known truth.

Under the situation in RCT with all-or-none treatment compliance, as we have described in Chapter 4, we assume that the population consists of three different

subpopulations or compliance strata (P): always-taker a, never-taker n, and complier c. These three populations can be identified based on subjects' status of the experimental treatment receipt under two treatment assignments, but only one of the treatment receipt (i.e., $S(1)$ and $S(0)$) can be observed. We assume that all covariates related to either $S_i(1)$ (if subject i is assigned to the treatment group) or $S_i(0)$ (if subject i is assigned to the control group) are correctly collected (no unmeasured confounders assumption) and dual propensity scores, r^1 and r^0, can be estimated for each subject, regardless his/her treatment assignment. We further assume r^1 follows uniform distribution $U[a^1, b^1]$ and r^0 follows uniform distribution $U[a^0, b^0]$. The distributions are assumed to be different for the three compliance strata a, n, and c, therefore, there are a total of six uniform distributions, presented as $U[a^{1a}, b^{1a}]$ (r^1 for always-taker a), $U[a^{1n}, b^{1n}]$ (r^1 for never-taker n), $U[a^{1c}, b^{1c}]$ (r^1 for complier c), and $U[a^{0a}, b^{0a}]$, $U[a^{0n}, b^{0n}]$ and $U[a^{0c}, b^{0c}]$. We randomly sample r^1 and r^0 directly from pre-specified uniform distributions for three compliance strata as an alternative way of estimating r^1 and r^0 by building predictive models based on covariates.

Our simulations also reflect the notion of counterfactual, as described in the previous chapters. Specifically, for each subject i, there are two potential outcomes; $Y_i(1, s)$ and $Y_i(0, s)$ exist for each subject i, but only one of the pair can be observed and the other becomes counterfactual. We assume that $Y_i(1, s)$ and $Y_i(0, s)$ follow normal distributions with a mean $\mu_{k,s}$ and a variance $\sigma^2_{k,s}$, where $k \in \{a, n, c\}$ and $s = 1$ or 0. Further, we assume that $Y_i(1, s)$ and $Y_i(0, s)$ are independent for subject i (Jin and Rubin, 2008). Therefore, there are total of six different distributions for three compliance strata. Following the exclusion restriction assumption for always-takers and never-takers from

Angrist, Imbens, and Rubin (1996), we assume $Y_i(1,1) = Y_i(0,1)$ for always-takers and $Y_i(1,0) = Y_i(0,0)$ for never-takers.

We consider nine (9) sets of simulation scenarios, as shown in Table 6-1. Some parameters are considered to be fixed across experimental conditions and some are varied. The parameters that we fix are outcome distributions $Y_i(1, s)$ and $Y_i(0, s)$. The parameters that we vary are the distributions of dual propensity scores r^1 and r^0, and the proportions of the principal compliance strata p^k (where $k \in \{a, n, c\}$).

Table 6-1 Simulation Specifications for All-or-none Compliance

S#	P	$S(1)$	$S(0)$	$Y(1, s) \sim$	$Y(0, s) \sim$	$r^1 \sim$	$r^0 \sim$	p^k
1	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,¾]$	$U[¼,1]$	1/4
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[¼,1]$	$U[0,¾]$	1/4
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[¼,1]$	$U[¼,1]$	1/2
2	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,⅔]$	$U[⅓,1]$	1/4
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[⅓,1]$	$U[0,⅔]$	1/4
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[⅓,1]$	$U[⅓,1]$	1/2
3	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,½]$	$U[½,1]$	1/4
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[½,1]$	$U[0,½]$	1/4
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[½,1]$	$U[½,1]$	1/2
4	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,¾]$	$U[¼,1]$	1/8
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[¼,1]$	$U[0,¾]$	1/8
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[¼,1]$	$U[¼,1]$	3/4
5	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,⅔]$	$U[⅓,1]$	1/8
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[⅓,1]$	$U[0,⅔]$	1/8
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[⅓,1]$	$U[⅓,1]$	3/4
6	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,½]$	$U[½,1]$	1/8
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[½,1]$	$U[0,½]$	1/8
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[½,1]$	$U[½,1]$	3/4
7	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,¾]$	$U[¼,1]$	1/20
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[¼,1]$	$U[0,¾]$	1/20
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[¼,1]$	$U[¼,1]$	9/10
8	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,⅔]$	$U[⅓,1]$	1/20
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[⅓,1]$	$U[0,⅔]$	1/20
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[⅓,1]$	$U[⅓,1]$	9/10
9	n	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,½]$	$U[½,1]$	1/20
	a	1	1	$N(8, 5^2)$	$N(8, 5^2)$	$U[½,1]$	$U[0,½]$	1/20
	c	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[½,1]$	$U[½,1]$	9/10

In Table 6-1, S# is the scenario number (from 1 to 9), *P* is the indicator of compliance strata, and the rest of the columns are defined in the previous section and will be explained in detail in the following section. One variable not listed in the table is the treatment assignment Z. Z is randomly generated in a way so that compliance strata across all scenarios are well balanced in terms of assignment to the treatment groups.

For outcome distributions across all scenarios, we set $Y_i(1,0)$ and $Y_i(0,0) \sim N(0, 5^2)$ for never-takers, $Y_i(1,1)$ and $Y_i(0,1) \sim N(8, 5^2)$ for always-takers, and $Y_i(1,1) \sim N(0, 5^2)$ and $Y_i(1,1) \sim N(0, 5^2)$ for compliers. We set all variances as equal for simplicity, i.e., $\sigma^2_{k,s} = 5^2$ where $k \in \{a, n, c\}$ and $s = 1$ or 0. We set the mean as 0 for distributions with $s = 0$ regardless of treatment assignment, which implies that compliers and never-takers have the same distribution in the control group ('no compliance effect for controls' (NCEC) assumption, Little, Long and Lin, 2008). For always-takers, we assume they have a larger mean than compliers do in the experimental group, implying that always-takers are the group who benefit most from treatment (e.g., sicker patients) so they take experimental treatment regardless of their treatment assignment (Jin and Rubin, 2008). The choices of $\mu_{c,1} = 5$ and $\sigma^2_{c,1} = 5^2$ are arbitrary. Initially, we set both as unity (1), but under partial compliance, some of $\mu_{k,s}$ have to be different fractions. To avoid fractions, we set all $\mu_{k,s}$ between 0 and 10 so $\mu_{c,1} = 5$, $\mu_{n,1} = 0$, and $\mu_{a,1} = 8$. Although the means (e.g, 0, 5 and 8) and variance we choose are arbitrary, there is no reason to believe that other choices would lead to a different conclusion. However, we have to acknowledge that different relationships between $\mu_{k,s}$ could yield different evaluation statistics but it would not change relative performance for those eight different estimators.

For dual propensity scores r^1 and r^0, their distributions are varied from Scenario #1 to #3 (we will refer to them as Scenario Group 1); the same pattern repeats from Scenario #4 to #6 (Scenario Group 2) and from Scenario #7 to #9 (Scenario Group 3). As we have discussed in Chapter 4, always-takers are observable under Z = 0, but never-takers and compliers cannot be differentiated. Therefore, we set $U[a^{0n}, b^{0n}]$ and $U[a^{0c}, b^{0c}]$ identically as $U[a^0, 1]$ to take into consideration that two distributions cannot be identified, and at the same time, set $U[a^{0a}, b^{0a}]$ differently as $U[0, b^0]$. We define overlap d as $d = b^0 - a^0$ and consider three different levels of overlap: *substantial* ($d = ½$), *moderate* ($d = ⅓$) and *none* ($d = 0$). Similarly, never-takers are observable under Z = 1, but not always-takers and compliers. Therefore, we set $U[a^{1a}, b^{1a}]$ and $U[a^{1c}, b^{1c}]$ identically as $U[a^1, 1]$ and set $U[a^{1n}, b^{1n}]$ differently as $U[0, b^1]$. We consider three different levels of overlap as well (substantial d = ½, moderate d = ⅓ and none d = 0), and define overlap similarly as $d = b^1 - a^1$. Notice that when there is no overlap ($d = 0$), compliers, in theory, can be identified in both the control and experimental group and the models can correctly predict the dual propensity scores.

We consider three different cases regarding the proportions of the compliance strata p^k, where $p^k = \Pr(P = k)$, $k \in \{a, n, c\}$, and $p^a + p^n + p^c = 1$. These three cases are as follows: 1) compliers compose only half of population with *substantial* non-compliance (50 %); 2) compliers consist of three quarter of population with *moderate* non-compliance (25 %); and 3) compliers make up nine-tenth of population with *mild* non-compliance (10 %). In all three cases, never-takers and always-takers evenly make up the rest of population. Therefore, in the first case we set $p^c = 1/2$ and $p^n = p^a = 1/4$; in the second case we set $p^c = 3/4$ and $p^n = p^a = 1/8$; and in the last case we set $p^c = 9/10$ and

$p^n = p^a = 1/20$. We expect all estimators would perform better when there is very mild non-compliance (Scenario Group 3) than where there exists substantial (Scenario Group 2) or moderate non-compliance.

6.1.2 Data-generating Process

Data are generated separately for each compliance stratum *a*, *n*, and *c*, all including the following variables: s_i^1, s_i^0, $Y_i(1, s)$, $Y_i(0, s)$, r_i^1, r_i^0, and Z_i. For instance, under Scenario #1 and for complier *c*, first we set $s_i^1 = 1$ and $s_i^0 = 0$, next we randomly generate two independent outcomes $Y_i(1,1)$ and $Y_i(0,0)$, where $Y_i(1,1) \sim N(5, 5^2)$ and $Y_i(0,0) \sim N(0, 5^2)$, then we randomly generate two propensity scores r_i^1 and r_i^0, where r_i^1 and $r_i^0 \sim U$ [¼,1], and finally we randomly generate treatment assignment variable Z_i (z = 0,1). We set Pr ($Z_i = 1$) = 0.5 so each subject is randomly assigned to either the treatment group or control group. Then, the true CACE in this case is 5. Keep in mind that only one of outcomes $Y_i(1,1)$ and $Y_i(0,0)$ are included in estimating the CACE, depending on Z_i, reflecting the counterfactual nature of the design.

Data are generated in the same way for each scenario and for each compliance stratum according to the specifications in Table 6-1. In each scenario, a total of 10,000 subjects are generated and evenly randomized to either the treatment group or control group. The sample size of 10,000 is chosen arbitrarily but it is considered sufficiently large for our purpose and it is in agreement with several published simulation studies conducted by other researchers on propensity score (e.g., Austin 2007 and Austin et al. 2007a; 2007b), on principal stratification (e.g., Gallop et al. 2009), and in other fields (e.g., Lefebvre, Delaney, and Platt, 2008). Because of varying proportion of compliance

strata, different numbers of subjects are generated for each stratum and scenario combinations, but the total sample size is fixed at 10,000. For example, under Scenario #1, we generate 5,000 compliers (N_c = 5000), 2,500 never-takers (N_n = 2500) and 2,500 always-takers ($N_{a,}$= 2500).

Each simulation scenario uses 1,000 replications. The number of simulations to perform is calculated based on the equation in Section 2.6 of the paper by Burton et al. (2006). A sample size of 500 is determined based on 80 % of power to detect a mean difference of 1 (with a standard deviation of 4) from the true value (which equals 5) as significant. The calculation is based on the assumption that the test statistics follows an approximately normal distribution. We also run a power analysis, as shown in Figure 6-1, in which we set a mean difference as 3, 2 and 1 each, with standard deviations as 4, 3, and 2, respectively. We finally double the number of simulated datasets needed to 1,000 to follow what have been used by Austin (Austin 2007) and his colleagues (Austin et al. 2007a; 2007b) in their simulations.

In summary, 1,000 datasets are randomly generated consisting of 10,000 subjects for each of nine simulation scenarios. The data-generating process and analyses are conducted using SAS version 9.1 (SAS Institute Inc., Cary NC).

6.1.3 Estimating CACE

Using each of the 1,000 simulated datasets, we estimate CACE using each of the eight methods described in Chapter 4 and 5, namely, ITT, NPI, REG, SEW, SPW, SRW, MDP, and MCPS. In the following section, we briefly review these estimators and their expected performances from the study design point of view.

ITT is the intention-to-treatment estimator, which ignores the compliance information so as to the dual propensity scores. ITT is considered to be the gold standard for estimating effectiveness. However, ITT usually underestimates efficacy (the true CACE). We expect the bias to be substantial.

NPI is the naive plug-in estimator, which identifies compliers based on observed treatment receipt and their counterfactual propensity score. NPI selects observed 'compliant' subjects first (e.g, *a* and *c* in the experimental group), and then removes the ones whose estimated counterfactual propensity score is below the pre-specified certain cut-off (e.g., $r^0 < 0.5$) so being considered as either *n* or *a*. For example, in the experimental group z =1, the compliers are identified by the following conditions: $s^1 = 1$ and $r^0 \geq 0.5$. Only the identified 'compliers' are included in the analysis to estimate the CACE. We use a cut-off of 0.5 based on the study design so we expect the NPI estimator to yield results close to the truth when there is no overlap (Scenario #3, #6, and #9).

REG is the regression estimator, estimated by fitting the regression model with treatment assignment z, dual propensity score r^1 and r^0, and the interaction between $r^1$and r^0. We expect REG to yield results close to the ITT estimates since both are considered as a measure of *marginal* causal effect (MCE), whereas CACE is considered as a *conditional* causal effect (CCE) for the subpopulation of compliers (Angrist, Imbens and Rubin, 1996)

SEW, SPW, and SRW are three estimators based on stratification. First, strata are constructed based on the quintiles of dual propensity score, then subjects are classified into 25 (5x5) 'compliance' strata. Once the strata are defined, the treatment effect is evaluated by comparing subjects directly between the two treatment groups within each

stratum. Then, the mean of the differences across strata is summarized using three different weighting schemes: equal weight, weight by the proportion (sample size of each stratum), and by the rank. We expect SEW and SPW would yield results close to ITT since both weighting schemes are standard strategies and do not give more weights to 'compliers' and less weights to 'non-compliers.' Recall that dual propensity scores are the probability of being treated in the experimental group and the probability of not being treated in the control group. The higher the scores, the higher the stratum rank, and the more likely the subject is a complier. Using the scheme with stratum ranking, SRW estimates CACE by overweighting the differences over the subset of 'compliers' and underweighting the differences for the rest of the strata. Therefore, we expect SRW would yield results less biased than results from SEW and SPW.

Last, MDPS and MCPS are two estimators created through matching. We use DPS and CPS matching to create a matched sample. For each subject in the experimental group, a control subject with the closest 'distance' is selected from the control group based on either DPS (MDPS, using caliper matching; caliper is calculated as 0.6 of the standard deviation of r^1 and r^0; see Austin, 2009) or CPS (MCPS, using a variation of nearest-neighbour matching.). Then, matched pairs are classified into predefined compliance principal stratifications. Finally, differences between the two treatment groups are computed conditional on the newly identified subgroups of 'compliers.' Therefore, we expect both MDPS and MCPS to yield results close to the true CACE.

6.1.4 Evaluation of Estimator Performance

For each method, we calculate the mean of estimated CACE (as the average of CACE over the 1,000 replications) and its standard deviation, using the standard two-sample procedure. Furthermore, we evaluate the performance of the eight methods against the true CACE by determining bias, relative bias (percentage bias and standardized bias), coverage of 95 % confidence intervals, and mean squared error (MSE) (see Burton et al. 2006 for computational details and additional performance measures).

The bias is defined as the difference between the average estimate of CACE and the true CACE. The percentage bias is defined as the percentage of bias over the truth. The standardized bias is calculated as the empirical bias divided by the standard error (estimated by standard deviation of simulated estimates). We refer to both the percentage bias and the standard bias as *relative bias*. The coverage of 95 % confidence intervals is the proportion of times that the estimated 95 % confidence interval contains the true CACE; that is to say, the proportion of the replications where the estimate of CACE is within 1.96 estimated standard errors from the truth. Since 95 % confidence intervals are calculated using 1,000 independent simulations, coverage between 0.936 and 0.964 is considered acceptable (Burton et al. 2006). Last, the MSE is calculated as the sum of the square of the bias and the empirical variance of an estimator over all simulations. Therefore, MSE allows one to quantify the variance-bias trade-off (Burton et al. 2006).

6.1.5 Simulation Results

Results of the Monte Carlo simulations are reported in Table 6-2 to 6-5, and are discussed in the following section. In general, simulation results indicate a common pattern shared by different evaluations with respect to the performance of the eight estimators, and the findings confirm our expectation of the relative performance between them.

The mean estimated CACE for each method are reported in Table 6-2. In all scenarios, ITT is underestimated true CACE roughly by the fraction of the proportion of compliers (p^c). It is not surprising because we assume that causal effects for always-takers and never-takers are zero; therefore, ITT is estimated as CACE but over the total population, which is set up as one time ($p^c = 1/2$), or one-third ($p^c = 3/4$), or one-tenth ($p^c = 9/10$) larger than the sample size of compliers, respectively. In scenarios with the same p^c (e.g., Scenario #1 - #3), ITT estimations remain identical and show that the varying overlaps of dual propensity scores have no impact on the ITT estimator because ITT ignores compliance information completely.

In contrast to the ITT estimator, NPI generates better results compared to the true CACE with negligibly higher values. As we have discussed in the previous section, one reason is that the cut-off (=0.5) we use in our simulation is known to be the 'best' based on the study design. That will not be granted in reality, since the cut-off is 'unknown' and has to be figured out one way or another. Especially in the cases of no overlap (Scenario #3, #6, and #9), compliers are identifiable through their dual propensity scores (e.g., in the treatment group, $s^1 = 1$ and $r^0 \geq 0.5$), so NPI is estimating the true 'compliers average causal effect.' In the cases of substantial and moderate overlap, NPI overestimates the

true CACE but with a different amount, e.g. 5.43 and 5.60 (p^c = 1/2), 5.15 and 5.22 (p^c = 3/4), and 5.05 to 5.08 (p^c = 9/10), respectively. The magnitude of overestimation increases when there is more overlap but shrinks when the proportion of compliers increases.

REG and two 'standard' stratification estimators, SEW and SPW, yield essentially identical results to ITT, confirming our belief that all three estimators provide marginal causal effects across all compliance strata as ITT does, where CACE is a conditional causal effect for the subpopulation of compliers. To facilitate our discussion, we refer to ITT, REG, SEW and SPW as the ITT-type estimators, and refer to NPI, SRW, MDPS and MCPS as the LATE-type estimators (LATE stands for local average causal effect).

The last stratification estimator, SRW, which uses ranks of the compliance strata as weight, also underestimates the true CACE but by a smaller amount compared with ITT and two other standard stratification estimators. The mean estimates range from 3.12 (Scenario #1 when there is substantial overlap and substantial non-compliance) to 4.82 (Scenario #9 when there is no overlap and mild non-compliance).

Two matching estimators, MDPS and MCPS, overestimate the true CACE as NPI does, and turn out to be the best estimators, as we expected. Both produce identical results, ranging from 5.71 (Scenario #1) to 5.00 (Scenario #9), regardless of whether DPS or CPS is used for the matching algorithm.

Table 6-2 Estimation of CACE (True CACE = 5)

#	d =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	2.51	5.60	2.50	2.52	2.50	3.13	5.71	5.71
2	⅓	1/2	2.51	5.43	2.51	2.56	2.51	3.45	5.54	5.55
3	0	1/2	2.51	5.00	2.50	2.82	2.50	3.97	5.04	5.01
4	½	3/4	3.75	5.22	3.75	3.75	3.75	4.15	5.30	5.29
5	⅓	3/4	3.75	5.15	3.75	3.76	3.75	4.33	5.23	5.22
6	0	3/4	3.75	5.00	3.75	3.79	3.75	4.57	5.02	5.00
7	½	9/10	4.50	5.08	4.50	4.50	4.50	4.68	5.11	5.10
8	⅓	9/10	4.50	5.05	4.50	4.50	4.50	4.75	5.08	5.08
9	0	9/10	4.50	5.00	4.50	4.50	4.50	4.82	5.00	5.01

Note: Each cell contains the mean estimated CACE.

The bias and relative biases for each method are reported in Table 6-3. The bias is substantial when there is a substantial overlap and/or substantial non-compliance. The bias decreases as the overlap and the proportion of the non-compliance decrease. In all scenarios, each of the four ITT-type estimators (ITT, REG, SEW and SPW) result in negatively biased estimation, with a bias ranging from -2.5 to -0.5. As expected, among the four estimators, the estimated CACE are biased substantially but comparable. On the other hand, three of the four LATE-type estimators (NPI, MDPS and MCPS) result in slightly upwards biased estimation, with a bias ranging from 0.7 to 0. The SRW results in minor negative bias when substantial overlap and non-compliance exist. As expected, the bias is negligibly different than zero in the cases of no overlap (Scenario #3, #6, and #9). Across all scenarios, the NPI is the best estimator among the four, and the SRW is the worst performer.

The relative bias decreases with increasing the proportion of the compliers in the populations or with decreasing the overlap. When the proportion of compliers consists of 90 % of the population, the relative bias ranges from -10 % to zero %; this is expected because all of the estimators are quite similar by design to the ITT estimator in this setting. For the ITT-type estimators, the relative bias ranges from a low of -50 % (- 22 % with the standardized bias) to a high of -10 % (- 5 % with the standardized bias). For the LATE-type estimators other than SRW, the relative bias ranges from a high of 14 % (5 % with the standardized bias) to a low of 0 % (0 % with the standardized bias). The SRW has a relative bias ranging from -37 % (-14 % with the standardized bias) to - 3.5 % (-1.5 % with the standard bias).

To summarize in terms of bias, the four LATE-type estimators result in unbiased estimations of the true CACE, while the four ITT-type estimators result in downwards biased estimations.

Table 6-3 Biases of the Estimated CACE (True CACE = 5)

#	D =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	-2.49 -49.9 -22.0	0.60 12.1 4.0	-2.49 -49.9 -22.0	-2.50 -49.9 -22.9	-2.48 -49.6 -22.5	-1.87 -37.4 -14.4	0.71 14.1 5.2	0.71 14.2 5.1
2	⅓	1/2	-2.49 -49.9 -22.0	0.43 8.5 2.9	-2.49 -49.9 -23.0	-2.44 -48.8 -22.9	-2.49 -49.9 -22.9	-1.55 -30.9 -12.1	0.54 11.0 3.9	0.55 11.0 4.1
3	0	1/2	-2.49 -49.9 -22.0	0.00 0.02 0.01	-2.49 -49.9 -22.0	-2.50 -49.9 -22.9	-2.48 -49.6 -22.5	-1.87 -37.4 -14.4	0.04 0.73 0.26	0.01 0.27 0.09
4	½	3/4	-1.25 -25.0 -11.9	0.22 4.5 1.7	-1.25 -25.0 -12.0	-1.25 -25.0 -12.0	-1.25 -25.0 -12.0	-0.85 -16.9 -7.0	0.30 6.0 2.6	0.29 5.9 2.4
5	⅓	3/4	-1.25 -25.0 -11.9	0.15 3.1 1.2	-1.25 -25.0 -12.1	-1.24 -24.9 -12.0	-1.25 -25.0 -12.1	-0.67 -13.5 -5.4	0.23 4.6 1.9	0.22 4.3 1.8
6	0	3/4	-1.25 -25.0 -11.9	0.0 -0.1 -0.03	-1.25 -25.0 -12.5	-1.21 -24.3 -12.0	-1.25 -25.0 -12.4	-0.43 -8.6 -3.5	0.02 0.32 0.14	0.00 0.04 0.02
7	½	9/10	-0.50 -10.1 -5.0	0.08 1.5 0.6	-0.50 -10.1 -5.0	-0.50 -10.1 -4.9	-0.50 -10.1 -5.0	-0.32 -6.4 -2.6	0.11 2.23 1.04	0.10 2.08 0.93
8	⅓	9/10	-0.50 -10.1 -5.0	0.05 1.0 0.4	-0.50 -10.1 -5.0	-0.50 -10.1 -5.0	-0.50 -10.1 -5.0	-0.25 -5.0 -2.0	0.08 1.7 0.8	0.08 1.6 0.7
9	0	9/10	-0.50 -10.1 -5.0	0.00 -0.09 -0.04	-0.50 -10.1 -5.0	-0.50 -10.0 -5.0	-0.50 -10.1 -5.0	-0.18 -3.5 -1.5	0.01 0.11 0.05	0.00 0.02 0.01

Note: Each cell contains the bias, the percentage bias and the standardized bias in order.

Moreover, we report the empirical coverage of 95 % confidence intervals of the estimators in Table 6-4. It is not a surprise to see that the ITT-type estimators provide no coverage at all across all scenarios, given the small amount of variance implied by the study design relative to the substantial bias. The same argument applies to the poor coverage of the LATE-type estimators as well. For the LATE-type estimators other than SRW, coverage is only acceptable in the cases of no overlap (Scenario #3, #6, and #9). In the cases of mild non-compliance, LATE-type estimators other than SRW maintain

considerable coverage from .816 to .950. However, when non-compliance is substantial, all estimators grossly undercover the true value with less than 5 % when the overlaps are not zero.

Table 6-4 Coverage of 95 Percent Confidence Intervals for CACE

#	d =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	.000	.022	.000	.000	.000	.000	.001	.001
2	⅓	1/2	.000	.022	.000	.000	.000	.000	.019	.024
3	0	1/2	.000	.947	.000	.000	.000	.000	.954	.947
4	½	3/4	.000	.360	.000	.000	.000	.000	.629	.305
5	⅓	3/4	.000	.545	.000	.000	.000	.000	.777	.530
6	0	3/4	.000	.975	.000	.000	.000	.000	.963	.950
7	½	9/10	.002	.835	.020	.000	.000	.007	.915	.816
8	⅓	9/10	.002	.886	.001	.000	.000	.029	.930	.864
9	0	9/10	.002	.936	.000	.000	.000	.085	.950	.938

Note: Each cell contains the coverage of 95 % confidence intervals

Table 6-5 shows the MSE of the estimated CACE. It is obvious that all MSEs for LATE-type estimators are considerably smaller than the corresponding MSEs for ITT-type estimators, which are consistent with the pattern of the biases demonstrated in the previous sections. Since the four ITT-type estimators are severely biased, their MSEs are overwhelmingly dominated by the bias while the standard errors of the estimators are negligibly small. The NPI, MDPS and MCPS perform equally well with the smallest MSEs in all settings. This is because for larger sample sizes, the biases have a larger contribution to the MSEs as the standard errors of the estimators are smaller.

Table 6-5 Mean Squared Error of Estimated CACE

#	d =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	6.21	0.38	6.26	6.16	6.26	3.51	0.52	0.52
2	⅓	1/2	6.21	0.21	6.21	6.00	6.21	2.42	0.31	0.32
3	0	1/2	6.21	0.02	6.26	4.76	6.26	1.08	0.02	0.02
4	½	3/4	1.57	0.07	1.57	1.57	1.57	0.74	0.10	0.10
5	⅓	3/4	1.57	0.04	1.57	1.55	1.57	0.46	0.07	0.06
6	0	3/4	1.57	0.01	1.57	1.47	1.57	0.20	0.01	0.01
7	½	9/10	0.26	0.02	0.26	0.26	0.26	0.12	0.02	0.02
8	⅓	9/10	0.26	0.02	0.26	0.26	0.26	0.08	0.02	0.02
9	0	9/10	0.26	0.01	0.26	0.26	0.26	0.05	0.01	0.01

Note: Each cell contains the mean squared error of estimated CACE.

In summary, the results obtained from the LATE-type estimators are performed well in terms of bias and coverage, whereas results from the ITT-type estimators are performed poorly with negatively biased estimations. Among the LATE-type estimators, the NPI, MDPS and MCPS perform equally well with minimal biases and the smallest mean of squared errors when there are moderate and mild non-compliance. The NPI is a slightly better performer than MDPS and MCPS when there is substantial non-compliance. The SRW is the worst performer among the four in all settings.

6.1.6 Sensitivity Analyses of NPI Using Different Cut-offs

Sensitivity analyses are conducted by using different cut-offs (0.3, 0.4, 0.6, and 0.7 for both r^1 and r^0) to evaluate how sensitive the NPI estimate is to the choice of cut-offs. Tables 6-6 presents the estimations of CACE using NPI with different cut-offs. The results show that the estimates are closer to the true value when there is mild or no overlap with mild non-compliance. For each scenario, the estimates are centred at the one using the 0.5 cut-off, with a deviation less than 10 %. The estimates using 0.7 as the cut-off are closer to the true value compared to the estimates using 0.3 as the cut-off. One explanation is that fewer always-takers would be misclassified as compliers in the experimental group with a larger cut-off, which would result in a smaller mean in the experimental group. On the other hand, fewer never-takers would be misclassified as compliers in the control group with a larger cut-off, which would have a neutral impact on mean estimation in the control group.

Table 6-6 Estimation of CACE Using NPI with Different Cut-offs

#	d =	p^c	Cut-off				
			0.5	0.3	0.4	0.6	0.7
1	½	1/2	5.60	5.73	5.68	5.48	5.23
2	⅓	1/2	5.43	5.65	5.55	5.23	5.00
3	0	1/2	5.00	5.50	5.27	5.00	5.00
4	½	3/4	5.22	5.29	5.26	5.17	5.07
5	⅓	3/4	5.15	5.25	5.21	5.08	5.00
6	0	3/4	5.00	5.18	5.09	5.00	5.00
7	½	9/10	5.08	5.10	5.09	5.06	5.02
8	⅓	9/10	5.05	5.09	5.07	5.02	5.00
9	0	9/10	5.00	5.06	5.03	5.00	4.99

Note: Each cell contains the mean estimated CACE.

Table 6-7 presents the bias and relative biases for each cut-off. The bias and relative bias are noticeable when there is substantial overlap and/or substantial non-compliance. The biases decrease as the overlap and the proportion of the non-compliance decrease. The relative bias ranges from a high of 15 % (6 % with the standardized bias) to a low of 0 % (0 % with the standardized bias).

Table 6-7 Biases of the Estimated CACE Using NPI with Different Cut-offs

#	d =	p^c	Cut-off				
			0.5	0.3	0.4	0.6	0.7
1	½	1/2	0.60 12.1 4.0	0.73 14.6 5.9	0.68 13.6 5.1	0.48 9.6 2.7	0.23 4.6 1.1
2	⅓	1/2	0.43 8.5 2.9	0.65 13.0 5.4	0.55 11.0 4.2	0.23 4.6 1.4	0.0 0.0 0.0
3	0	1/2	0.0 0.02 0.01	0.50 10.0 4.0	0.27 5.4 2.0	0.0 0.0 0.0	0.0 0.0 0.0
4	½	3/4	0.22 4.5 1.7	0.29 5.8 2.6	0.26 5.2 2.2	0.17 3.4 1.1	0.07 1.4 0.4
5	⅓	3/4	0.15 3.1 1.2	0.25 5.0 2.3	0.21 4.2 1.8	0.08 1.6 0.5	0.0 0.0 0.0
6	0	3/4	0.0 -0.1 -0.03	0.19 3.8 1.7	0.09 1.8 0.8	0.0 0.0 0.0	-0.01 -0.2 -0.1
7	½	9/10	0.08 1.5 0.6	0.10 2.0 1.0	0.09 1.8 0.8	0.06 1.2 0.4	0.02 0.4 0.1
8	⅓	9/10	0.05 1.0 0.4	0.09 1.8 0.9	0.07 1.4 0.6	0.02 0.4 0.1	-0.01 -0.2 -0.1
9	0	9/10	0.01 -0.1 -0.0	0.06 1.2 0.6	0.03 0.6 0.3	0.0 0.0 0.0	-0.01 -0.2 -0.1

Note: Each cell contains the bias, the percentage bias and the standardized bias in order.

Table 6-8 presents the empirical coverage of 95 % confidence intervals of the estimators. In the cases of mild non-compliance, coverage ranges from .844 to .959. In

the cases of moderate non-compliance, coverage ranges from 0.294 to 0.963. When non-compliance is substantial and overlap is substantial (Scenario #1), coverage ranges from 0 to .815. The best coverage is observed for estimates with 0.7 cut-offs, ranging from .815 to .961.

Table 6-8 Coverage of 95 % CI of CACE Using NPI with Different Cut-offs

#	d =	p^c	Cut-off				
			0.5	0.3	0.4	0.6	0.7
1	½	1/2	.022	.000	.001	.234	.815
2	⅓	1/2	.194	.001	.016	.763	.960
3	0	1/2	.954	.024	.489	.959	.955
4	½	3/4	.629	.294	.435	.801	.938
5	⅓	3/4	.777	.395	.584	.913	.948
6	0	3/4	.963	.608	.871	.955	.961
7	½	9/10	.915	.844	.884	.929	.959
8	⅓	9/10	.930	.869	.895	.944	.933
9	0	9/10	.950	.910	.947	.953	.954

Note: Each cell contains the coverage of 95 % confidence intervals

Table 6-9 shows the MSE of the estimated CACE. The MSEs are relatively small in all settings, ranging from 0.55 to 0. The smallest MSEs are observed for estimates using the cut-off of 0.7, and the largest MSEs are observed for the cut-off of 0.3.

Table 6-9 Mean Squared Error of CACE Using NPI with Different Cut-offs

#	d =	p^c	Cut-off				
			0.5	0.3	0.4	0.6	0.7
1	½	1/2	0.38	0.55	0.48	0.26	0.10
2	⅓	1/2	0.21	0.44	0.32	0.08	0.04
3	0	1/2	0.02	0.27	0.09	0.02	0.03
4	½	3/4	0.07	0.10	0.08	0.05	0.04
5	⅓	3/4	0.04	0.07	0.06	0.03	0.03
6	0	3/4	0.01	0.05	0.02	0.02	0.02
7	½	9/10	0.02	0.02	0.02	0.03	0.03
8	⅓	9/10	0.02	0.02	0.02	0.02	0.03
9	0	9/10	0.01	0.01	0.01	0.01	0.02

Note: Each cell contains the mean squared error of estimated CACE.

In summary, the sensitivity analyses for NPI using different cut-offs show that the estimates are not sensible to different cut-offs. The relative performance is also based on the outcome distributions of always-takers, never-takers and compliers. One limitation is that we use the same cut-offs for r^1 and r^0.

6.2 FCACE in RCTs with Full-Partial-None Compliance

6.2.1 Simulation Specifications

In this section, we use the same definitions, notations, and assumptions as described in Section 6.1.1, unless specified otherwise. Under the settings in RCT in the presence of partial compliance as described in Chapter 5, we assume that the population consists of six different compliance strata P (assuming there are no defiers): partial-none-complier *pnc*, full-none-complier *fnc*, full-partial-complier *fpc*, partial-taker *p*, never-taker *n*, and always-taker *a*. Under two treatment assignments, $S(1)$ and $S(0)$ can be 1, ½, or 0. Accordingly, there will be three possible outcome distributions for each of two assignments. We maintain the assumption of the exclusion restriction for always-takers, never-takers and partial takers. Therefore, the causal effects are all zero for *a*, *n*, and *p*. Further, we assume that the causal effects for three compliers groups, *fnc*, *pnc*, and *fpc*, are none-zero; that is, FCACE, PCACE and FPCACE are not zero. In this simulation study, we focus on FCACE as we did in Chapter 5.

We consider six sets of simulation scenarios as shown in Table 6.6. We define overlap d the same way as in Section 6.1.1 and consider three different levels of overlap: substantial ($d = ½$), moderate ($d = ⅓$) and none ($d = 0$), under both treatment assignments. Given there are six compliance strata, let $p^k_i = \Pr(P_i = k)$ for $k \in \{fnc, pnc, fpc, a, p, n\}$ and assume $p^a + p^p + p^n + p^{fnc} + p^{pnc} + p^{fpc} = 1$. We consider two settings: in the first, compliers compose half of population (50 %), and in the second, compliers consist of three quarters of the population (75 %). In both cases, the five strata other than the complier evenly make up the rest of the population. We do not consider the case in which compliers make up nine-tenths of population, as there then would be very few

subjects for each of the five other strata, and the ordinary ITT estimator is expected to be nearly equivalent to the CACE. Specifically, we set $p^{fnc} = 1/2$ and $p^a = p^p = p^n = p^{pnc} = p^{fpc} = 1/10$ in the first case; we set $p^{fnc} = 3/4$ and $p^a = p^p = p^n = p^{pnc} = p^{fpc} = 1/20$ in the second case. Table 6-10 summarizes the data structure of our simulation studies.

Table 6-10 Simulation Specifications for Partial Compliance

#	*P*	*S*(1)	*S*(0)	$Y(1, s) \sim$	$Y(0, s) \sim$	$r^1 \sim$	$r^0 \sim$	Pr
1	*n*	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,¾]$	$U[¼,1]$	1/10
	a	1	0	$N(8, 5^2)$	$N(8, 5^2)$	$U[¼,1]$	$U[0,¾]$	1/10
	p	½	½	$N(3, 5^2)$	$N(3, 5^2)$	$U[¼,¾]$	$U[¼,¾]$	1/10
	fpc	1	½	$N(5, 5^2)$	$N(3, 5^2)$	$U[¼,1]$	$U[¼,¾]$	1/10
	pnc	½	0	$N(3, 5^2)$	$N(0, 5^2)$	$U[¼,¾]$	$U[¼,1]$	1/10
	fnc	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[¼,1]$	$U[¼,1]$	1/2
2	*n*	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,⅔]$	$U[⅓,1]$	1/10
	a	1	0	$N(8, 5^2)$	$N(8, 5^2)$	$U[⅓,1]$	$U[0,⅔]$	1/10
	p	½	½	$N(3, 5^2)$	$N(3, 5^2)$	$U[⅓,⅔]$	$U[⅓,⅔]$	1/10
	fpc	1	½	$N(5, 5^2)$	$N(3, 5^2)$	$U[⅓,1]$	$U[⅓,⅔]$	1/10
	pnc	½	0	$N(3, 5^2)$	$N(0, 5^2)$	$U[⅓,⅔]$	$U[⅓,1]$	1/10
	fnc	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[⅓,1]$	$U[⅓,1]$	1/2
3	*n*	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,½]$	$U[½,1]$	1/10
	a	1	0	$N(8, 5^2)$	$N(8, 5^2)$	$U[½,1]$	$U[0,½]$	1/10
	p	½	½	$N(3, 5^2)$	$N(3, 5^2)$	$U[⅓,⅔]$	$U[⅓,⅔]$	1/10
	fpc	1	½	$N(5, 5^2)$	$N(3, 5^2)$	$U[½,1]$	$U[⅓,⅔]$	1/10
	pnc	½	0	$N(3, 5^2)$	$N(0, 5^2)$	$U[⅓,⅔]$	$U[½,1]$	1/10
	fnc	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[½,1]$	$U[½,1]$	1/2
4	*n*	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,¾]$	$U[¼,1]$	1/20
	a	1	0	$N(8, 5^2)$	$N(8, 5^2)$	$U[¼,1]$	$U[0,¾]$	1/20
	p	½	½	$N(3, 5^2)$	$N(3, 5^2)$	$U[¼,¾]$	$U[¼,¾]$	1/20
	fpc	1	½	$N(5, 5^2)$	$N(3, 5^2)$	$U[¼,1]$	$U[¼,¾]$	1/20
	pnc	½	0	$N(3, 5^2)$	$N(0, 5^2)$	$U[¼,¾]$	$U[¼,1]$	1/20
	fnc	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[¼,1]$	$U[¼,1]$	3/4
5	*n*	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,⅔]$	$U[⅓,1]$	1/20
	a	1	0	$N(8, 5^2)$	$N(8, 5^2)$	$U[⅓,1]$	$U[0,⅔]$	1/20
	p	½	½	$N(3, 5^2)$	$N(3, 5^2)$	$U[⅓,⅔]$	$U[⅓,⅔]$	1/20
	fpc	1	½	$N(5, 5^2)$	$N(3, 5^2)$	$U[⅓,1]$	$U[⅓,⅔]$	1/20
	pnc	½	0	$N(3, 5^2)$	$N(0, 5^2)$	$U[⅓,⅔]$	$U[⅓,1]$	1/20
	fnc	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[⅓,1]$	$U[⅓,1]$	3/4
6	*n*	0	0	$N(0, 5^2)$	$N(0, 5^2)$	$U[0,½]$	$U[½,1]$	1/20
	a	1	0	$N(8, 5^2)$	$N(8, 5^2)$	$U[½,1]$	$U[0,½]$	1/20
	p	½	½	$N(3, 5^2)$	$N(3, 5^2)$	$U[⅓,⅔]$	$U[⅓,⅔]$	1/20
	fpc	1	½	$N(5, 5^2)$	$N(3, 5^2)$	$U[½,1]$	$U[⅓,⅔]$	1/20
	pnc	½	0	$N(3, 5^2)$	$N(0, 5^2)$	$U[⅓,⅔]$	$U[½,1]$	1/20
	fnc	1	0	$N(5, 5^2)$	$N(0, 5^2)$	$U[½,1]$	$U[½,1]$	3/4

6.2.2 Data-generating Process

Data are generated separately for each of the six compliance strata in the same way as done in the previous section. The true FCACE in this case is 5, and the true FPCACE and PCACE are 1 and 2 respectively. 1,000 datasets are randomly generated consisting of 10,000 subjects for each of six simulation scenarios. The data-generating process and analyses are conducted using SAS version 9.1 (SAS Institute Inc., Cary NC).

6.2.3 Estimation of FCACE

Using each of the 1,000 simulated datasets, we estimate FCACE using each of the eight methods described in Chapter 4 and 5. All the algorithms remain the same as described in the previous section and consistent with our discussion in Chapter 5.

6.2.4 Evaluation of Estimator Performance

For each method, we calculate the mean of estimated FCACE (as the average of FCACE among the 1,000 simulated datasets) and its standard deviation. Furthermore, we evaluate the performance of the eight methods against the true FCACE by determining bias, relative bias (percentage bias and standardized bias), coverage of 95 % confidence intervals, and mean squared error, all defined in Section 6.1.4.

6.2.5 Simulation Results

The results of the Monte Carlo simulations are reported in Table 6-11 - 6-14, and are discussed in the following sections. Simulation results indicate a common pattern

shared by different evaluations with respect to the performance of the eight estimators, and the findings confirm our expectation of the relative performance between them.

We report the mean estimated FCACE for each method in Table 6-11. In all scenarios, ITT underestimates true FCACE by a fraction of 3/5 and 4/5 respectively, which no longer equals the populations' proportions of compliers as shown in Table 6-2. This is because of the existence of two other complier groups, *fpc* and *pnc,* with causal effects (FPCACE and PCACE) of 1 and 2 respectively. We observe the ITT estimation remain identical when p^c is constant since the ITT ignores compliance information so that varying dual propensity scores have no impact on the estimation. Not surprisingly, the REG, SEW and SPW yield almost exactly identical results to the ITT. They all underestimate the true FCAC (which equals 5), by a fraction of 3/5 (~ 3.00 for the first three scenarios) and 4/5 (4.00 for the last three scenarios). Moreover, the SRW also underestimate the true FCACE but with a smaller amount than ITT and the other two standard stratification estimators. The estimations range from 3.42 (Scenario #1) to 4.62 (Scenario #6).

In contrast to the ITT estimator, the NPI overestimates the true FCACE in at least five scenarios ranging from 5.23 to 5.00, with one 4.99 essentially identical to the true value. The magnitude of overestimation decreases with a decreasing overlap or an increasing proportion of compliers. Two matching estimators, MDPS and MCPS, overestimate the true FCACE as NPI does. Both yield identical results, ranging from 5.27 (Scenario #1) to 4.99 (Scenario #3), regardless of whether DPS or CPS is used for the matching algorithm. No biases are observed when there is no overlap for all three estimators.

Table 6-11 Estimation of FCACE (True FCACE = 5)

#	d =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	2.99	5.23	2.99	2.99	2.99	3.45	5.27	5.26
2	⅓	1/2	2.99	5.16	2.99	3.00	2.99	3.67	5.19	5.18
3	0	1/2	2.99	4.99	2.99	3.04	2.99	4.10	4.99	4.99
4	½	3/4	4.00	5.09	4.00	4.00	4.00	4.30	5.12	5.11
5	⅓	3/4	4.00	5.06	4.00	4.00	4.00	4.44	5.09	5.08
6	0	3/4	4.00	5.00	4.00	4.00	4.00	4.62	5.00	5.00

Note: Each cell contains the mean estimated FCACE.

We report the bias and the relative biases in Table 6-12. The bias is substantial when there are substantial overlaps and/or substantial non-compliance. The bias decreases as the overlap and the proportion of the non-compliance decrease. In all scenarios, each of the four ITT-type estimators (ITT, REG, SEW and SPW) result in negatively biased estimation, with a bias either -2 or around -1. The NPI, MDPS and MCPS result in slightly upwards biased estimation, with a bias ranging from 0.27 to 0. The SRW results in minor negative bias when substantial overlap and non-compliance exist, ranging from -1.55 to -0.38.

The relative bias decreases with increasing the proportion of the compliers in the populations or with decreasing the overlap. For the ITT-type estimators, the relative bias ranges from a low of -50 % (- 22 % with the standardized bias) to a high of -10 % (- 5 % with the standardized bias). For the LATE-type estimators other than SRW, the relative bias ranges from a high of 14 % (5 % with the standardized bias) to a low of 0 % (0 % with the standardized bias). The SRW has a relative bias of either -40 % (about -20 % with the standardized bias) or - 20 % (about -10 % with the standard bias).

As observed in Section 6.1.5, the four LATE-type estimators result in unbiased estimations of the true FCACE, while the four ITT-type estimators results in downwards biased estimations.

Table 6-12 Biases of the Estimated FCACE (True FCACE = 5)

#	D =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	-2.01 -40.2 -18.9	0.23 4.5 1.4	-2.01 -40.2 -19.1	-2.01 -40.1 -19.0	-2.01 -40.2 -19.1	-1.55 -31.0 -12.1	0.27 5.4 1.9	0.26 5.2 1.8
2	⅓	1/2	-2.01 -40.2 -18.9	0.16 3.2 1.1	-2.01 -40.2 -19.2	-2.01 -40.0 -19.1	-2.01 -40.2 -19.2	-1.33 -26.6 -12.1	0.19 3.8 1.4	0.18 3.6 1.3
3	0	1/2	-2.01 -40.2 -18.9	-0.01 -0.18 -0.07	-2.01 -40.1 -19.6	-1.96 -39.2 -19.1	-2.01 -40.2 -19.5	-0.90 -17.9 -7.4	-0.01 -0.15 -0.05	-0.01 -0.15 -0.05
4	½	3/4	-1.00 -20.0 -9.8	0.09 2.8 0.7	-1.00 -20.0 -9.8	-1.00 -20.0 -9.8	-1.00 -20.0 -9.8	-0.70 -14.0 -5.6	0.12 2.4 1.1	0.11 2.2 0.9
5	⅓	3/4	-1.00 -20.0 -9.8	0.06 1.1 0.4	-1.00 -20.0 -9.8	-1.00 -20.0 -9.8	-1.00 -20.0 -9.8	-0.56 -11.2 -4.6	0.09 1.7 0.7	0.08 1.8 0.7
6	0	3/4	-1.00 -20.0 -9.8	0.0 -0.04 -0.02	-1.00 -20.0 -9.9	-1.00 -20.0 -9.8	-1.00 -20.0 -9.8	-0.38 -7.7 -3.2	0.00 -0.04 -0.01	0.00 -0.04 -0.02

Note: Each cell contains the bias, the percentage bias and the standardized bias in order.

We report the empirical coverage of 95 % confidence intervals of the estimators in Table 6-13. Once more, ITT-type estimators, as well as the SRW, provide no coverage at all across all scenarios, given the small amount of variance implied by the study design relative to the substantial bias that exists. For the LATE-type estimators other than SRW, coverage is acceptable (between 0.942 and 0.970) when there is no overlap. In the cases of moderate non-compliance, LATE-type estimators other than SRW maintain

considerable coverage from .808 to .948. Poor coverage is observed when there are both substantial overlap and substantial non-compliance.

Table 6-13 Coverage of 95 Percent Confidence Intervals for FCACE

#	d =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	.000	.692	.000	.000	.000	.000	.544	.592
2	⅓	1/2	.000	.817	.000	.000	.000	.000	.748	.798
3	0	1/2	.000	.961	.000	.000	.000	.000	.962	.970
4	½	3/4	.000	.897	.000	.000	.000	.000	.808	.852
5	⅓	3/4	.000	.926	.000	.000	.000	.000	.882	.888
6	0	3/4	.000	.948	.000	.000	.000	.002	.942	.948

Note: Each cell contains the coverage of 95 % confidence intervals

We report MSE of the estimated FCACE in Table 6-14. Since the four ITT-type estimators are severely biased, their MSEs are overwhelmingly dominated by the bias while the standard errors of the estimators are negligibly small, as we previously discussed. Therefore, we observe much higher MSEs for ITT-type estimators. The NPI, MDPS and MCPS perform equally well with the lowest MSEs in all settings.

Table 6-14 Mean Squared Error of Estimated CACE

#	d =	p^c	ITT	NPI	REG	SEW	SPW	SRW	MDPS	MCPS
1	½	1/2	4.05	0.08	4.05	4.05	4.05	2.42	0.09	0.09
2	⅓	1/2	4.05	0.05	4.05	4.05	4.05	1.79	0.06	0.05
3	0	1/2	4.05	0.02	4.05	3.85	4.05	0.83	0.02	0.02
4	½	3/4	1.01	0.03	1.01	1.01	1.01	0.51	0.03	0.03
5	⅓	3/4	1.01	0.02	1.01	1.01	1.01	0.33	0.02	0.02
6	0	3/4	1.01	0.01	1.01	1.01	1.01	0.16	0.01	0.01

Note: Each cell contains the mean squared error of estimated FCACE.

As summarized in the previous section, the LATE-type estimators performed well in terms of bias and coverage, whereas the ITT-type estimators performed poorly with negatively biased estimation. Among the LATE-type estimators, the NPI, MDPS and MCPS perform equally well with minimal biases and the smallest mean squared errors when there is moderate non-compliance. The NPI is a slightly better performer than MDPS and MCPS when there is substantial non-compliance. The SRW is the worst performer among the four in all settings.

6.3 Conclusions and Limitations

In this chapter, we investigate the behaviour of the eight estimators under different scenarios using simulated data. The objective is to compare the performance of the proposed methods in estimating CACE and FCACE with respect to bias, mean-squared error of the estimators, and the empirical coverage of confidence intervals. We summarize our finding as follows. We demonstrate that three LATE-type methods (NPI, MDPS, and MCPS) result in estimation with minimal bias across all scenarios under both settings of all-or-none compliance and partial compliance. In the cases of substantial overlaps and/or non-compliance, they tend to overestimate the true values (see detailed discussions in the following section). In contrast, the ITT-type estimators (ITT, REG, SEW, and SPW) yield a very conservative and negatively biased estimate, as expected in all scenarios. The magnitude of bias increases with an increasing of overlap and/or non-compliance. The SRW estimator results in estimations not substantially biased in the cases of no overlap or moderate/mild non-compliance. Therefore, we recommend that MDPS and MCPS should be used in RCTs in the presence of partial non-compliance.

Several reasons contribute to our recommendation of the matching based algorithms. First, as we have shown in the simulation studies, MDPS and MCPS result in estimation with the least bias and smallest mean squared error. Second, it is attractive to applied researchers to be able to directly compare baseline characteristics of 'compliers' between the experimental group and the control group in the matched sample to assess the comparability between the two. Third, matching algorithms do not require that an outcomes model be correctly specified.

Although simulation shows that NPI performs similarly to (if not better than) MCPS and MDPS, in practice there is no way to verify the cut-offs of NPI. Therefore, sensitivity analysis must be used for inference purposes.

6.3.1 Limitations

We evaluate the performance of eight methods over a variety of populations and subsequently identify methods that seem to perform well under most scenarios. However, we must acknowledge that many key assumptions have been made throughout the simulation and there are many limitations to the current simulation studies. For instance, we do not simulate the performance of the methods when the key assumptions (e.g., SUTVA, exclusion restriction, no unmeasured confounders) are violated. These assumptions were made when the foundations of this methodology were developed, and will be reviewed in detail in Chapter 7. Since the primary goal of the simulation studies is to compare the relative performance for the proposed methods, we think it is reasonable to test them under a set of commonly accepted assumptions. There are also assumptions made specifically for the simulation studies, including the assumption of

equal variances for potential outcomes and equal proportion for strata other than compliers. Some of these assumptions may be overly simplistic and are reviewed in details as follow:

1. Assumption of normal distributions with a common variance for all potential outcomes across principal strata. We assume that all potential outcome variables follow normal distributions with a mean $\mu_{k,s}$ and a variance $\sigma^2_{k,s}$, where $k \in \{c, pnc, fpc, a, p, n\}$ and $s = 1$, ½, or 0 (we use c to replace *fnc* to simplify the notation for this discussion). We set $\sigma^2_{k,s} = 5^2$ for all k and s. Since our focus is to assess the accuracy of our estimation procedures, we feel confident that relaxing this assumption and allowing heterogeneity in the variance will not affect the relative performance of the eight estimators.
2. Assumption of 'no compliance effect for controls' (NCEC) (Little, Long and Lin, 2008). We assume that the mean outcome under the control treatment is the same (and equals zero) for compliers and never-takers, i.e., $\mu_{n,0} = \mu_{c,0} = 0$. This assumption implies that the mean outcomes are the same for subjects who do not receive the experimental treatment regardless of their compliance membership. NCEC has been considered strong and unacceptable in some cases because compliers and never-takers may differ on various characteristics related to the outcome under the control treatment. However, in our case, it is reasonable to assume that $\mu_{n,0} = \mu_{c,0} = 0$.
3. Relaxing the assumption of 'no compliance effect for treatment' (NCET) (Little, Long and Lin, 2008). This assumption asserts that the mean outcome under the experimental treatment is the same for compliers and always-takers. We do *not*

make this assumption in our simulation. Instead, we set $\mu_{a,1} = 8$ and $\mu_{c,1} = 5$, so always-takers have a larger mean than compliers do in the experimental group. We follow the argument of Lin and Rubin (2008) that always-takers might be the group who benefit most from treatment (e.g., sicker patients), so they take experimental treatment regardless of their treatment assignment.

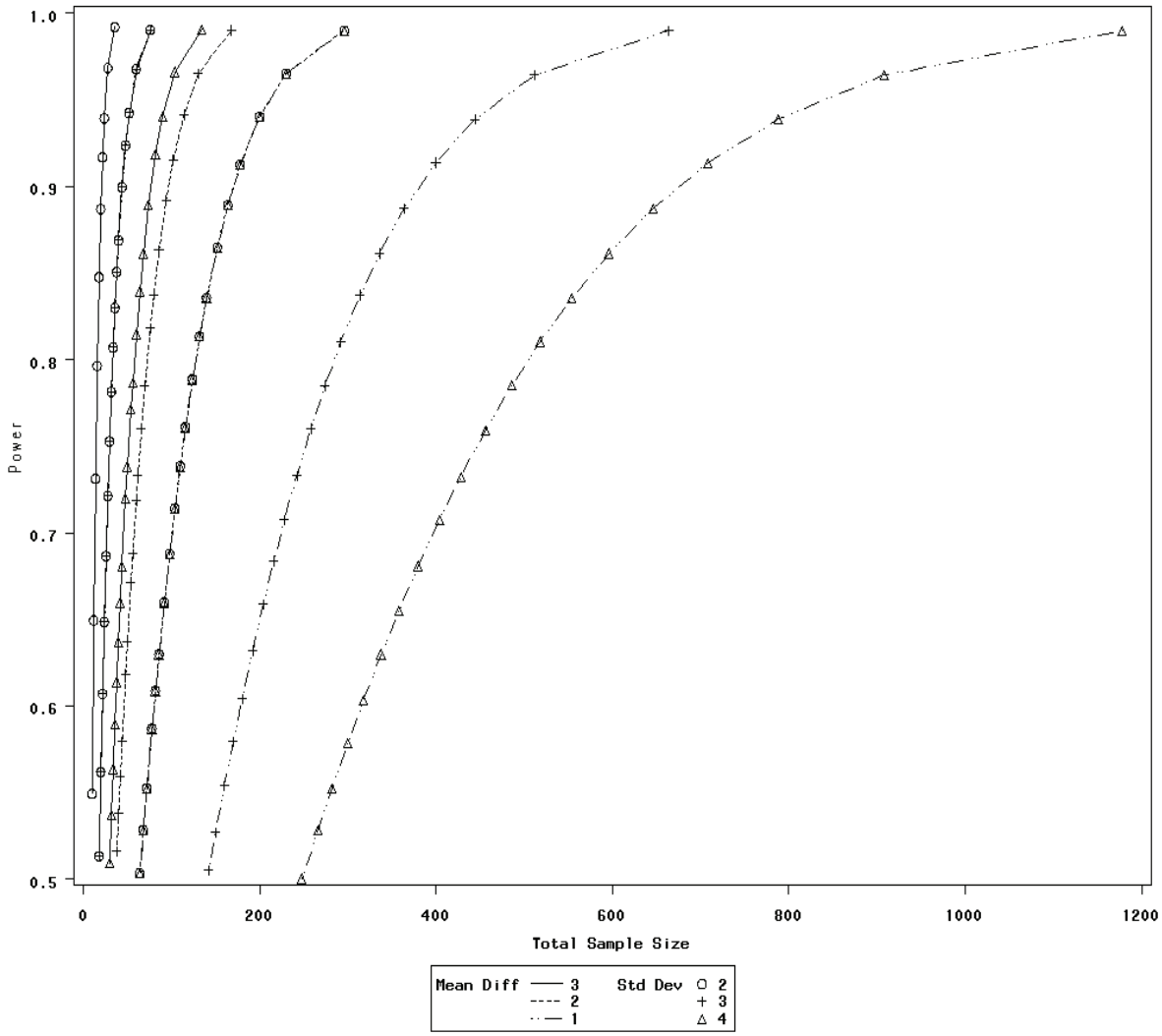
4. Assumption of independence of potential outcomes. We assume that the correlation between $Y_i(1, s)$ and $Y_i(0, s)$ is zero given s. Jin and Rubin (2008) conducted a sensitivity analysis with different values of correlation and concluded that the results changed only slightly compared with the results assuming no correlation.
5. Assumption of uniform distribution of r^1 and r^0. We assume that r^1 and r^0 follow uniform distributions under each principal stratum and we randomly sample r^1 and r^0 directly from pre-specified distributions rather than estimated from predictive models based on simulated covariates. Our reason for not simulating covariates directly is that the focus of the simulations is to validate and compare the proposed methods. We believe that the whole process of simulating covariates and exposure, building up the predictive models, and estimating the dual propensity scores, will add another layer of complexity and my weaken the evidence and interpretation of the results. Relaxing this assumption will increase the variability of r^1 and r^0 and impact the performance of all estimators except ITT. However, we evaluate the cases of 'substantial overlap' scenarios in our simulations, which consider r^1 and r^0 are mixed for different principal strata. So there is no reason to believe that the conclusion will be different and we believe

matching-based algorithm may still be the best choice. Another limitation is that we use the same cut-off (e.g., 0.5) for r^1 and r^0. In practice, the cut-offs have to be determined based on the distributions of r^1 and r^0, respectively. And it is likely that they would differ from each other.

6. Assumption of an equal proportion. We set $p^a = p^p = p^n = p^{pnc} = p^{fpc}$ under the partial compliance and $p^a = p^n$ under all-or-none compliance for simplicity. Relaxing this assumption may affect the estimates of CACE and FCACE when there is substantial misclassification. However, we do not expect any significant impact on the relative performance of eight estimators.

In conclusion, DPS- and CPS-based methods allow one to estimate CACE and FCACE in the presence of all-or-none compliance or partial compliance, and with always-takers. We recommend that matching-based estimators (MDPS and MCPS) be used in RCTs with complicated non-compliance issues.

Figure 6-1 Power versus Sample Size for Simulation Study



Chapter 7 CONCLUSIONS

In this thesis, we have demonstrated the use of dual propensity scores and principal stratification theory in RCTs in the presence of partial non-compliance. First, we defined compliance stratification effects based on principal stratification theory and presented theoretically justified strategies to estimate the compliance stratification effects. Second, we developed the dual propensity scores approach as a tool to identify compliance strata and to estimate the compliance stratification effects. Third, we evaluated the performance of eight methods through simulations and identified that a matching-base algorithm performed well across different scenarios. In addition, we applied newly developed methods to PROBIT and concluded that prolonged and exclusive breastfeeding had positive effects on infant growth for the first 3 months, but that differences disappeared (or even reversed) by 12 months of age.

7.1 Contributions

This research project focuses on a situation in RCTs where non-compliance occurs in the form of partial compliance and can occur in both treatment groups - subjects in the experimental group may take only a portion of the experimental treatment, while subjects in the control group may somehow also obtain the experimental treatment, fully or partially. This would not be the case in many clinical trials of new drugs or other experimental treatments in which subjects randomized to either group usually have no access to the 'other' treatment. Often, non-compliance only occurs in the treatment arm, while subjects in the control group (i.e., placebo group) have no choice but to comply

because it is extremely unlikely those subjects would receive the experimental treatment. Although we agree that it is one of the key limitations of our proposed DPS methods, we consider it as a strength rather than weakness. Actually, many experimental studies face the situation where non-compliance can occur in either treatment group, especially studies with so-called encouragement design (e.g., Hirano, Imbens et al., 2000; Ten Have, Elliott et al., 2004). In addition, our methods are not limited to estimating 'compliance' principal effects; they can be easily apply to estimate general principal effects with a post-treatment variable S^{obs} no longer 'compliance.' In those cases, it is common that a subgroup, which is equivalent to always-takers, does exist.

The original methodological contributions of this thesis include: adapting the principal stratification to partial compliance and defining compliance principal strata; developing the innovative dual propensity score framework to identify compliance principal strata; determining proofs of sub-theorems related to the properties of the proposed estimators, adapting ordinal logistic regression to dual propensity score estimation; developing weighting by stratum ranking and matching the algorithm to identify compliance principal strata with dual propensity scores and to estimate principal effect. Weighting is key to producing a LATE-type estimate without explicitly identifying the principal compliance strata, and it is very close to the inverse-probability-of-treatment weighting used in observational studies, which links to structural nested mean models (Robins et al., 2004). On the other hand, matching strategies allow the identification of pre-defined compliance strata and lead to the estimations of LATE-type estimators (i.e., CACE and FCACE).

The DPS methodology is quite different from the standard propensity score approaches in many ways. First, the aim of the single propensity score is to balance all observed and relevant covariates while the aim of the DPS is to identify compliance strata. Second, the DPS is two-dimensional and contains two scores - actual and counterfactual - for each subject while the single propensity score is one-dimensional, with a single score for each subject. Third, conventional propensity score methodology in RCTs ignores treatment assignment (randomization) and focuses on the observed treatment exposure directly (as-treated approach), while DPS methodology considers the treatment assignment as an instrumental variable and keeps randomization intact. Finally, the ultimate interest in estimating of propensity scores is the marginal causal effect (ITT-type) while the ultimate interest of DPS is the conditional causal effects (CACE or FCACE). We think ITT-type estimate alone may be inadequate to answer explanatory questions regarding efficacy. Sheiner and Rubin (1995), among others, go even further by claiming that assessment of efficacy that accounts for subject compliance is more important than an assessment of effectiveness by an ITT analysis, and that the latter provides a biased estimate of the former.

7.2 Assumptions and Limitations

Although we conclude that the DPS-based methods offer a practical technique for assessing principal effects, there are many assumptions being made and certain limitations exist. Similar to all models based on potential outcomes framework and principal stratification theory, our methods require strong assumptions to be appropriately

specified, including SUTVA, randomization, exclusion restriction, monotonicity, weakly unconfoundedness, among others. We review the key assumptions as following:

1. SUTVA. The assumption of 'no interference between units,' part of SUTVA is needed to define compliance principal effects for individual units without reference to other individuals in the study. Departure from this assumption may occur when treatments are administrated at the provider level, such as in the PROBIT study. The consistency assumption of SUTVA is needed for estimating the effects by linking the potential outcomes to the observed outcomes. It implies that the observed outcome variable will equal one of the potential outcome variables even if the administration of treatment assignment and treatment itself vary slightly (Rubin, 1986). Violations of this assumption may occur when there are different forms of treatment administration and/or treatment itself. Although SUTVA is implausible for most of RCTs, it is not clear how to address violations of SUTVA and little research about this has been done.
2. Randomization. The randomization assumption is necessary for estimation of compliance principal effects in combination with SUTVA, and to relate the models for the observed outcomes to the models of their respective potential variables. The DPS-based methods retain the randomization and comparisons are always made between two randomized (sub-) groups. In RCT settings, this assumption always holds true.
3. The exclusion restriction (ER) assumption. This assumption implies that any effect of treatment assignment on the potential outcomes must be exclusively through the actual treatment received for the whole population. ER is required so

that the causal effects are zero for always-takers, never-takers and partial takers. When the ER is unlikely, the causal effect of treatment becomes unidentifiable (Robins and Rotnitzkv, 2004). However, using DPS matching methods, relaxing this assumption may not be so critical as long as the matching is successfully done and principal strata can be identified. Nonzero causal effect in one stratum (e.g., always-takers) would not affect the causal effect in other strata (e.g., compliers).

4. Monotonicity assumption. This assumption is required to rule out the existence of defiers. Relaxing this assumption increases the number of the compliance principal strata, especially under partial compliance, and makes the complier stratum unobservable. It is reasonable to believe that this assumption holds true for most RCTs. However, Ten Have et al. (2004) indicated that departure from this assumption may exist, as they showed with their examples.
5. Weak unconfoundedness assumption. This assumption states that if receipt of treatment and outcome are independent given the observed covariates, then treatment receipt and outcome are independent given dual propensity score. This is the key assumption to allow the efficacy of a treatment to be able to be estimated by adjusting for DPS. This assumption is an equivalent to a commonly made 'no unmeasured confounder' assumption.
6. Correctly specified models assumption. Even when all relevant confounders have been measured, an unbiased estimate can be obtained *only if* the model itself reflects the true relationship among treatment exposure and confounders. Outside of simulation studies, we can never know whether or not the model we have constructed accurately depicts those relationships. Therefore, correct

specification of predicted model to estimate DPS is typically an unverifiable assumption. On the other hand, the DPS methods use the exact same model for outcomes as the ITT does. Therefore, the DPS methods do not require an additional outcomes model be 'correctly' specified. However, when covariates need to be included, the DPS methods also require a correctly specified model for the outcome with respect to covariates, especially when specific principal strata are expected to have a different outcome: covariates and treatment relationship compared with the relationship among the whole population.

7. Free-of-measurement-error assumption. The DPS methods entail additional and accurate information about compliance or treatment exposure, and covariates. Without the correct information, the DPS methods will not be able to provide reliable statistical estimates as they are supposed to.

The DPS methods require strong assumptions and the fact that most of these assumptions are unverifiable suggests caution should be taken in the implementation of the proposed methods. In addition, there are other limitations. One limitation is that the DPS methods are mostly useful when there are complicated non-compliance issues. However, in many drug trials, non-compliance only occurs in the treatment arm, while subjects in the control group have no access to the experimental treatment. In this case, CACE can be estimated more straightforwardly or even be identifiable, and the new methods may not be useful. Another limitation is that so far we ignore sampling variability by restricting attention to estimands, so any extrapolation from the individuals in the trial sample to the population may be questionable. One point we want to

emphasize is that the DPS methods work only when the assignment mechanism is truly unconfounded, given the observed covariates and all covariates related to the treatment receipt should have been collected. Therefore, plan should be in place at the design stage in order to use the DPS-based approaches at the analysis stage.

7.3 Applications and Extensions

DPS methodology can be generalized to all applications of principal stratification, including surrogate endpoints, biomarkers, direct and indirect causal effects, and censoring by death. In the recent work, other than the non-compliance issue using principal stratification, Barnard et al. (2003) estimate effects of school vouchers in student performance; Hill et al. (2003) evaluate the effects of high participation in an early intervention for low-birth-weight premature infants; and Zhang and Rubin (2003) show how to address censoring of outcomes by death.

Extensions to binary and survival outcomes will be straightforward. Once the principal strata have been identified, the same model which is used to obtain the ITT estimator can be easily applied to the newly identified subgroups. Another extension of DPS methods is to model time-varying or dynamic treatment receipt. One possible approach is to model DPS at each interval over time, using baseline covariates and covariates observed prior to this interval. Thus, a sequence of DPS, instead of one set of DPS at the baseline, will be used to estimate the principal compliance effects at each time interval. For each subject, the membership to compliance strata may change from one time to another. Recently, Moodie et al. (2008) reanalyzed PROBIT data using methods

related to optimal dynamic treatment regimes in order to consider the effect of breastfeeding on infant growth at one year of age.

In conclusion, DPS- and CPS-based methods allow one to estimate CACE and FCACE in the presence of all-or-none compliance or partial compliance, and with always-takers. We recommend that matching-based estimators (MDPS and MCPS) be used in RCTs with complicated non-compliance issues. We believe the dual propensity score approach is an innovative and useful tool to estimate the principal compliance effects in RCTs with partial non-compliance.

BIBLIOGRAPHY

Agresti, A. (2002). *Categorical Data Analysis* (2nd edn). New York:Wiley, Inc.

Albert, J. M., and Demets, D. L. (1994). On a model-based approach to estimating efficacy in clinical trials. *Statistics in Medicine* **13**, 2323-2335.

Angrist, J., and Imbens, G. W. (1995). Two-stage least squares esstimation of average causal effects in models with variable treatment intersity. *Journal of the American Statistical Association* **90**, 431-442.

Angrist, J., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91,** 444-472.

Armitage, P. (1998). Attitudes in clinical trials. *Statistics in Medicine* **17**, 2675-2683.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007a). 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.

Austin, P. C., Grootendorst, P., Normand, S. L. T., and Anderson, G. M. (2007b). Conditioning on the propensity score can result in biased estimation of common meausure of treatment effect: A Monte Carlo study. *Statistics in Medicine* **26**, 754-768.

Austin, P. C. and Mamdani, M. M. (2006). A Comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* **25**, 2084-2106.

Austin, P. C., Mamdani, M. M., Stukel, T. A., Anderson, G. M., and Tu, J. V. (2005). The use of the propensity score for estimating treatment effects: Administrative versus clinical data. *Statistics in Medicine* **24**, 1563-1578.

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* **26**, 3078-3094.

Austin, P. C. (2008). A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* **27**, 2037-2049.

Austin, P. C. (2009). Some methods of propensity score matching had superior perfromance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal* **51**, 1-14.

Balke, A., and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92**, 1171-1176.

Bang, H., and Davis, C. E. (2007). On estimating treatment effects under non-compliance in randomized clinical trials: are intent-to-treat or instrumental variable analyses perfect solutions? *Statistics in Medicine* **26**, 954-964.

Bang, H., and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 692-972.

Barnard, J., Frangakis, C. E., Hill, J., and Rubin, D. B. (2003). A principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City (with discussion). *Journal of the American Statistical Association* **98**, 299-323.

Bellamy, S. L., Lin, J. Y., and Ten Have, T. R. (2007). An introduction to causal modeling in clinical trials. *Clinical Trials* **4**:58-73.

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Sturmer, T. (2006). Variable selection for propensity score models. *Amearican Journal of Epidemiology* **163**, 1149-1156.

Brumback, B. A., Hernán, M. A., Haneuse, S. J. P. A., and Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measure. *Statistics in Medicine* **23**, 749-767.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine* **25**,4279-4292.

Cochrane, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**, 295-313.

Cox, D. R. (1992). Causality: some statistical aspects. *Journal of the Royal Statistical Society A* **155**, part 2, 291-301.

D'Agostino Jr., R. B. (1998). 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.

D'Agostino Jr., R. B., and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* **95**, 749-759.

Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association* **95**, 407-448.

Dehejia, R. H., and Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053-1062.

Dehejia, R. H., and Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics* **84**, 151-161.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1-38.

Dewey, K., Heinig, M., Nommsen, L., Peerson, J., and Lonnerdal, B. (1992). Growth of breast-fed and formula-fed infants from 0 to 18 months: the DARLING study. *Pediatrics* **89**, 1035-1041.

Donner, A., and Klar, N. (2000). *Cluster randomization trials in health research.* Arnold: London, 2000.

Drake, C. (1993). Effects of misspcification of the propensity score on estimators of treatment effect. *Biometrics* **49**, 1231-1236.

Duncan, O. D. (1974). *Introduction to Structural Equation Models*. New York: Academic.

Efron, B., and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* **86**, 9-26.

Efron, B., and Tibshirani, R. J. (1998). *An introduction to the Boostrap.* Chapman & Hall: London, 1998.

Fischer-Lapp, K., and Goetghebeur, E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials* **20**, 531-546.

Follmann, D. A. (2000). On the effect of treatment among would-be treatment compliers: an analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association* **95**, 1101-1109.

Frangakis, C. E., and Baker, S. G. (2001). Compliance subsampling designs for comparative research: estimation and optimal planning. *Biometrics* **57**, 899-908.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association* **99**, 365-379.

Frangakis, C. E., and Rubin, D. B. (1999). Addressing compliations of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent mssing outcomes. *Biometrika* **86**, 365-379.

Frangakis, C. E., and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21-29.

Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms (with discussion). *Biostatistics* **3**,147-164.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., Ten Have T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine* online, DOI: 10.1002/sim.3533.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulations using multiple sequences. *Statistical Science* **7**, 457-511.

Goetghebeur, E., and Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* **91**, 928-934.

Goetghebeur, E., and Shapiro, S. H. (1996). Analysing noncompliance in clinical trials: ethical imperative or mission impossible? *Statistics in Medicine* **15**, 2813-2826.

Goetghebeur, E., and van Houwelingen, H. C., eds. (1998). Analyzing noncompliance in clinical trials. *Statistics in Medicine* **17**, 247-389.

Greenland, S. (2000a). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29**, 722-729.

Greenland, S. (2000b). Causal analysis in the health sciences. *Journal of the American Statistical Association* **95,** 286-289.

Greenland, S., and Brumback, B. A. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology* **31**, 1030-1037.

Greenland, S., Lanes, S., and Jara, M. (2008). Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. *Clinical Trials* **5**: 5-13.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37-48.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statisical Science* **14**, 29-46.

Greevy, R., Silber, J. H., Cnaan, A., and Rosenbaum, P. R. (2004). Randomization inference with imperfect compliance in the ACE-Inhibitor after anthracycline randomized trial. *Journal of the American Statistical Association* **99,** 7-15.

Gu, X. S., and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*. **2**, 405-420.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315-331.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Heckman, J. J. (1996). Comment on " Identification of causal effects using instrumental variables". *Journal of the American Statistical Association* **91,** 459-462.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies* **64**, 605-654.

Heitjan, D. F. (1999). Ignorability and bias in clinical trial. *Statistics in Medicine* **18**, 1821-2434.

Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health* **58**, 265-271.

Hernán, M. A., Brumback, B. A., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* **96,** 440-448.

Hernán, M. A., and Robins, J. M. (2006a). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* **60**, 578-586.

Hernán, M. A., and Robins, J. M. (2006b). Instruments for causal inference: an epidemiologist's dream? *Epidemiology* **17**, 360-372.

Hill, J. L., Brooks-Gunn, J., and Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology* **39**, 730-744.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161-1189.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69-88.

Holland, P. (1986). Statistics and Causal Inference (with discussion). *Journal of the American Statistical Association* **81**, 945-970

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of Royal Statistics Society. A*. **171**: 481-502.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **83**, 706-710.

Imbens, G. W., and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467-476.

Imbens, G. W., and Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals Statistics* **25**, 305-327.

Imbens, G. W., and Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies* **64**, 555-574.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association* **103,** 101-111.

Joffe, M., and Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology* **150**, 327-333.

Joffe, M. M., Ten Have, T. R., and Brensinger, C. (2003). The compliance score as a regressor in randomized trials. *Biostatistics* **4**, 327-340.

Kenna, L. A., and Sheiner, L. B. (2004). Estimating treatment effect in the presence of non-compliance measured with error: precision and robustness of data analysis methods. *Statistics in Medicine* **23**, 3561-3580.

Kramer, M. S., Chalmers, B., Hodnett, E. D., Sevkovskaya, Z., Dzikovich, I., Shapiro, S. et al. (2001). Promotion of breastfeeding intervention trial (PROBIT): A randomized trial in the Republic of Belarus. *Journal of the American Medical Association* **285**, 413-420.

Kramer, M. S., Chalmers, B., Hodnett, E. D., Sevkovskaya, Z., Dzikovich, I., Shapiro, S. et al. (2002). Promotion of breastfeeding intervention trial (PROBIT): A cluster-randomized trial in the Republic of Belarus. In: Koletzko B, Michaelsen KF, Hernell O, editors. *Short and long term effects of breastfeeding on child health*. New York: Kluwer Academic/Plenum Publishers.

Kramer, M. S., Guo, T., Platt, R. W., Shapiro, S., Collet, J., Chalmers, B., Hodnett, E., Sevkovskaya, Z., Dzikovish, I., and Vanilovish, I. (2002). Breastfeeding and infant growth, biology or bias? *Pediatrics* **110**, 343-347.

Kramer, M. S., Guo, T., Platt, R. W., Sevkovskaya, Z., Dzikovish, I., Collet, J., Shapiro, S., Chalmers, B., Hodnett, E., and Vanilovish, I., Mezen, I., Ducruet, T., Shishko, G., and Bogdanovich, N. (2003). Infant growth and health outcomes associated with 3 compared with 6 months of exclusive breastfeeding. *American Journal of Clinical Nutrition* **78**, 291-295.

Kramer, M. S., Guo, T., Platt, R. W., Sevkovskaya, Z., Dzikovish, I., Collet, J., Shapiro, S., Chalmers, B., Hodnett, E., and Vanilovish, I., Mezen, I., Ducruet, T., Shishko, G., and Bogdanovich, N. (2004a). Does previous infection protect against atopic eczema and recurrent wheeze in infancy? *Clinical and Experimental Allergy* **34**, 753-756.

Kramer, M. S., Guo, T., Platt, R. W., Vanilovich, I., Sevkovskaya, Z., and Dzikovish, I., Michaelsen, K. F., and Dewey, K. (2004b). Feeding effects on growth during infancy. *Journal of Pediatrics* **145**, 600-605.

Last, J. M., Spasoff, R. A., Harris, S. S., and Thuriaux, M. C. (2000). eds. *A dictionary of epidemiology.* 4th edition. Oxford University Press.

Lee, J., Ellenberg, J., Hirtz D, and Nelson, K. (1991). Analysis of clinical trials by treatment actually received: is it really an option? *Statistics in Medicine* **10**,1595-1605.

Lefebvre, G., Delancy, J. A. C., and Platt, R. W. (2008). Impact of mis-specification of the treatment model on estimates from a marginal structure model. *Statistics in Medicine* **27**,3629-3642.

Levis, J. A., and Machin, D. (1993). Intention-to-treat – who should use ITT? *British Journal of Cancer* **68**, 647-650.

Li, F., Frangakis, C. E., and Varadhan, R. (2004). Polydesigns for partially controlled studies: motivation and definition. *American Statistical Association Proceedings of the Biopharmaceutical Section.*

Lin, D. Y., Fleming, T. R., and De Gruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **16**,1515-1527.

Lin, J. Y., Ten Have, T. R., and Elliott, M. R. (2008). Longitudinal nested compliance class model in the presence of time-varying noncompliance. *Journal of the American Statistical Association* **103**, 462-473.

Little, R. J. A., Long, Q., and Lin, X. H. (2008). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics* **64**, 118-124.

Little, R. J. A., and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review Public Health* **21**: 121-145.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data,* 2nd eds. New York: Wiley.

Little, R. J. A., and Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psych methods* **3**, 147-159.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* **96**, 1245-1253.

Lunceford, J. K., and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23**, 2937-2960.

McCullagh, P. (1980). Regression Models for ordinal data (with discussion). *Journal of the Royal Statistical Soceity Series B* **42**, 109-142.

Ming, K., and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56**, 118-124.

Mealli, F., Imbens, G. W., Ferro, S., and Biggeri, A. (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* **5**: 207-222.

Moodie, E. E. M., Platt, R. W., Kramer, M. S. (2008). Estimating response-maximized decision rules with applications to breastfeeding. *Journal of the American Statistical Association* in press.

Nagelkerke, N., Fidler, V., Bernsen, R., and Borgdor, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* **19**: 1849-1864.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. English translation of excerpts by D. M. Dabrowska and T. P. Speed (1990). *Statistical Science* **5**, 465-480.

Nielsen, G., Thomsen, B., and Michaelsen, K. (1998). Influence of breastfeeding and complementary food on growth between 5 and 10 months. *Acta Paediatr* **87**, 911-917.

O'Malley, A. J., and Normad S. T. (2005). Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. *Biometrics* **61**: 325-334.

Parson, L. S. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques. *Proceeding of the Twenty-sixth Annual SAS Users Group International Conference.* SAS Institute Inc., 214-216.

Pearl, J. (2000). *Causality.* Cambridge University Press.

Peng, Y., Little, R. J. A., and Raghunathan, T. E. (2004). An extended general location model for causal inference from data subject to noncompliance and missing values. *Biometrics* **60**, 598-607.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431-440.

Robins, J. M. (1994). Correcting for noncompliance in randomized trials using structural nested mean models. *Communications in Statistics* **23**, 2379-2412.

Robins, J. M. (1997). Causal inference from complex longitudinal data. *Latent Variable Modeling with Applications to Causality*. ed. M. Berkane, New York: Springer-Verlag.

Robins, J. M. (1998a). Correction for non-compliance in equivalence trials. *Statistics in Medicine* **17**, 269-302.

Robins, J. M. (1998b). Structural nested failure time models. *The Encyclopedia of Biostatistics*, P. Armitage and T. Colton, Editors, 4372-4389. New York: Wiley.

Robins, J. M. (1999a). Association, causation, and marginal structural models. *Synthese* **121**, 151-179.

Robins, J. M. (1999b). Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology*: *The Environment and Clinical Trials.* Halloran, M. E. and Berry, D., eds. NY: Springer-Verlag, pp. 95-134.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* **1999**; 6-10.

Robins, J. M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3,** 143-155.

Robins, J. M., and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association* **89,** 737-749.

Robins, J. M., and Greenland, S. (1996). Comment on: Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 456-458.

Robins, J. M., Greenland, S., and Hu, F. C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association* **94,** 687-712.

Robins, J. M., Hernán, M. A., and Brumback, B. A. (2000). Marginal structural models and causal inference in epidemiology. E*pidemiology* **11,** 550-560.

Robins, J. M., and Rotnitzky, A. (2004). Estimation of treatment effects in randomized trials with non-compliance and dictomous outcome using structural mean models. *Biometrika* **91**, 763-783.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846-866.

Robins JM, Scharfstein D, Rotnitzky A. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. *Statistical Models in Epidemiology*: *The Environment and Clinical Trials.* Halloran, M. E. and Berry, D., eds. NY: Springer-Verlag, pp. 1-94

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* **79**, 565-574.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association* **84**, 1024-1032.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55.

Rosenbaum, P. R., and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**:516-524.

Rosenbaum, P. R., and Rubin, D. B. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33-38.

Rosenbaum, P. R., and Rubin, D. B. (1985b). The bias due to incomplete matching. *Biometrics* **41**, 103-116.

Rothman, K. J., and Greenland, S. (1998). *Modern Epidemiology*, 2nd eds. Philadelphia: Lippincott.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29**, 185-203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688-701.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* **6**; 34-58.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74,** 318-324.

Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics* **36**, 293-298.

Rubin, D. B. (1986). Statistics and causal inference. Comment: which ifs have causal answers. *Journal of the American Statistical Association* **81,** 961-962.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Stat. Plan. Inf.* **25**, 279-292.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **47**, 1213-1234.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91,** 473-520.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127**, 757-763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* **2**, 169-188.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association* **100,** 322-331.

Rubin, D. B. (2007a). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* **26**, 20-36.

Rubin, D. B. (2007b). Causal inference through potential outcomes and prinicipal stratification. *Statistical Science* **21**, 299-309.

Rubin, D. B., and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics* **52**, 249-264.

Rubin, D. B., and Thomas, N. (2000). Combining propensity score mathcing with additional adjustment for prognostic covariates. *Journal of the American Statistical Association* **95,** 573-585.

SAS Institue Inc. (2002-2003). *SAS Version 9.1*. SAS Institute, Inc., Cary, NC.

Sheiner, L. B., and Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clin Pharm Therap* **57**, 6-15.

Small, D., Ten Have, T. R., Joffe, M. M., Cheng, J. (2006). Random effects models for analysing efficacy of a longitudinal randomzied treatment with non-adherence. *Statistics in Medicine* **25**, 1981- 2007.

Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27**, 325-353.

Smith, J., and Todd, P. (1997). Dose matching overcome LaLonde's critique of non-experimental estimators? *Journal of Econometrics*, **125**, 305-353.

Sobel, M. E. (1990). Effect analysis and causation in linear structure equation models. *Psychometrika* **55**, 495-515.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In *Handbook of Satistical Modeling for the Social and Behavioral Sciences.* ed. Arminger. New York: Plenum.

Sommer, A., and Zeger, S. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10**, 45-52.

Tanner, M., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528-550.

Taylor, L., and Zhou, X. H. (2008). Multiple imputation methods for treatment non-compliance and nonreponse in randomized clinical trials. *Biometrics* **64**, 1-8.

Ten Have, T. R., Joffe, M. M., Lynch, K., Maisto, S., Brown G., and Beck, A. (2007). Causal mediation analyses with rank preserving models. *Biometrics* **63**, 926-934.

Ten Have, T. R., Elliott, M. R., Joffe, M. M., Zanutto, E., and Datto, C. (2004). Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association* **99**, 16-25.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals Statistics* **22**, 1701-1762.

Wei, G., and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association* **85,** 699-704.

White, I. R., (2005). Uses and limitations of randomized-based efficacy estimators. *Statistical Methods in Medical Research* **14**: 327-347.

Yau, L., and Little, R. J. A. (2001). Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* **96**, 1232-1244.

Zeger, S. (1998). Adjustment for non-compliance. in *Encyclopedia of Biostatistics 4*, eds. P. Armitage and T. Colton, New York: Wiley, pp. 3006-3009.

Zhang, J. L., and Rubin, D. B. (2003). Estimation of causal effects via principal stratificaiton when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics* **28**: 353-368.

Zhou, X. H., and Li, S. (2005). ITT analysis of randomized encouragement design studies with missing data. *Statistics in Medicine* **23:** 1991-2003.

APPENDIX: FIGURES - FOREST PLOTS BY MONTH

Figure A1-1 Effect of Prolonged and Exclusive BF on Weight Gain (g) at 1 Month

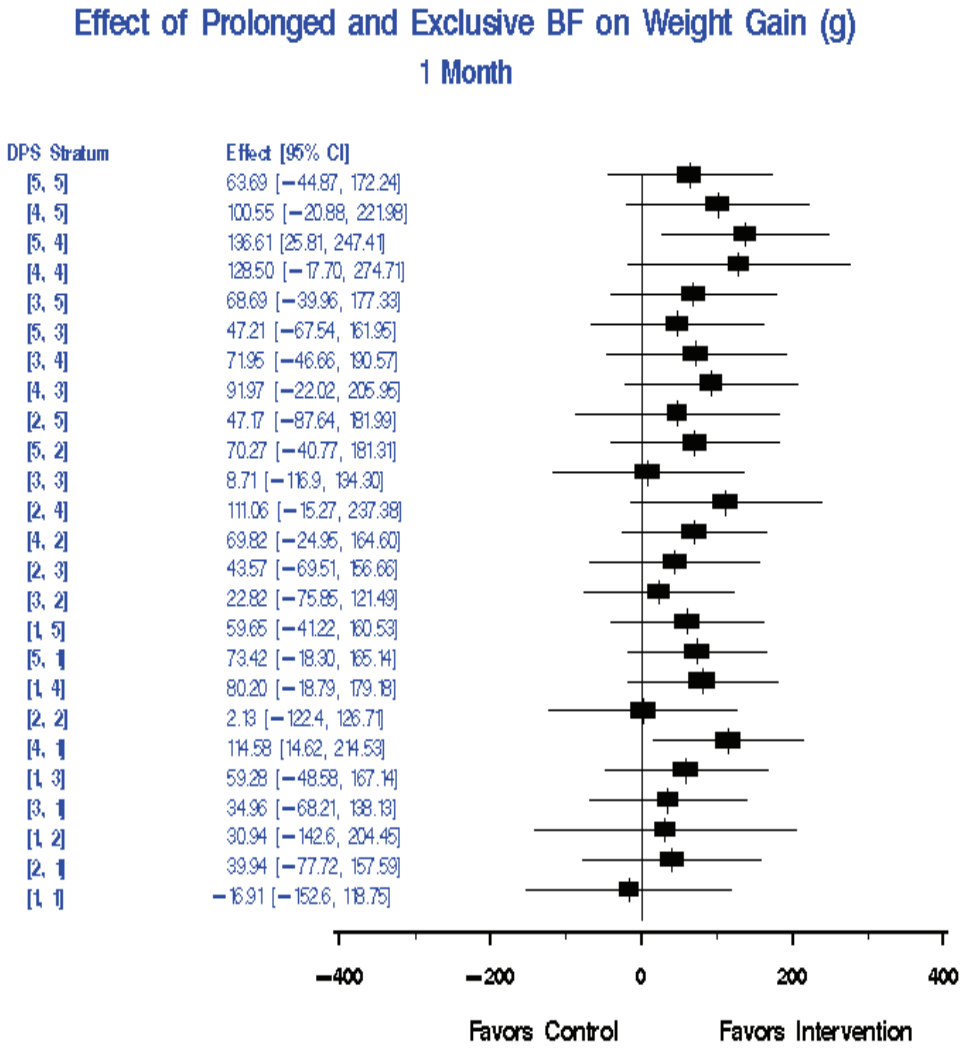


Figure A1-2 Effect of Prolonged and Exclusive BF on Weight Gain (g) at 2 Months

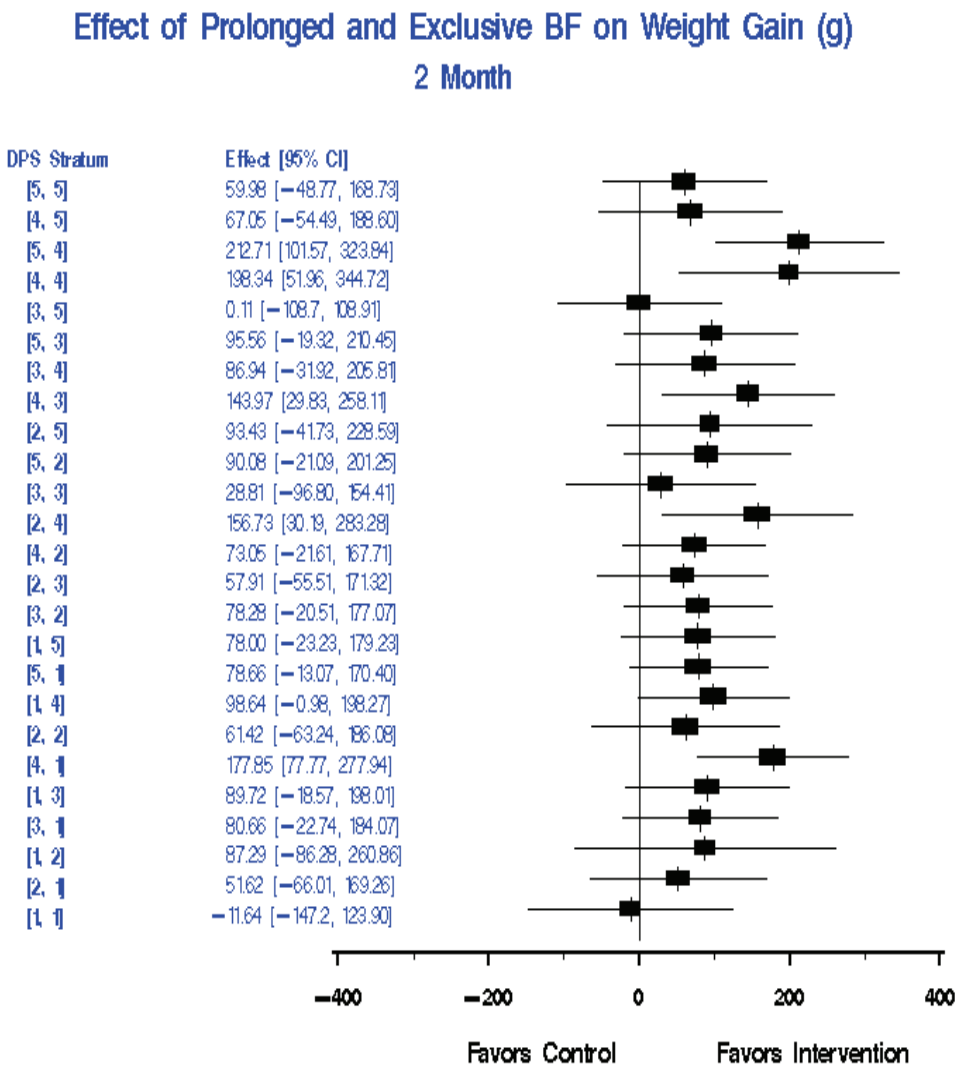


Figure A1-3 Effect of Prolonged and Exclusive BF on Weight Gain (g) at 3 Months

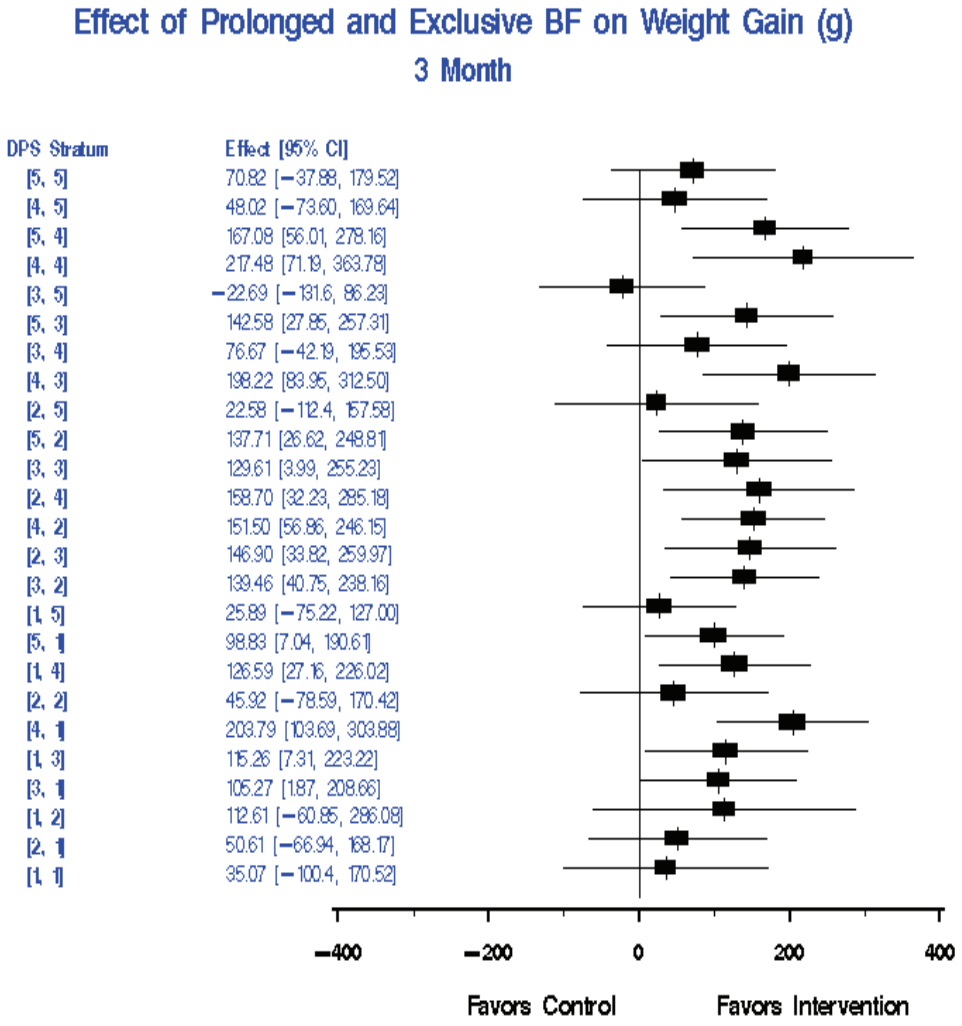


Figure A1-4 Effect of Prolonged and Exclusive BF on Weight Gain (g) at 6 Months

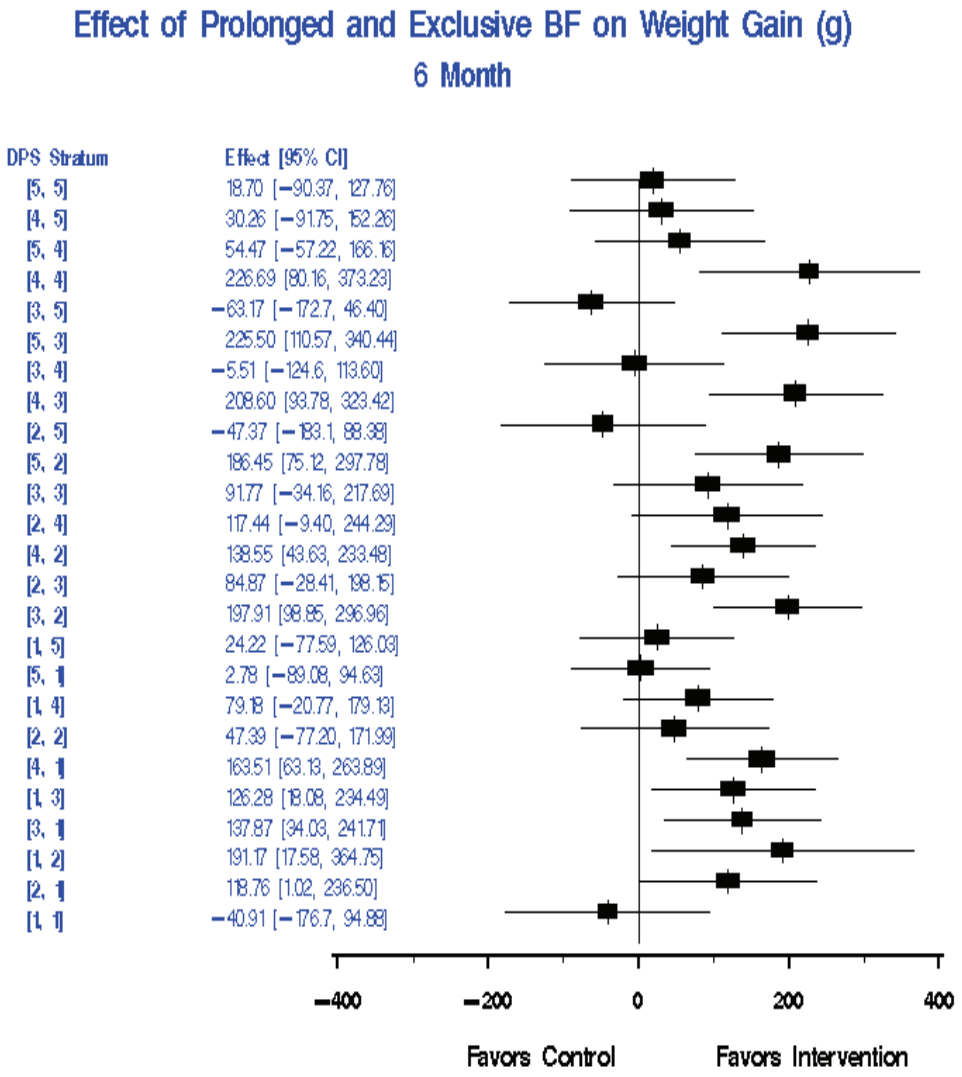


Figure A1-5 Effect of Prolonged and Exclusive BF on Weight Gain (g) at 9 Months

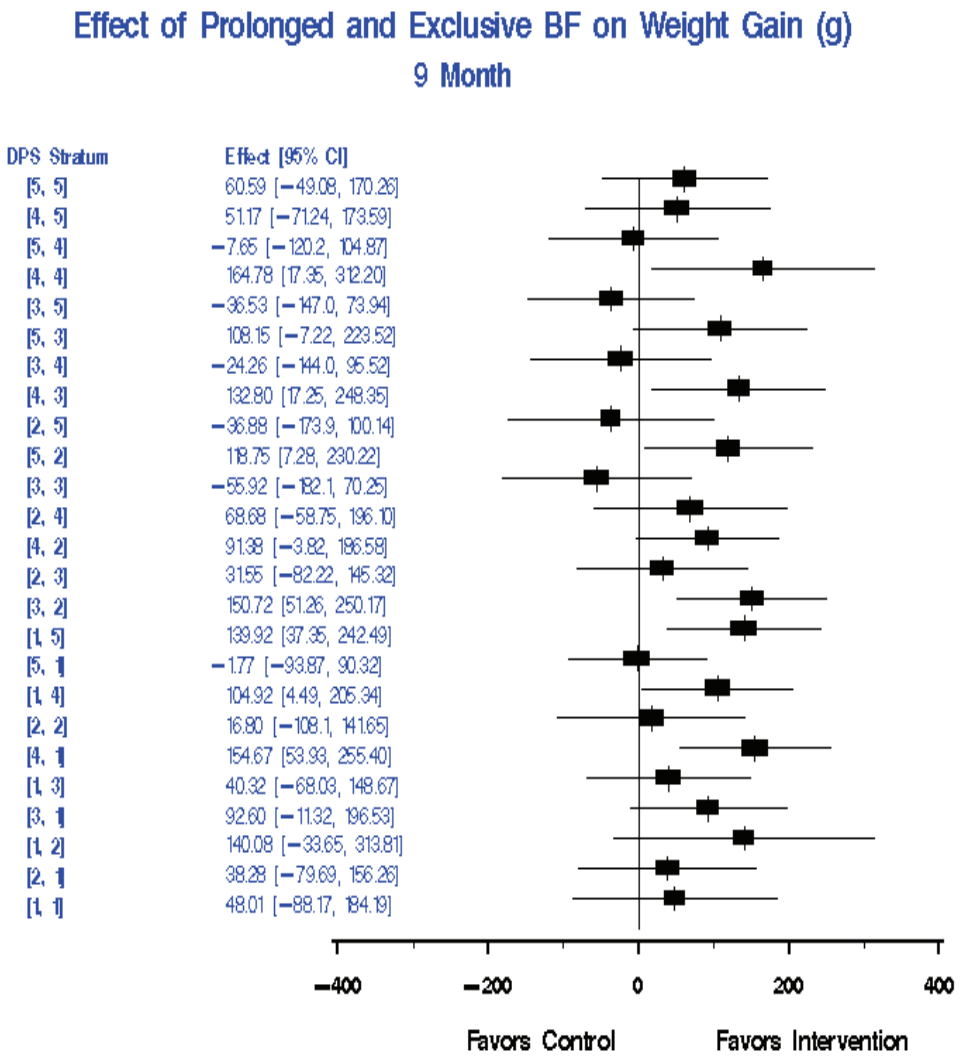


Figure A1-6 Effect of Prolonged and Exclusive BF on Weight Gain (g) at 12 Months

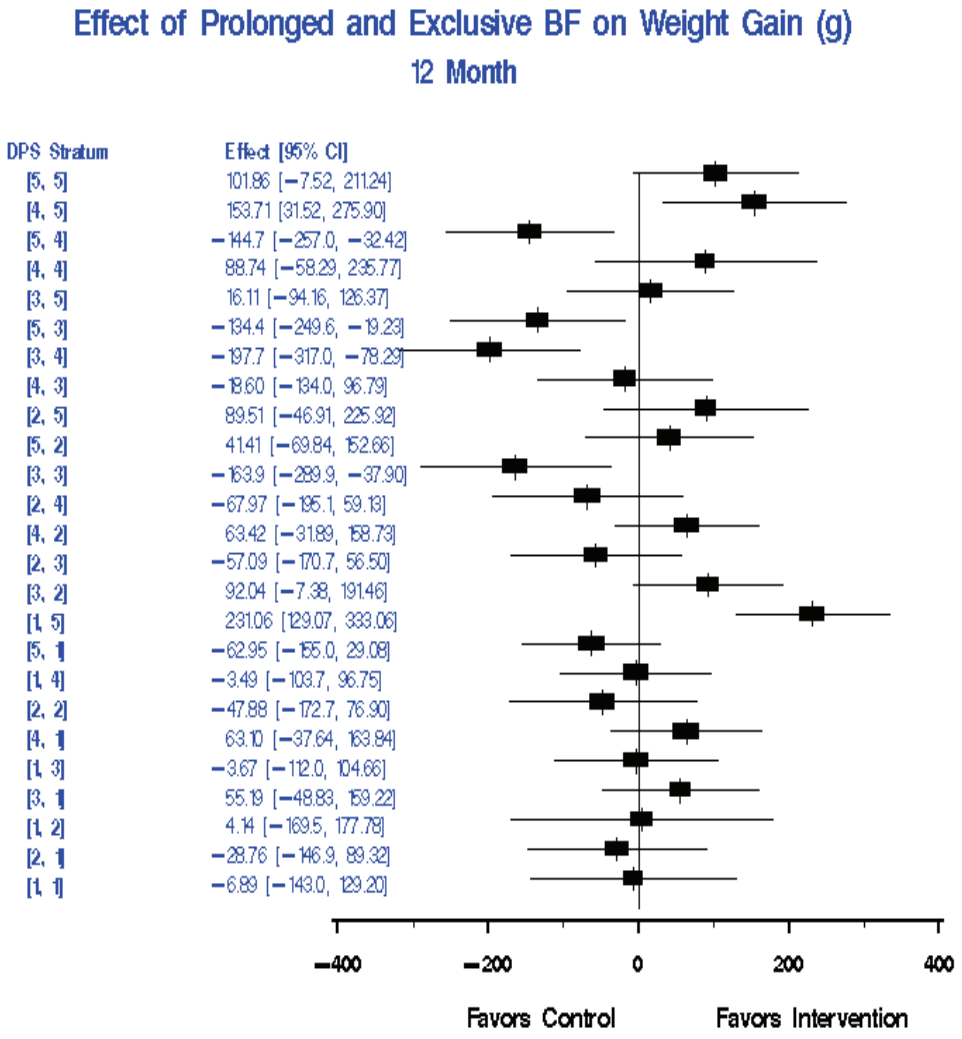


Figure A1-7 Effect of Prolonged and Exclusive BF on Length Gain (cm) at 1 Month

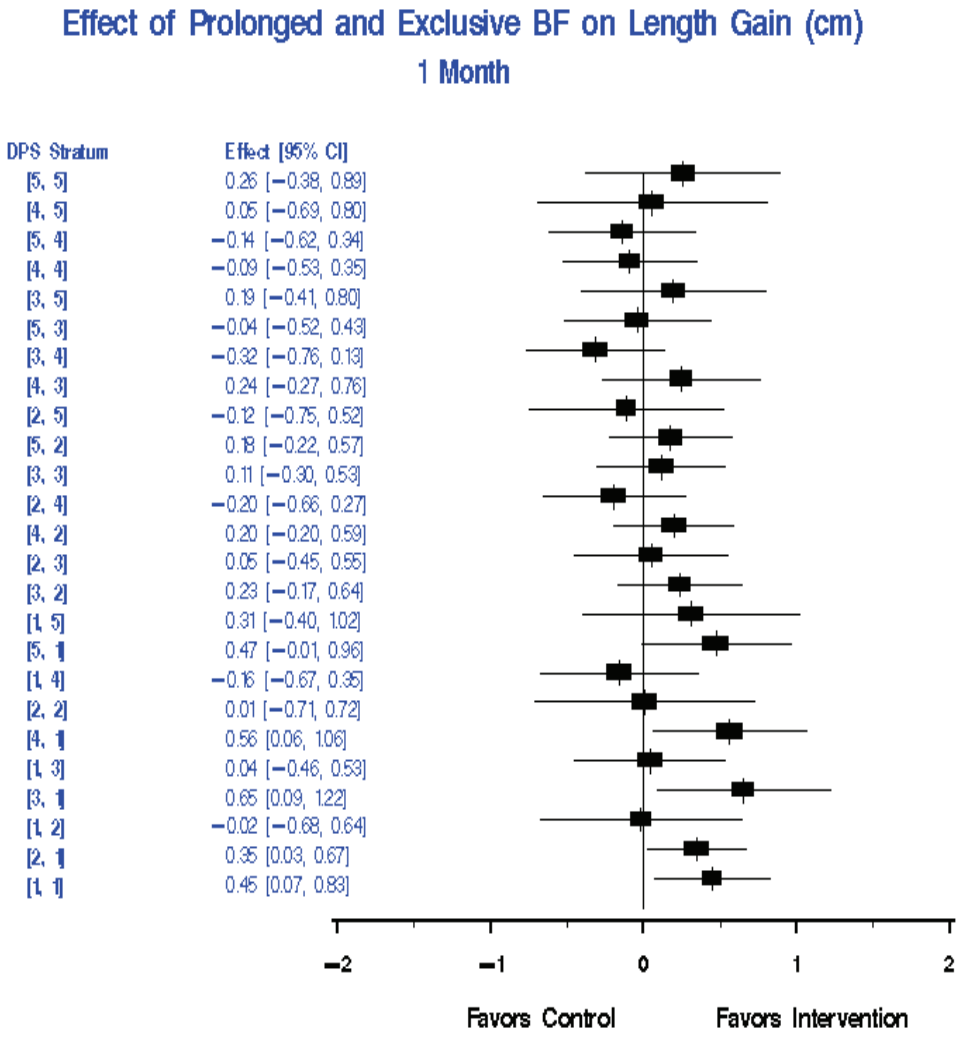


Figure A1-8 Effect of Prolonged and Exclusive BF on Length Gain (cm) at 2 Months

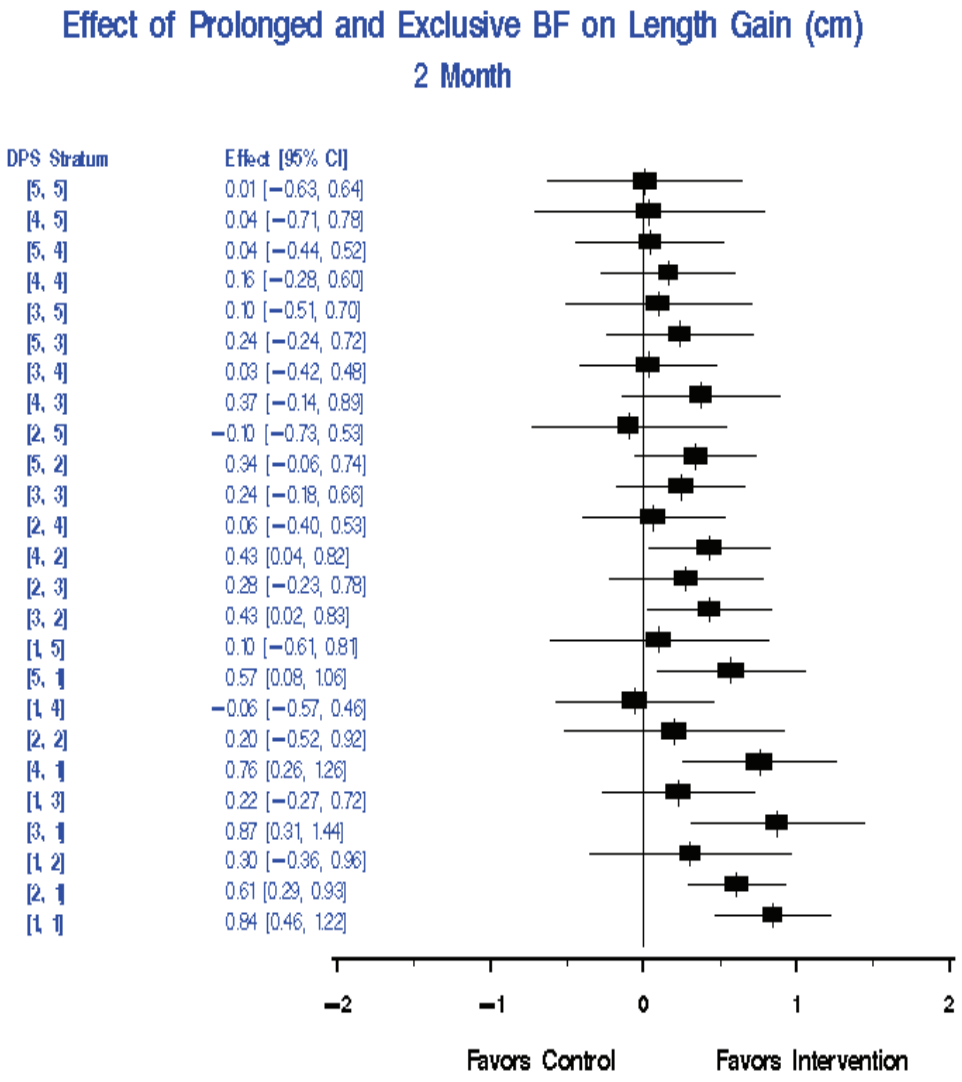


Figure A1-9 Effect of Prolonged and Exclusive BF on Length Gain (cm) at 3 Months

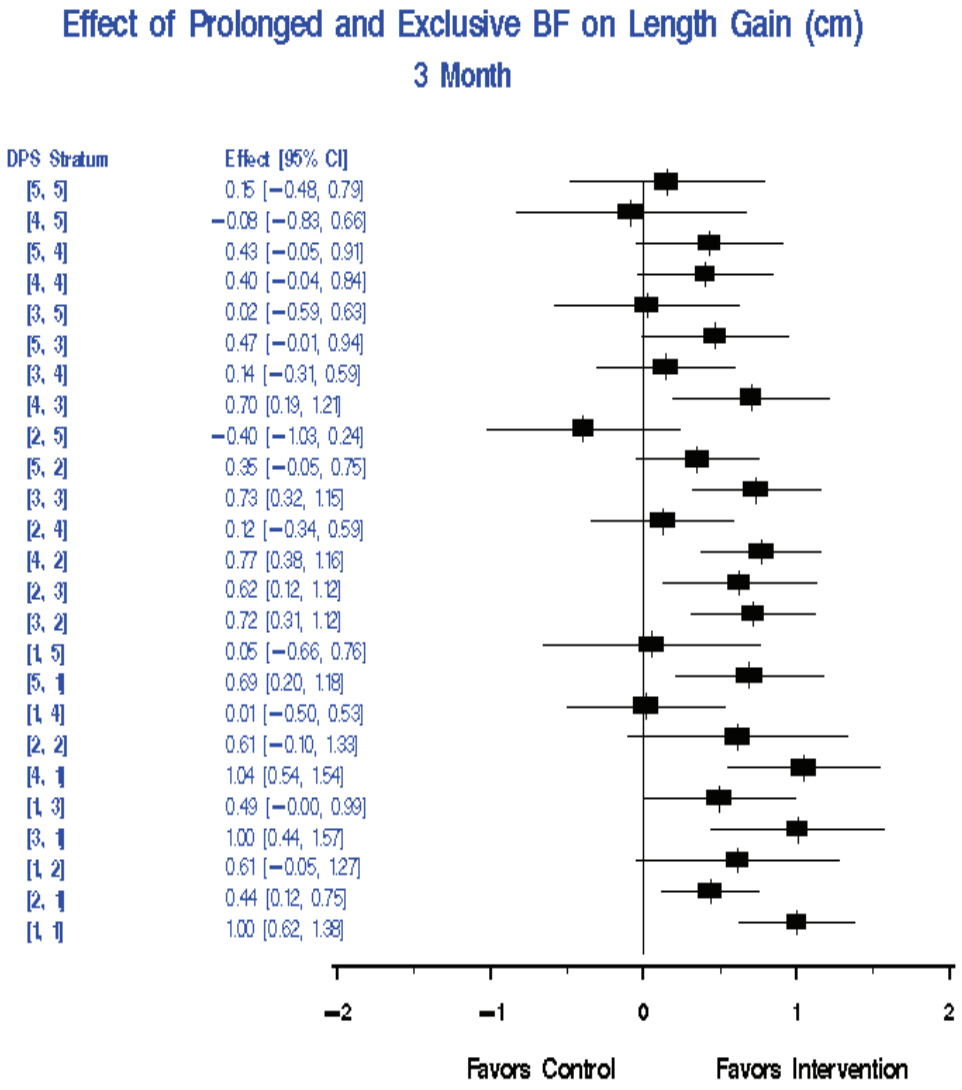


Figure A1-10 Effect of Prolonged and Exclusive BF on Length Gain (cm) at 6 Months

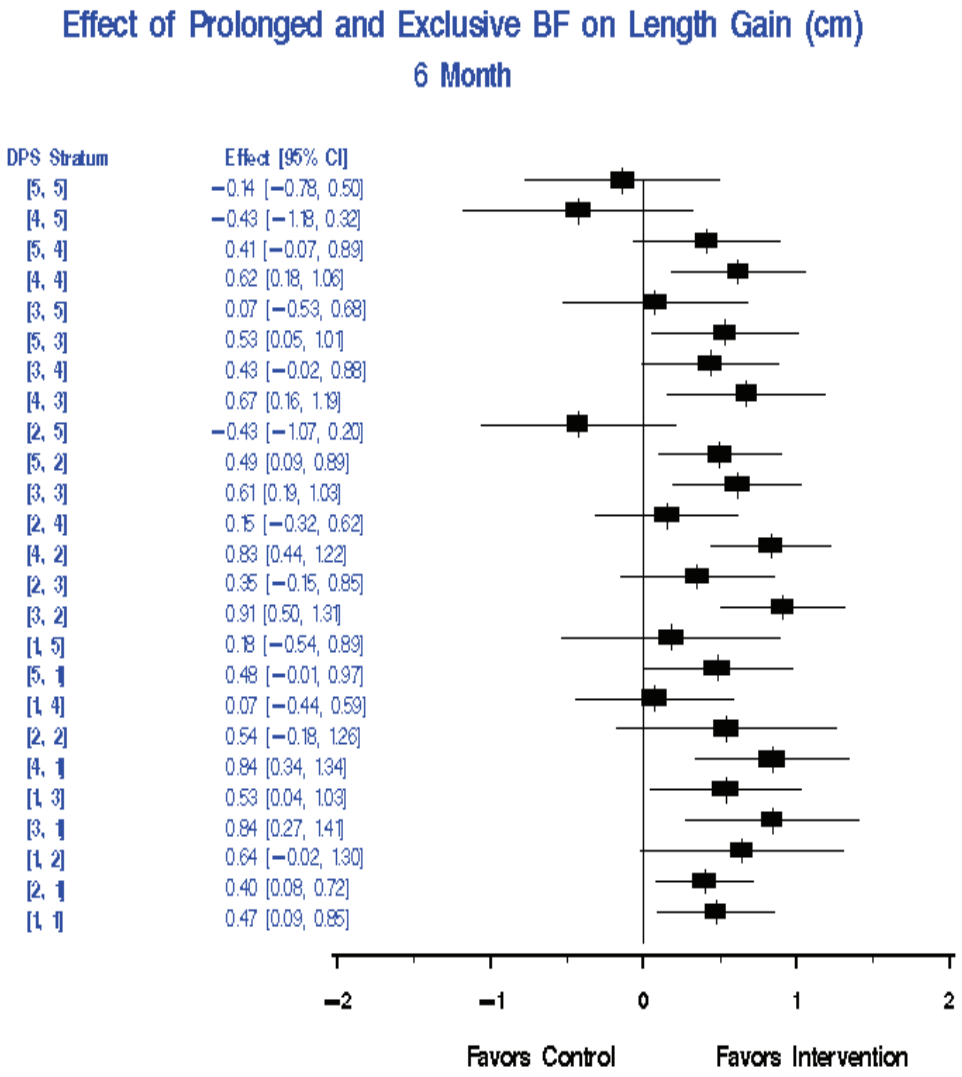


Figure A1-11 Effect of Prolonged and Exclusive BF on Length Gain (cm) at 9 Months

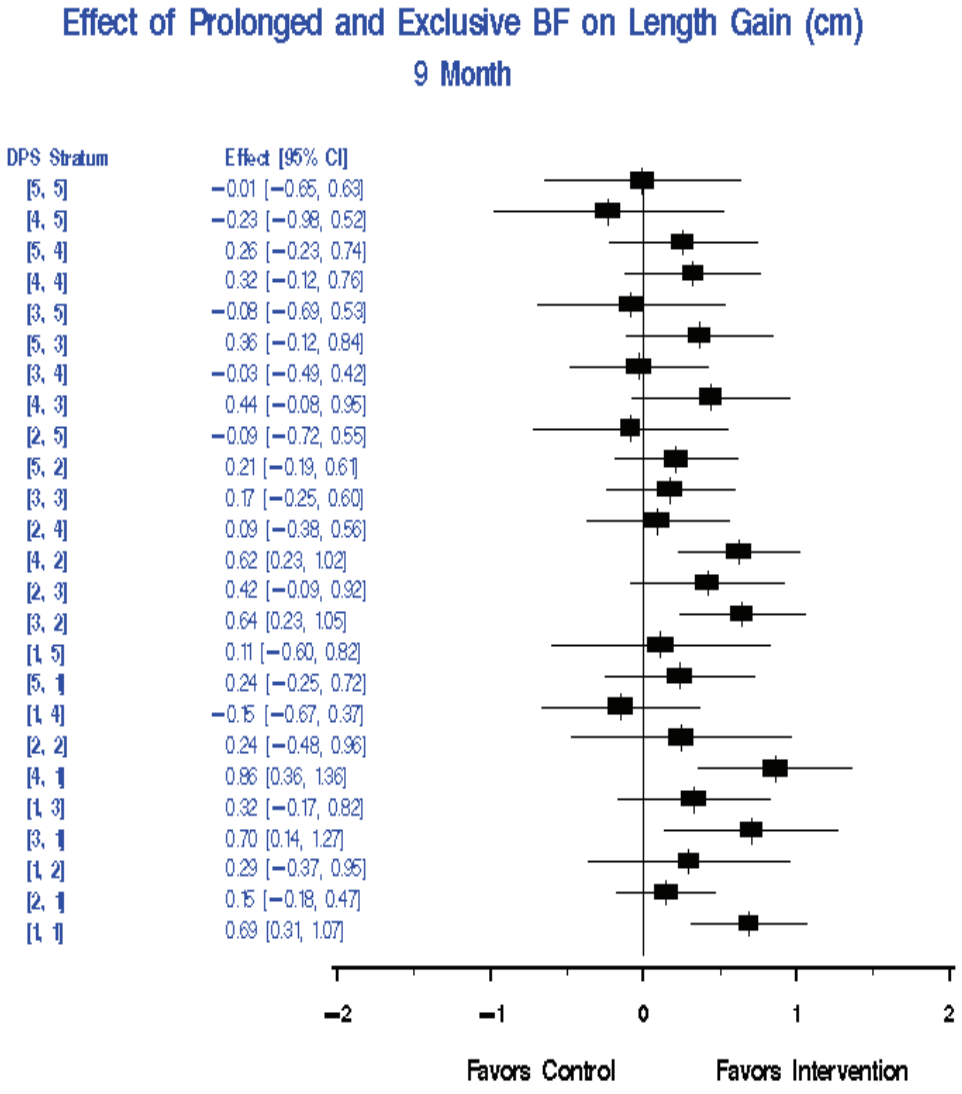


Figure A1-12 Effect of Prolonged and Exclusive BF on Length Gain (cm) at 12 Months