1-1-2014

Causal inference Methods for Addressing Censoring by Death and Unmeasured Confounding Using Instrumental Variables

Fan Yang
University of Pennsylvania, yangfan@sas.upenn.edu

Causal inference Methods for Addressing Censoring by Death and Unmeasured Confounding Using Instrumental Variables

Abstract
This thesis considers three problems in causal inference. First, for the censoring by death problem, we propose a set of ranked average score assumptions making use of survival information both before and after the measurement of a non-mortality outcome to tighten the bounds on the survivor average causal effect (SACE) obtained in the previous literature that utilized survival information only before the measurement. We apply our method to a randomized trial study of the effect of mechanical ventilation with lower tidal volume vs. traditional tidal volume for acute lung injury patients. Our bounds on the SACE are much shorter than the bounds obtained using only the survival information before the measurement of the non-mortality outcome. Second, for the IV method with nonignorable missing covariates problem, we develop a method to estimate the causal effect of a treatment in observational studies using an IV when there are nonignorable missing covariates, i.e., missingness depending on the partially observed compliance class besides the fully observed outcome, covariates and IV. We apply our method to a motivating study in neonatal care to study the effectiveness of high level compared to low level NICUs. Third, besides the association with the treatment, there are two key assumptions for the IV to be valid: (i) the IV is essentially random conditioning on observed covariates, (ii) the IV affects outcomes only by altering the treatment, the so-called ``exclusion restriction". These two assumptions are often said to be untestable; however, that is untrue if testable means checking the compatibility of assumptions with other things we think we know. A test of this sort may result in an aporia. We discuss this subject in the context of our on-going study of the effects of delivery by cesarean section on the survival of extremely premature infants of 23-24 weeks gestational age.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Applied Mathematics

First Advisor
Dylan Small

Second Advisor
Edward George

Keywords
Aporia, causal inference, censoring by death, nonignorable missing data

CAUSAL INFERENCE METHODS FOR ADDRESSING CENSORING BY DEATH AND UNMEASURED CONFOUNDING USING INSTRUMENTAL VARIABLES

Fan Yang

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Dylan Small
Professor
Department of Statistics

Co-Supervisor of Dissertation

_____

Edward George
Professor
Department of Statistics

Graduate Group Chairperson

_____

Charles Epstein
Professor, Department of Mathematics

Dissertation Committee:

Paul Rosenbaum, Professor, Department of Statistics

Acknowledgments

My deepest gratitude is to Professors Dylan Small and Edward George for their continuous support of my PhD study and research, for their excellent guidance, patience, motivation and providing me with an excellent atmosphere for doing research.

I am deeply grateful to Professor Paul Rosenbaum for many stimulating ideas and for valuable discussions. I would like to thank Professor Scott Lorch for supporting my research, giving his best suggestions and guidance to help develop my background in medicine. I would also like to thank Professor Jose Zubizarreta for many helpful discussions.

My sincere thanks also go to Professors Abba Krieger and Mark Low for their encouragement during my graduate studies and valuable comments during my job search. I would also like to thank Professor Charles Epstein for providing me this wonderful opportunity to study in this program and for his continuous support.

Many friends have helped me through these years. I greatly value their friendship.

Last but not the least, I want to thank my husband Xingtan and my parents, for their support, encouragement, and love.

ABSTRACT

CAUSAL INFERENCE METHODS FOR ADDRESSING CENSORING BY DEATH AND UNMEASURED CONFOUNDING USING INSTRUMENTAL VARIABLES

Fan Yang

Dylan Small, Edward George

This thesis considers three problems in causal inference. First, for the censoring by death problem, we propose a set of ranked average score assumptions making use of survival information both before and after the measurement of a non-mortality outcome to tighten the bounds on the survivor average causal effect (SACE) obtained in the previous literature that utilized survival information only before the measurement. We apply our method to a randomized trial study of the effect of mechanical ventilation with lower tidal volume vs. traditional tidal volume for acute lung injury patients. Our bounds on the SACE are much shorter than the bounds obtained using only the survival information before the measurement of the non-mortality outcome. Second, for the IV method with nonignorable missing covariates problem, we develop a method to estimate the causal effect of a treatment in observational studies using an IV when there are nonignorable missing covariates, i.e., missingness depending on the partially observed compliance class besides the fully observed outcome, covariates and IV. We apply our method to a motivat-

ing study in neonatal care to study the effectiveness of high level compared to low level NICUs. Third, besides the association with the treatment, there are two key assumptions for the IV to be valid: (i) the IV is essentially random conditioning on observed covariates, (ii) the IV affects outcomes only by altering the treatment, the so-called "exclusion restriction". These two assumptions are often said to be untestable; however, that is untrue if testable means checking the compatibility of assumptions with other things we think we know. A test of this sort may result in an aporia. We discuss this subject in the context of our on-going study of the effects of delivery by cesarean section on the survival of extremely premature infants of 23-24 weeks gestational age.

Contents

Chapter 1

Introduction

Causal inference is a central aim of many medical studies and social science studies where the goal is to identify the impact of a treatment or exposure on outcomes of interest, for instance, the effect of smoking on lung cancer. In this thesis, we focus on three problems that complicate the analysis of experiments and observational studies to draw causal conclusions, namely, the censoring by death problem, the IV method with nonignorable missing covariates and testing the validity of an IV.

Many clinical studies where the effect of treatment on a non-mortality outcome is of interest are complicated by censoring by death – for those patients who die before the measurement of non-mortality outcome, their non-mortality outcomes are not measured or well-defined. This is a special type of missing data. Even in randomized experiments with perfect compliance, a direct comparison of the non-mortality outcomes among the survivors in treatment vs. control could be biased

because censoring by death is typically informative since patients who die usually would have had bad non-mortality outcomes compared to those who did not die even if the dead patients could have somehow been kept alive. To address this problem, we focus on a well defined causal estimand-the survivor average causal effect (SACE) (Rubin, 2000; Frangakis and Rubin, 2002)- which is the effect of treatment on the non-mortality outcome among subjects who would survive under both treatment and control to the time point when the non-mortality outcome is measured. The SACE is not point identified without strong untestable assumption; however, with reasonable assumptions, it can be bounded (Zhang and Rubin, 2003; Imai, 2008; Chiba, 2012). In the previous literature on bounding the SACE, only the survival information before the measurement on the non-mortality outcome has been used. With limited information used, the bound is generally wide. In most clinical studies, the information is not restricted to the treatment the patient received, the outcome of interest and the survival information before the measurement of this outcome. In this thesis, based on the fact that the survival information after measurement of non-mortality outcome is also informative as a proxy of the severity of conditions of subjects, we proposed a set of ranked average score with two stage survival information assumptions which are plausibly satisfied in many quality of life studies and developed a two-step linear programming approach to obtain the closed form of the bound of the SACE under our assumptions. By utilizing both the survival information before and after the measurement of non-mortality outcome, inferences

on the SACE could be sharpened. Both numerical examples and an application to a critical care study illustrate the benefit of utilizing the further outcome information of survival. We also extend our method to bound the complier survivor average causal effect (CSACE) in randomized trials with noncompliance or observational studies where a valid instrumental variable is available.

The work on the IV method with nonignorable missing covariates is motivated by an observational study of neonatal care that aims to estimate the effect on mortality of premature babies being delivered in a high level NICU (neonatal intensive care units that have the capacity for sustained mechanical assisted ventilation and high volume) vs. a low level NICU. This study is complicated by unmeasured confounders as well as nonignorable missing covariates. To control for unmeasured confounders, we adopt the IV approach. An IV is a variable that is (i) associated with the treatment, (ii) has no direct effect on the outcome, and (iii) is independent of unmeasured confounders conditional on measured confounders. In the neonatal care study, we consider the use of excess travel time as an instrumental variable (IV) to control for unmeasured confounders. However, some confounders of the IV - outcome relationship we must condition on in order for our IV to be valid are not completely recorded and the missingness of those covariates may depend on only partially observed compliance status (i.e., whether the choice of treatment complies with the IV encouragement) besides the fully observed outcome, fully observed covariates and IV. Many observational studies face similar issues of unmeasured

confounding and nonignorable missing covariates, for example, in comparative effectiveness studies, it is a concern that the missingness of important lab values might be related with compliance status. Thus, developing an approach that can account for both issues is of real need. Previous literature on IV with nonignorable missing data focused on outcomes (Frangakis and Rubin, 1999; Mealli et al., 2004; Peng et al.(2004); Chen et al., 2009; Small and Cheng, 2009). In this literature, it has been argued that ignorability of the missing outcome may only be plausible after conditioning on the covariates *and* the partially observed compliance status. Methods have been developed for estimating causal effects under this "latent ignorability". However, no literature we are aware of addressed the problem of IV with nonignorable missing covariates. In this thesis, a method is developed to address the above issues. We proposed a series of models to estimate the causal effect of a treatment when a valid IV is available under our nonignorable missingness assumption which assumes that the missingness of covariates is ignorable conditional on the fully observed outcome, the fully observed covariates, the IV as well as the partially observed compliance behavior. Simulation studies indicate that when the missingness of covariates is related to the partially observed compliance behavior, even if the missing rate of covariates is low and the effect of compliance status on the missingness is only moderate, the commonly used estimation methods, complete case analysis and multiple imputation by chained equations assuming missingness at random, provide substantially biased estimates, while our method, which is de-

signed to deal with nonignorable missingness of covariates, provide approximately unbiased estimates. By extending the proposed series of models to allow for an unmeasured confounder's effects on both outcome and missingness, one can assess the sensitivity of the causal conclusions to a deviation from our nonignorable missingness assumption.

The IV method is widely used in observational studies (Angrist and Krueger, 1991; Baiocchi et al., 2010; Yang, Lorch and Small, 2014). Besides the association with the treatment, there are two key assumptions for the IV to be valid: (i) the IV is haphazard or essentially random once adjustments have been made for observed covariates, (ii) the IV affects outcomes only by altering the treatment, the so-called "exclusion restriction". These two assumptions are often said to be untestable (e.g., Morgan and Winship 2007, p.196). Our point of view is that if one confined attention to the information in the target study itself - namely, the confounders measured, the IV, treatment and outcome - then perhaps there is no way to test IV assumptions, however, it is often possible to check the IV assumptions against other things we think we know. In this thesis, we suggest that a test of IV assumptions may lead neither to rejection of the assumptions nor to acceptance but rather to an aporia. An aporia is a collection of propositions such that each one is plausible on its own but they are jointly inconsistent. We discuss this subject in the context of an ongoing study I am collaborating on about the effect of delivery by cesarean section on the survival of extremely premature infants of 23-24 weeks gestational

age. We proposed as an IV the cesarean section rate at the hospital at which the baby was delivered; similar IVs of how often a procedure is performed at a hospital or in a geographic region have been used in many health care studies (Brookhart and Schneeweiss, 2007) We applied to the data a new bipartite matching algorithm to strengthen this instrumental variable. Under the assumption that the IV is valid, we found strong evidence that cesarean sections increase the survival of premature infants of 23-24 weeks gestational age. To test the validity of our IV, we used the fact that the literature claims that for older preterm babies of 30-34 weeks gestational age, there is no benefit from cesarean section (Werner et al., 2013; Malloy, 2009). We used the same IV and applied the same matching procedure on older preterm babies, say 30-34 weeks gestational age. There are two possible results: (i) there is no evidence of the benefit from cesarean section for infants of 30 - 34 weeks gestational age using the IV proposed which agrees with the literature, then we are more confident about the validity of our IV thus more comfortable with the causal conclusion we obtained for the infants of 23-24 weeks gestational age; (ii) there is evidence from the IV analysis that cesarean sections benefit 30-34 week old babies, which contradicts to the current literature. In our analysis of the data, we obtained result (ii). This creates an aporia – the assumption that the IV is valid and the literature that says that cesarean sections don't benefit 30-34 week old babies cannot both be right. This is an advance in understanding, albeit recognizing an aporia is an uncomfortable one, but it is certainly better than believing each claims

without recognizing their aporetic status.

The rest of this thesis is organized as follows.

In Chapter 2, we discuss the problem of censoring by death. Section 2.1 provides a detailed introduction to this problem. In section 2.2, we introduce notation and assumptions to set up the causal framework. In section 2.3, we present the derivations of the bounds of SACE and provide some numerical examples to compare the bounds derived with the bound using one set of assumptions in Zhang and Rubin (2003). We extend our method to IV settings in section 2.4. In section 2.5, we discuss how to check the plausibility of our assumptions for the "large sample" data as well as the sample data. We discuss the confidence intervals for bounds in section 2.6, and we apply our approach to the tidal volume study in section 2.7. Conclusions and discussions are presented in section 2.8.

In Chapter 3, we discuss the IV method with nonignorable missing covariates. Section 3.1 provides a detailed introduction to this problem. In section 3.2, we introduce notation and assumptions to set up the causal framework. In section 3.3, we present our model for inference about complier average causal effect (CACE) and EM algorithm to estimate parameters involved in the model. A simulation study is be provided in section 3.4, and we study how the estimates of CACE can be affected by wrong assumptions about missing mechanisms. Then, we apply our approach to the neonatal care data in section 3.5, and conduct the sensitivity analysis in section 3.6. Conclusions are presented in section 3.7.

In Chapter 4, we discuss the aporetic conclusions when testing the validity of an IV. Section 4.1 provides an introduction to this problem. In section 4.2, we discuss some backgroud, present the IV analysis, test the IV assumptions and discuss in detail resulting in an aporia. A summary is presented in section 4.3.

Chapter 2

Two-stage Censoring by Death

2.1 Introduction

In many clinical studies, researchers are interested in the effect of a treatment on a non-mortality outcome such as complications or quality of life in addition to mortality. However, the assessment of the causal effect on non-mortality outcomes of interest is often complicated by censoring by death. This censoring by death occurs because, by the time the non-mortality outcome is measured, some patients have died and thus the non-mortality outcome cannot be measured or is not well defined for these dead patients. For example, suppose we want to study the effect on intraventricular hemorrhage (IVH) of premature babies being delivered in a high-level neonatal intensive care unit (NICU) vs. a lower-level NICU. IVH is rarely present at birth but usually occurs in the first several days of life (See Lee, 2013).

If the baby died before being born (a fetal death) or shortly after birth, then whether the baby had IVH is not well-defined. Another example is that in cancer studies, quality of life outcomes that might be measured six months or a year after treatment like incidence of fatigue, myelosuppression and treatment side-effects (e.g., Motzer et al., 2013) are important outcomes considered to assess the efficacy of a treatment. However, patients may die before the measurement of the quality of life outcomes; for those patients, the quality of life outcomes are not well-defined. Censoring by death is typically informative – patients who die usually would have had worse quality of life than those who did not die even if the dead patients could have somehow been kept alive (Cox et al., 1992). Furthermore, those patients who are saved by a treatment are often sicker patients on average than those patients who would live under both treatment and control. Consequently, a direct comparison of the non-mortality outcomes among the survivors in treatment vs. control would be biased. To address the fundamental problems that the non-mortality outcomes are not well defined for those who die before measurement and that the censoring of the measurement is informative, Rubin (2000), and Frangakis and Rubin (2002) proposed a well defined causal estimand – the survivor average causal effect (SACE) – which is the effect of treatment on the non-mortality outcome among patients who would survive under both treatment and control to the time point when the non-mortality outcome is measured.

Without strong untestable assumptions, the SACE is not point identified; how-

ever, with reasonable assumptions, we can obtain an interval in which SACE will lie. Zhang and Rubin (2003) discussed various assumptions (ranked average score assumptions) that can be made to bound the SACE, and derive large sample bounds in a randomized trial. Imai (2008) provided an alternative proof that the bounds obtained in Zhang and Rubin (2003) are sharp and generalized the proof to obtain sharp bounds on the quantile treatment effect. Chiba (2012) proposed a number of assumptions that are different from the ranked average score assumptions in Zhang and Rubin (2003) and derived the corresponding bounds. Another stream of work on drawing inference about the SACE is through sensitivity analysis procedures, for instance, Hayden et al. (2005), Egleston et al. (2007), and Chiba and VanderWeele (2011). A problem similar to censoring by death arises in evaluating the effect of vaccine vs. placebo on post-infection outcomes. Hudgens, Hoering and Self (2003) developed tests for the causal effect on viral load among the individuals who would be infected no matter whether they received the vaccine regimen or a placebo regimen. Gilbert, Bosch and Hudgens (2003) proposed a class of models indexed by an interpretable sensitivity parameter, where the SACE is identified given the sensitivity parameter.

In the previous literature on bounding the SACE, only the survival information before the measurement on the non-mortality outcome has been used. However, survival information after measurement may be informative. In this chapter, we develop a method to use both the survival information before and after the mea-

surement of non-mortality to sharpen inferences on the SACE in the setup of randomized experiments. We will also present an extension of our method to bound the complier survivor average causal effect (CSACE) in a randomized trial with noncompliance or an observational study where an instrumental variable (IV) is available.

We will apply our method to the ARDSNet study, a randomized clinical trial on the effect of mechanical ventilation with lower tidal volumes vs. traditional tidal volumes for patients suffering from acute lung injury (The Acute Respiratory Distress Syndrome Network, 2000). The trial found evidence that lower tidal volumes reduce mortality. The investigators were also interested in assessing the effect of lower tidal volumes on a quality of life (QOL) outcome, whether the patient was able to breathe without assistance by day 28. In the data, both survival at day 28, when the QOL is measured, and whether the patient was ultimately discharged home alive, post-QOL measurement survival information, are recorded. Utilizing the post QOL measurement survival information in addition to the pre-QOL measurement survival information, we are able to substantially sharpen the bounds on the SACE for the effect of lower tidal volume on being able to breathe without assistance by day 28.

2.2 Notation and Assumptions: Randomized Experiment with Perfect Compliance

In this section and the following, we focus on two arm randomized experiments where the subjects are randomly assigned to either treatment or control. The method is extended to IV settings in section 2.4.

2.2.1 Notation

We use the potential outcomes approach to define causal effects. Let D_i represent the binary treatment for the i^{th} subject; we call level 1 "the treatment" and level 0 "the control". Let $\boldsymbol{D}$ denote the vector of treatment assignment indicators for all subjects. Let $S_{1i}(\boldsymbol{d})$ be the potential survival indicator of subject i that would be observed at the first time point after which the measurement of non-mortality outcome is taken, with 0 indicating death, 1 if alive. Let $Y_i(\boldsymbol{d})$ represent the potential non-mortality binary outcome (for instance, complication of babies, QOL of participants) that would be observed under treatment assignment $\boldsymbol{d}$. The non-mortality outcome is measured after the first time point, thus if the subject would die before that time point ($S_{1i}(\boldsymbol{d}) = 0$), $Y_i(\boldsymbol{d})$ is not defined. For convenience, we assume that level 1 of the non-mortality outcome is worse than level 0 of the outcome, e.g., in the ARDSNet study, level 1 indicates that the patient was not able to breathe without assistance by day 28 and level 0 indicates the patient was able to

breathe without assistance by day 28. We further define $S_{2i}(\boldsymbol{d})$ to be the potential indicator of survival at the second time point for subject i that would be observed if under treatment assignment $\boldsymbol{d}$. If $S_{1i}(\boldsymbol{d}) = 0$, then $S_{2i}(\boldsymbol{d}) = 0$ by definition. We use D_i, S_{1i}, Y_i and S_{2i} to denote respectively the observed treatment received, observed survival indicator at the first time point, observed non-mortality outcome and observed survival indicator at the second time point for subject i.

2.2.2 Assumptions

We assume that the following assumptions hold for randomized experiments.

Assumption 1. Stable unit treatment value assumption (SUVTA).

- If $d_i = d_i'$, then $S_{1i}(\boldsymbol{d}) = S_{1i}(\boldsymbol{d'})$, $S_{2i}(\boldsymbol{d}) = S_{2i}(\boldsymbol{d'})$, and $Y_i(\boldsymbol{d}) = Y_i(\boldsymbol{d'})$

SUVTA means that there is no interference between subjects so that a subject's outcome only depends on the subject's own treatment. Under SUVTA, each subject has two potential first time point survival outcomes $(S_{1i}(1), S_{1i}(0))$, based on values of which we can classify subjects into four groups:

- $11 = \{i|S_{1i}(1) = 1, S_{1i}(0) = 1\}$, always survivors: the subjects that would survive at least to the first time point under both treatment arms,

- $10 = \{i|S_{1i}(1) = 1, S_{1i}(0) = 0\}$, protected: the subjects that would survive at least to the first time point under treatment, but would die before then under control;

- $01 = \{i|S_{1i}(1) = 0, S_{1i}(0) = 1\}$, harmed: the subjects that would die before the first time point under treatment, but would survive at least to the first time point under control;

- $00 = \{i|S_{1i}(1) = 0, S_{1i}(0) = 0\}$, never survivors: the subjects that would die before the first time point under both treatment arms;

Assumption 2. The assignment D_i of each subject is independent of his/her potential outcomes.

Assumption 3. Monotonicity: $S_{1i}(1) \geq S_{1i}(0), S_{2i}(1) \geq S_{2i}(0)$. There is no 01 (harmed) group.

The monotonicity assumption says that the treatment does not cause death, which is often plausible in practice. Under this assumption, subjects could either be "always survivors", "protected" or "never survivors". The most meaningful inference of causal effect of treatment on Y can be drawn only for the "always survivors", because it is the only group for which both $Y_i(1)$ and $Y_i(0)$ are well defined, see Rubin (2000), Frangakis and Rubin (2002). Define the survivor average causal effect (SACE) as $\mathrm{E}(Y_i(1) - Y_i(0) \mid 11)$, which is our quantity of interest.

We further create finer strata based on the possible combinations of potential survival at both the first (QOL measurement point) and second (post-QOL measurement point) time points, which is described in Table 2.1.

The always survivors at time point 1 are divided into the following three subgroups: 1111, always survivors who would live at least to the second time point

Table 2.1: Fine Strata

Probability	$S_{1i}(1)$	$S_{1i}(0)$	$S_{2i}(1)$	$S_{2i}(0)$	Principal Strata at Time Point 1
π_{1111}	1	1	1	1	Always survivors
π_{1110}	1	1	1	0	Always survivors
π_{1100}	1	1	0	0	Always survivors
π_{1010}	1	0	1	0	Protected
π_{1000}	1	0	0	0	Protected
π_{0000}	0	0	0	0	Never survivors

under both treatment arms; 1110, always survivors who would survive at least to the second time point under treatment, but would die before then under control; 1100, always survivors who although they can live at least to the first time point, would die before the second time point under both treatment arms. The protected at time point 1 are combinations of the following two subgroups: 1010, subjects who would live at least to the second time point under treatment, but would die before the first time point under control; 1000, subjects who if they receive treatment would live at least to the first time point but would die before the second time point, but if they receive control, would die even before the first time point. Never survivors comprise a single subgroup which we denote as 0000 because the second time point death indicator provides no additional information for them.

In terms of our fine strata, the SACE is expressed as:

$$
\begin{aligned}
SACE &= \mathrm{E}(Y_i(1) - Y_i(0) \mid S_{1i} = S_{1i}(0) = 1) \\
&= P(Y_i(1) = 1 \mid S_{1i}(1) = S_{1i}(0) = 1) - P(Y_i(0) = 1 \mid S_{1i}(1) = S_{1i}(0) = 1) \\
&= \frac{(\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100))}{\pi_{1111} + \pi_{1110} + \pi_{1100}} \\
&- \frac{(\pi_{1111}\mathrm{E}(Y_i(0) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(0) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(0) \mid 1100))}{\pi_{1111} + \pi_{1110} + \pi_{1100}}
\end{aligned} \tag{2.2.1}
$$

Plausible assumptions can be made on data to tighten the bounds of SACE. Zhang and Rubin (2003) proposed the assumption that when assigned treatment, on average, the outcome for "always survivors" is better than "protected", in our case, that is to say $P(Y_i(1) = 1 \mid 11) \leq P(Y_i(1) = 1 \mid 10)$, recalling that we use 1 to denote worse outcome for Y. This uses only the information on death before the measurement of the non-mortality outcome. In the rest of this chapter, we will refer to this assumption as the ranked average score with one stage survival information assumption. Survival information after measurement of the non-mortality outcome may deliver finer information, making use of which can help us make more reasonable assumptions and sharpen inferences. We will refer to the following set of assumptions as ranked average score with two stage survival information assumptions.

Assumption 4. Among always survivors at time point 1, the probability of worse outcome for group 1111 is the lowest, whereas the probability of worse outcome for group 1100 is the highest under both treatment arms:

$$
P(Y_i(1) = 1 \mid 1111) \leq P(Y_i(1) = 1 \mid 1110) \leq P(Y_i(1) = 1 \mid 1100) \tag{2.2.2}
$$

$$P(Y_i(0) = 1 \mid 1111) \leq P(Y_i(0) = 1 \mid 1110) \leq P(Y_i(0) = 1 \mid 1100) \tag{2.2.3}$$

Assumption 5. Among protected at time point 1, the probability of worse outcome for group 1010 is no higher than that for group 1000 under treatment:

$$P(Y_i(1) = 1 \mid 1010) \leq P(Y_i(1) = 1 \mid 1000) \tag{2.2.4}$$

Assumption 6. Under treatment, the probability of worse outcome for group 1100 is not lower than that for group 1010, but not higher than that for group 1000, and the probability of worse outcome for group 1110 is not higher than that for group 1010:

$$P(Y_i(1) = 1 \mid 1110) \leq P(Y_i(1) = 1 \mid 1010) \leq P(Y_i(1) = 1 \mid 1100) \leq P(Y_i(1) = 1 \mid 1000) \tag{2.2.5}$$

Assumptions 4, 5 and 6 are plausibly satisfied in many QOL studies. Consider the ARDSNet study of the effect of lower tidal volumes (treatment) vs. traditional tidal volumes (control) on being able to breathe without assistance by day 28 in the ICU described in the introduction, where the post-QOL measurement survival time point is being discharged home alive. Assumption 4 says, among patients who would survive to day 28 under both treatment and control, those patients who would be discharged home alive under both treatment and control are healthiest at day 28 on average, and those who would be discharged home alive under treatment but not control are healthier at day 28, than those who would die in the hospital under both treatment and control. Assumption 5 says, among patients who would survive to day 28 only under treatment, those patients who would ultimately be

discharged home alive under treatment are healthier on average than patients who would ultimately die in the hospital. Assumptions 4 and 5 are plausible because being discharged home alive is a proxy for health at day 28. Assumption 6 is a comparison of the 1010 patients who would die before day 28 under control but survive to day 28 and be discharged home alive under treatment, to the 1100 patients who would survive to day 28 under both treatment and control but die in the hospital after day 28 under both treatment and control. Assumption 6 says that under the treatment, the 1010 patients tend to be healthier than the 1100 patients at day 28. This is plausible for the ARDSNet study for the following reasons. The 1100 patients are likely to be fairly sick by day 28 under the treatment since these patients will die in the ICU. In contrast, the 1010 patients are likely to be less sick on day 28 under the treatment because they will be (or already have) discharged home alive. An example of a 1010 patient would be a patient who was healthy but suffered a gunshot wound that caused an acute lung injury. When the patient arrives at the ICU, the patient is in critical condition and only the treatment will save the patient, but if the patient receives the treatment, the patient's health before the gunshot wound will enable the patient to recover well and be regaining his or her health by day 28. In summary, assumptions 4, 5 and 6 are plausible for the ARDSNet study.

The ranked average score with one stage survival information assumption is different from our ranked average score with two stage survival information as-

sumptions. The major difference is that the one-stage survival assumption assumes that always survivors, on average, have better QOL outcome than the protected, whereas our two-stage survival assumptions assume that one particular always survivors group, 1100, has worse QOL outcome than a particular protected group, 1010, on average under treatment, which is a more reasonable assumption for the ARDSNet study. The differences in the bounds obtained under the ranked average score with one stage survival information assumption and our two stage survival information assumptions are presented in numerical examples and the analysis of ARDSNet study in section 2.3.3 and 2.7 respectively.

2.3 Derivations of Bounds

Under assumptions 1-6, the SACE is not point identified based on the knowledge of the observable joint distribution of $(D_i, S_{1i}, S_{2i}, Y_i)$. However, we can use that joint distribution to obtain an interval in which the SACE must lie. We first derive the bounds for the proportions in each stratum, then for fixed proportions we derive the bounds for the SACE. In this section, we assume that the joint distribution of $(D_i, S_{1i}, S_{2i}, Y_i)$ is known; in section 2.6, we will discuss forming confidence intervals for the bounds that acknowledge sample uncertainty.

2.3.1 Bounds for proportions of each stratum

Notice that the observable strata of (D_i, S_{1i}, S_{2i}) are mixtures of fine strata (Table 1). Thus we can express the proportions of strata of (D_i, S_{1i}, S_{2i}) by proportions of fine strata. Combining this with the fact that all the proportions in the fine strata must lie between 0 and 1, we can obtain the bounds for each fine stratum's proportion. We use $p_{s_1 s_2|d}$ to denote $P(S_{1i} = s_1, S_{2i} = s_2 \mid D_i = d)$. The following equations hold:

$$p_{11|1} = \pi_{1111} + \pi_{1110} + \pi_{1010} \tag{2.3.1}$$

$$p_{10|1} = \pi_{1100} + \pi_{1000} \tag{2.3.2}$$

$$p_{00|1} = \pi_{0000} \tag{2.3.3}$$

$$p_{11|0} = \pi_{1111} \tag{2.3.4}$$

$$p_{10|0} = \pi_{1110} + \pi_{1100} \tag{2.3.5}$$

$$p_{00|0} = \pi_{1010} + \pi_{1000} + \pi_{0000} \tag{2.3.6}$$

Further we have,

$$0 \leq \pi_{1111}, \pi_{1110}, \pi_{1100}, \pi_{1010}, \pi_{1100}, \pi_{1000}, \pi_{0000} \leq 1 \tag{2.3.7}$$

Given (2.3.1)-(2.3.6), we can express each π by functions of $p_{ss|d}$ and π_{1100}:

$$\pi_{1111} = p_{11|0}$$

$$\pi_{1110} = p_{10|0} - \pi_{1100}$$

$$\pi_{1010} = p_{11|1} - p_{11|0} - p_{10|0} + \pi_{1100}$$

$$\pi_{1000} = p_{10|1} - \pi_{1100}$$

$$\pi_{0000} = p_{00|1}$$

and subject to the constraint of (2.3.7), we have,

$$max\{0, p_{11|0} + p_{10|0} - p_{11|1}\} \leq \pi_{1100} \leq min\{p_{10|0}, p_{10|1}\} \tag{2.3.8}$$

2.3.2 Bounds for the SACE

In this step, we first derive the bounds for the SACE with known proportions of each fine stratum, then will combine the result with the bounds obtained in section 2.3.1 to construct the final bounds for the SACE.

The observable strata of $(Y_i, S_{1i}, S_{2i} \mid D_i)$ are mixtures of potential outcomes from the fine strata. Letting $q_{ys_1s_1|d}$ denote $P(Y_i = y, S_{1i} = s_1, S_{2i} = s_2 \mid D_i = d)$, we have the following identities:

$$q_{111|1} = \pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1010}\mathrm{E}(Y_i(1) \mid 1010) \tag{2.3.9}$$

$$q_{110|1} = \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100) + \pi_{1000}\mathrm{E}(Y_i(1) \mid 1000) \tag{2.3.10}$$

$$q_{111|0} = \pi_{1111}\mathrm{E}(Y_i(0) \mid 1111) \tag{2.3.11}$$

$$q_{110|0} = \pi_{1110}\mathrm{E}(Y_i(0) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(0) \mid 1100) \tag{2.3.12}$$

Recall that

$$SACE = \frac{(\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100))}{\pi_{1111} + \pi_{1110} + \pi_{1100}} - \frac{(\pi_{1111}\mathrm{E}(Y_i(0) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(0) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(0) \mid 1100))}{\pi_{1111} + \pi_{1110} + \pi_{1100}} \tag{2.3.13}$$

Given $\pi's$, $\frac{(\pi_{1111}\mathrm{E}(Y_i(0)|1111)+\pi_{1110}\mathrm{E}(Y_i(0)|1110)+\pi_{1100}\mathrm{E}(Y_i(0)|1100))}{\pi_{1111}+\pi_{1110}+\pi_{1100}} = \frac{q_{111|0}+q_{110|0}}{\pi_{1111}+\pi_{1110}+\pi_{1100}}$ which is point identified. Thus to bound the SACE, we only need to bound $\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)$, which defines a linear programming problem:

$$\min / \max \qquad (\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100} \tag{2.3.14}$$

Subject to:

$$q_{111|1} = \pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1010}\mathrm{E}(Y_i(1) \mid 1010) \tag{2.3.15}$$

$$q_{110|1} = \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100) + \pi_{1000}\mathrm{E}(Y_i(1) \mid 1000) \tag{2.3.16}$$

$$\mathrm{E}(Y_i(1) \mid 1111) \leq \mathrm{E}(Y_i(1) \mid 1110) \leq \mathrm{E}(Y_i(1) \mid 1100) \tag{2.3.17}$$

$$\mathrm{E}(Y_i(1) \mid 1010) \leq \mathrm{E}(Y_i(1) \mid 1000) \tag{2.3.18}$$

$$\mathrm{E}(Y_i(1) \mid 1110) \leq \mathrm{E}(Y_i(1) \mid 1010) \leq \mathrm{E}(Y_i(1) \mid 1100) \leq \mathrm{E}(Y_i(1) \mid 1000) \tag{2.3.19}$$

$$0 \leq \mathrm{E}(Y_i(1) \mid 1111), \mathrm{E}(Y_i(1) \mid 1110), \mathrm{E}(Y_i(1) \mid 1100), \mathrm{E}(Y_i(1) \mid 1010), \mathrm{E}(Y_i(1) \mid 1000) \leq 1 \tag{2.3.20}$$

where constraints (2.3.17)-(2.3.19) are imposed by assumptions 4-6.

The above linear programming problem has a solution if and only if $\frac{q_{110|1}}{p_{10|1}} \geq \frac{q_{111|1}}{p_{11|1}}$, which is an inequality that must be satisfied based on assumptions 4-6. For each possible value of π_{1100}, we solve the above linear programming problem; then, combining this result with the bound for π_{1100} derived in section 2.3.1 that $\pi_{1100} \in I$, where $I = [max\{0, p_{11|0} + p_{10|0} - p_{10|1}\}, \quad min\{p_{10|0}, p_{10|1}\}]$, we have,

$$\begin{aligned}\min SACE &= \min_{\pi_{1100}\in I}[\frac{\min((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100}) - (q_{111|0} + q_{110|0})}{\pi_{1111} + \pi_{1110} + \pi_{1100}}] \\ &= \begin{cases} max\{\frac{q_{111|1}+q_{110|1}-p_{11|1}-p_{10|1}+p_{11|0}+p_{10|0}}{p_{11|0}+p_{10|0}}, \frac{q_{111|1}}{p_{11|1}}\} - \frac{q_{111|0}+q_{110|0}}{p_{11|0}+p_{10|0}}, & \text{if } p_{11|0} + p_{10|0} - p_{11|1} \geq 0 \\ max\{0, \frac{q_{111|1}p_{10|1}+q_{110|1}(p_{11|0}+p_{10|0}-p_{11|1})}{p_{10|1}(p_{11|0}+p_{10|0})}\} - \frac{q_{111|0}+q_{110|0}}{p_{11|0}+p_{10|0}}, & \text{if } p_{11|0} + p_{10|0} - p_{11|1} < 0 \end{cases}\end{aligned} \tag{2.3.21}$$

$$\begin{aligned}\max SACE &= \max_{\pi_{1100}\in I}[\frac{\max((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100}) - (q_{111|0} + q_{110|0})}{\pi_{1111} + \pi_{1110} + \pi_{1100}}] \\ &= \frac{q_{111|1}}{p_{11|1}} - \frac{q_{111|0} + q_{110|0}}{p_{11|0} + p_{10|0}} + \frac{q_{110|1}p_{11|1} - q_{111|1}p_{10|1}}{p_{10|1}p_{11|1}(p_{11|0} + p_{10|0})} \cdot min\{p_{10|0}, p_{10|1}\}\end{aligned} \tag{2.3.22}$$

The details of the calculation for the bounds of SACE are provided in the Appendix.

2.3.3 Numerical Examples

Example 1

Assume that the underlying truth about the population is described by Table 2.2. The SACE = 0.05, meaning that the treatment will increase the probability of the worse non-mortality outcome by 0.05 among always survivors who will survive at least to the first time point under both treatment and control.

Suppose that we have an infinite sample, then we would observe that

$$p_{11|1} = 0.65 \quad p_{10|1} = 0.2 \quad p_{00|1} = 0.15 \quad p_{11|0} = 0.5 \quad p_{10|0} = 0.15 \quad p_{00|0} = 0.35 \tag{2.3.23}$$

Table 2.2: Setup 1

% of population	Fine Strata	% of $Y_i(1) = 1$	% of $Y_i(0) = 1$
50	1111	10	5
10	1110	20	15
5	1100	40	35
5	1010	30	-
15	1000	50	-
15	0000	-	-

$$q_{111|1} = 0.085 \quad q_{110|1} = 0.095 \quad q_{111|0} = 0.025 \quad q_{110|0} = 0.0325 \tag{2.3.24}$$

Given the constraints imposed by the observed data (2.3.23)-(2.3.24) and assumptions 4-6, we obtain the bound for SACE: [0.042, 0.122], showing that the treatment increases the probability of the worse non-mortality outcome.

However, if we don't use the second time point survival information, the observed data would be:

$$P(S_{1i} = 1|D_i = 1) = 0.85 \quad P(S_{1i} = 1|D_i = 0) = 0.65 \tag{2.3.25}$$

$$P(Y_i = 1, S_{1i} = 1|D_i = 1) = 0.18 \quad P(Y_i = 1, S_{1i} = 1|D_i = 0) = 0.0575 \tag{2.3.26}$$

Then, given the constraints imposed by the observed data (2.3.25)-(2.3.26) and the ranked average score with one stage survival assumption, the bound we would obtain for the SACE is [-0.088, 0.123], according to which we wouldn't know whether or not the treatment increases the probability of the worse non-mortality outcome even

Table 2.3: Setup 2

% of population	Fine Strata	% of $Y_i(1) = 1$	% of $Y_i(0) = 1$
50	1111	10	5
15	1110	20	15
5	1100	40	35
5	1010	30	-
10	1000	50	-
15	0000	-	-

though the true SACE is positive. From this example, we see that making use of the survival information after measurement may provide us with more information and narrow the bounds on the SACE.

Example 2

Through elementary calculation, one can easily prove that the lower bound for the SACE under our assumptions 4-6 will be at least equal to or larger than the lower bound for SACE under the ranked average score with one stage survival information assumption. However, the upper bound under our two stage survival information assumption is not comparable with the upper bound under one-stage survival assumption. Our upper bound can be smaller as shown in Example 1, but it can also be larger as we show below. Assume that the underlying truth about the population is described by the following Table 2.3.

The true SACE is 0.05. If we have an infinite sample, then we would have the following observed data:

$$p_{11|1} = 0.7 \quad p_{10|1} = 0.15 \quad p_{00|1} = 0.15 \quad p_{11|0} = 0.5 \quad p_{10|0} = 0.2 \quad p_{00|0} = 0.3 \tag{2.3.27}$$

$$q_{111|1} = 0.095 \quad q_{110|1} = 0.07 \quad q_{111|0} = 0.025 \quad q_{110|0} = 0.04 \tag{2.3.28}$$

Given the constraints imposed by the observed data (2.3.27)-(2.3.28) and assumptions 4-6, we obtain the bounds for the SACE: [0.043, 0.114]. If we don't utilize the second time survival information, we would observe the following data:

$$P(S_{1i} = 1|D_i = 1) = 0.85 \quad P(S_{1i} = 1|D_i = 0) = 0.7 \tag{2.3.29}$$

$$P(Y_i = 1, S_{1i} = 1|D_i = 1) = 0.165 \quad P(Y_i = 1, S_{1i} = 1|D_i = 0) = 0.065 \tag{2.3.30}$$

Then, given the constraints imposed by the observed data (2.3.29)-(2.3.30) and the ranked average score with one stage survival information assumption, the bound we would obtain for the SACE is [-0.071, 0.101]. In this setup, the upper bound under the ranked average score with two stage survival information assumption (Assumption 4-6) is larger than that of the ranked average score with one stage survival information assumption. The reason is that the ranked average score with two stage survival information assumptions allow for the possibility that the always survivors' (1111, 1110, 1100) probability of bad outcome exceed the protecteds' (1010, 1000) probability of bad outcome which contradicts the ranked average score with one stage survival information assumption.

2.4 Extension to IV settings

The idea of using second time point survival information to sharpen the inference of SACE under randomized trials with perfect compliance can be naturally extended to randomized trials with noncompliance or observational studies with a valid IV to obtain inference about the complier survivor average causal effect (CSACE). In a randomized trial with noncompliance, the assignment of treatment can be used as an IV to assess the effects of receiving the treatment on the outcome. In observational studies, natural experiments such as a person's draft lottery number, randomly assigned federal judges or quarter of birth have been used as IVs. (Angrist, 1990; Angrist and Krueger, 1991; Kling, 1999). For more literatures on IV, see Angrist, Imbens, and Rubin (1996), Abadie (2002), Hernan and Robins (2006), Tan (2006), Brookhart and Schneeweiss (2007), Cheng (2009), and Clarke and Windmeijer (2012).

Let Z_i represent the binary IV; 1 encourages the treatment for the i^{th} subject and 0 does not provide encouragement of the treatment. We use $\boldsymbol{Z}$ to denote the vector of IV for all subjects. Let $D_i(\boldsymbol{z})$ be the potential binary treatment variable that would be observed under IV assignment $\boldsymbol{z}$ for subject i; 1 being the treatment and 0 denotes the control. Let $S_{1i}(\boldsymbol{z})$ be the potential survival indicator of subject i that would be observed at the first time point after which the measurement of non-mortality outcome is taken; with 0 indicating death, 1 if alive. Let $Y_i(\boldsymbol{z})$ represent the potential non-mortality binary outcome that would be observed under

IV assignment $\boldsymbol{z}$. Again, the non-mortality outcome would be measured after the first time point, thus if the subject would die before that time point ($S_{1i}(\boldsymbol{z}) = 0$), $Y_i(\boldsymbol{z})$ is not defined; otherwise $S_{1i}(\boldsymbol{z}) = 1$ and $Y_i(z) = 1$ or 0, 1 indicating a worse outcome. We further define $S_{2i}(\boldsymbol{z})$ to be the potential indicator of survival at the second time point for subject i that would be observed if under IV assignment $\boldsymbol{z}$. As in section 2.2, if $S_{1i}(\boldsymbol{z}) = 0$, then $S_{2i}(\boldsymbol{z}) = 0$ by definition. We use Z_i, D_i, S_{1i}, Y_i and S_{2i} to denote respectively the observed IV, treatment received, observed survival indicator at the first time point, observed non-mortality outcome and observed survival indicator at the second time point for subject i.

2.4.1 Assumptions

We assume the following assumptions hold for the IV setup. These assumptions combine those of Angrist, Imbens and Rubin (1996) for the IV setup and the ranked average score with two stage survival information assumptions of section 2.2.

Assumption IV-1. Stable unit treatment value assumption (SUVTA).

- If $z_i = z_i'$, then $D_i(\boldsymbol{z}) = D_i(\boldsymbol{z'})$, $S_{1i}(\boldsymbol{z}) = S_{1i}(\boldsymbol{z'})$, $S_{2i}(\boldsymbol{z}) = S_{2i}(\boldsymbol{z'})$, and $Y_i(\boldsymbol{z}) = Y_i(\boldsymbol{z'})$

SUVTA means that a subject's potential treatments and outcomes are not affected by other individuals' IV status and means that we can write $D_i(\boldsymbol{z})$ as $D_i(z_i)$, $S_{1i}(\boldsymbol{z})$ as $S_{1i}(z_i)$, $S_{2i}(\boldsymbol{z})$ as $S_{2i}(z_i)$ and $Y_i(\boldsymbol{z})$ as $Y_i(z_i)$

Assumption IV-2. Nonzero average causal effect of Z on D. The average causal effect of Z on D, E$[D_i(1) - D_i(0)]$, is not equal to zero.

Assumption IV-3. Independence of the instrument from unmeasured confounders: the random vector $(D(1), D(0), S_1(1), S_1(0), S_2(1), S_2(0), Y(1), Y(0))$ is independent of Z.

Based on subjects' compliance behavior, we can first partition the population into four groups:

$$U_i = \begin{cases} 00, & \text{if } D_i(1) = D_i(0) = 0 \\ 10, & \text{if } D_i(1) = 1,\ D_i(0) = 0 \\ 11, & \text{if } D_i(1) = D_i(0) = 1 \\ 01, & \text{if } D_i(1) = 0,\ D_i(0) = 1 \end{cases} \tag{2.4.1}$$

where 00, 10, 11, and 01 represent never taker, complier, always taker and defier, respectively. Because $D_i(1)$ and $D_i(0)$ are never observed jointly, the compliance behavior of a subject is unknown.

Assumption IV-4. Monotonicity of effect of IV on treatment: $D(1) \geq D(0)$. There is no U=01 group.

Assumption IV-5. Monotonicity of effect of IV on survival: $S_{1i}(1) \geq S_{1i}(0)$, $S_{2i}(1) \geq S_{2i}(0)$.

The monotonicity of the effect of the IV on the survival will hold if the treatment never causes death and assumption IV-4 holds if the IV has a monotone effect on treatment.

Table 2.4: Principal Strata

$D_i(1)$	$D_i(0)$	$S_{1i}(1)$	$S_{1i}(0)$	Principal Strata
1	0	1	1	Complier, always survivors
1	0	1	0	Complier, protected
1	0	0	0	Complier, never survivors
1	1	1	1	Never taker, always survivors
1	1	0	0	Never taker, never survivors
0	0	1	1	Always taker, always survivors
0	0	0	0	Always taker, never survivors

Assumption IV-6. Exclusion restrictions among never-takers and always-takers: $S_{1i}(1) = S_{1i}(0), S_{2i}(1) = S_{2i}(0), Y_i(1) = Y_i(0)$, for $U_i = 00$ or 11.

This means that the IV only affects the outcomes through treatment and has no direct effect on outcomes.

Based on the possible joint combinations of $(D_i(1), D_i(0), S_{1i}(1), S_{1i}(0))$ under the above assumptions, we can define principal strata as shown in Table 2.4.

Different from the case of randomized experiments with perfect compliance, the principal strata in the IV setup are defined with respect to IV levels, for example, the "complier, always survivors" are compliers(comply with their IV encouragement of treatment) who would survive under both IV levels. Among all the principal strata, the "complier, always survivors" (1011) group is the only group that we can observe the outcome under treatment if IV is 1, as well as the outcome under

Table 2.5: Fine Strata

Probability	$D_i(1)$	$D_i(0)$	$S_{1i}(1)$	$S_{1i}(0)$	$S_{2i}(1)$	$S_{2i}(0)$	Principal Strata at Time Point 1
π_{101111}	1	0	1	1	1	1	Complier, always survivors
π_{101110}	1	0	1	1	1	0	Complier, always survivors
π_{101100}	1	0	1	1	0	0	Complier, always survivors
π_{101010}	1	0	1	0	1	0	Complier, protected
π_{101000}	1	0	1	0	0	0	Complier, protected
π_{100000}	1	0	0	0	0	0	Complier, never survivors
π_{111111}	1	1	1	1	1	1	Always takers, always survivors
π_{111100}	1	1	1	1	0	0	Always takers, always survivors
π_{110000}	1	1	0	0	0	0	Always takers, never survivors
π_{001111}	0	0	1	1	1	1	Never takers, always survivors
π_{001100}	0	0	1	1	0	0	Never takers, always survivors
π_{000000}	0	0	0	0	0	0	Never takers, never survivors

control if IV is 0, and that would survive under both treatment such that the non-mortality outcome Y is well defined in both cases. Thus, it is the only group for which variation in the IV can identify the causal effect of the treatment on the non-mortality outcome: $CSACE = \mathrm{E}(Y_i(1) - Y_i(0) \mid 1011)$.

Similarly to the case of randomized experiments with perfect compliance (Section 2.2), we can further incorporate the information of second time survival indicator to create finer strata as shown in Table 2.5.

In terms of the fine strata in Table 2.5, the CSACE is expressed as:

$$
\begin{aligned}
CSACE &= \mathrm{E}(Y_i(1) - Y_i(0) \mid 1011) \\
&= P(Y_i(1) = 1 \mid 1011) - P(Y_i(0) = 1 \mid 1011) \\
&= \frac{(\pi_{101111}\mathrm{E}(Y_i(1) \mid 101111) + \pi_{101110}\mathrm{E}(Y_i(1) \mid 101110) + \pi_{101100}\mathrm{E}(Y_i(1) \mid 101100))}{\pi_{101111} + \pi_{101110} + \pi_{101100}} \\
&- \frac{(\pi_{101111}\mathrm{E}(Y_i(0) \mid 101111) + \pi_{101110}\mathrm{E}(Y_i(0) \mid 101110) + \pi_{101100}\mathrm{E}(Y_i(0) \mid 101100))}{\pi_{101111} + \pi_{101110} + \pi_{101100}}
\end{aligned}
\tag{2.4.2}
$$

The same assumptions are made for compliers as we made for subjects under randomized trials with perfect compliance (Assumptions 4-6 in Section 2.2).

Assumption IV-7. Among "complier, always survivors", the probability of worse outcome for group 101111 is the lowest, whereas the probability of worse outcome for group 101100 is the highest under both treatment arms:

$$P(Y_i(1) = 1 \mid 101111) \leq P(Y_i(1) = 1 \mid 101110) \leq P(Y_i(1) = 1 \mid 101100) \tag{2.4.3}$$

$$P(Y_i(0) = 1 \mid 101111) \leq P(Y_i(0) = 1 \mid 101110) \leq P(Y_i(0) = 1 \mid 101100) \tag{2.4.4}$$

Assumption IV-8. Among "complier, protected", the probability of worse outcome for group 101010 is no higher than that for group 101000 under treatment:

$$P(Y_i(1) = 1 \mid 101010) \leq P(Y_i(1) = 1 \mid 101000) \tag{2.4.5}$$

Assumption IV-9. Under treatment, the probability of worse outcome for group 101100 is not lower than that for group 101010, but not higher than that for group 101000, and the probability of worse outcome for group 101110 is not higher than

that for group 101010:

$$P(Y_i(1) = 1 \mid 101110) \leq P(Y_i(1) = 1 \mid 101010) \leq P(Y_i(1) = 1 \mid 101100) \leq P(Y_i(1) = 1 \mid 101000) \tag{2.4.6}$$

2.4.2 Derivations of Bounds

As for the SACE in randomized experiments setup, the CSACE is not point identified without further assumptions based on the observable joint distribution of $(Z_i, D_i, S_{1i}, S_{2i}, Y_i)$, but can be bounded. We will again adopt the two step method we used in section 2.3 to obtain the bound.

The observable strata of $(Z_i, D_i, S_{1i}, S_{2i})$ are mixtures of fine strata, if we use $p_{s_1 s_2 d|z}$ to denote $P(S_{1i} = s_1, S_{2i} = s_2, D_i = d \mid Z_i = z)$, we have the following identities:

$$p_{111|1} = \pi_{101111} + \pi_{101110} + \pi_{101010} + \pi_{111111} \tag{2.4.7}$$

$$p_{101|1} = \pi_{101100} + \pi_{101000} + \pi_{111100} \tag{2.4.8}$$

$$p_{001|1} = \pi_{100000} + \pi_{110000} \tag{2.4.9}$$

$$p_{110|1} = \pi_{001111} \tag{2.4.10}$$

$$p_{100|1} = \pi_{001100} \tag{2.4.11}$$

$$p_{000|1} = \pi_{000000} \tag{2.4.12}$$

$$p_{110|0} = \pi_{101111} + \pi_{001111} \tag{2.4.13}$$

$$p_{100|0} = \pi_{101100} + \pi_{001100} + \pi_{101110} \tag{2.4.14}$$

$$p_{000|0} = \pi_{100000} + \pi_{000000} + \pi_{101010} + \pi_{101000} \tag{2.4.15}$$

$$p_{111|0} = \pi_{111111} \tag{2.4.16}$$

$$p_{101|0} = \pi_{111100} \tag{2.4.17}$$

$$p_{001|0} = \pi_{110000} \tag{2.4.18}$$

and the constraint

$$0 \leq \pi_{101111}, \pi_{101110}, \pi_{101100}, \pi_{101010}, \pi_{101000}, \pi_{100000}, \pi_{111111}, \pi_{111100}, \pi_{110000}, \pi_{001111}, \pi_{001100}, \pi_{000000} \leq 1 \tag{2.4.19}$$

Given (2.4.7)-(2.4.18), we can express each π in terms of $p_{s_1 s_2 d|z}$ and π_{101100}:

$$\pi_{000000} = p_{000|1}$$

$$\pi_{001111} = p_{110|1}$$

$$\pi_{001100} = p_{100|1}$$

$$\pi_{111111} = p_{111|0}$$

$$\pi_{110000} = p_{001|0}$$

$$\pi_{111100} = p_{101|0}$$

$$\pi_{100000} = p_{001|1} - p_{001|0}$$

$$\pi_{101111} = p_{110|0} - p_{110|1}$$

$$\pi_{101000} = p_{101|1} - p_{101|0} - \pi_{101100}$$

$$\pi_{101110} = p_{100|0} - p_{100|1} - \pi_{101100}$$

$$\pi_{101010} = p_{111|1} + p_{110|1} + p_{100|1} - p_{110|0} - p_{100|0} - p_{111|0} + \pi_{101100}$$

and subject to the constraint of (2.4.19), we have,

$$\max\{0, p_{110|0}+p_{100|0}+p_{111|0}-p_{111|1}-p_{110|1}-p_{100|1}\} \leq \pi_{101100} \leq \min\{p_{101|1}-p_{101|0}, p_{100|0}-p_{100|1}\} \tag{2.4.20}$$

Bounds for the CSACE

For fixed $\pi's$, let $q_{ys_1s_2d|z}$ denote $P(Y_i = y, S_{1i} = s_1, S_{2i} = s_2, D_i = d \mid Z_i = z)$. We have the following identities based upon the observable strata of $(Y_i, S_{1i}, S_{2i}, D_i, Z_i)$:

$$q_{1111|1} = \pi_{101111}\mathrm{E}(Y_i(1) \mid 101111)+\pi_{101110}\mathrm{E}(Y_i(1) \mid 101110)+\pi_{101010}\mathrm{E}(Y_i(1) \mid 101010)+\pi_{111111}\mathrm{E}(Y_i(1) \mid 111111) \tag{2.4.21}$$

$$q_{1101|1} = \pi_{101100}\mathrm{E}(Y_i(1) \mid 101100) + \pi_{101000}\mathrm{E}(Y_i(1) \mid 101000) + \pi_{111100}\mathrm{E}(Y_i(1) \mid 111100) \tag{2.4.22}$$

$$q_{1110|1} = \pi_{001111}\mathrm{E}(Y_i(1) \mid 001111) \tag{2.4.23}$$

$$q_{1100|1} = \pi_{001100}\mathrm{E}(Y_i(1) \mid 001100) \tag{2.4.24}$$

$$q_{1111|0} = \pi_{111111}\mathrm{E}(Y_i(0) \mid 111111) \tag{2.4.25}$$

$$q_{1101|0} = \pi_{111100}\mathrm{E}(Y_i(0) \mid 111100) \tag{2.4.26}$$

$$q_{1110|0} = \pi_{101111}\mathrm{E}(Y_i(0) \mid 101111) + \pi_{001111}\mathrm{E}(Y_i(0) \mid 001111) \tag{2.4.27}$$

$$q_{1100|0} = \pi_{101110}\mathrm{E}(Y_i(0) \mid 101110) + \pi_{101100}\mathrm{E}(Y_i(0) \mid 101100) + \pi_{001100}\mathrm{E}(Y_i(0) \mid 001100) \tag{2.4.28}$$

Recall that

$$\begin{aligned} CSACE = &\frac{(\pi_{101111}\mathrm{E}(Y_i(1) \mid 101111) + \pi_{101110}\mathrm{E}(Y_i(1) \mid 101110) + \pi_{101100}\mathrm{E}(Y_i(1) \mid 101100))}{\pi_{101111} + \pi_{101110} + \pi_{101100}} \\ &- \frac{(\pi_{101111}\mathrm{E}(Y_i(0) \mid 101111) + \pi_{101110}\mathrm{E}(Y_i(0) \mid 101110) + \pi_{101100}\mathrm{E}(Y_i(0) \mid 101100))}{\pi_{101111} + \pi_{101110} + \pi_{101100}} \end{aligned} \tag{2.4.29}$$

Given $\pi's$,
$\frac{(\pi_{101111}\mathrm{E}(Y_i(0)|101111)+\pi_{101110}\mathrm{E}(Y_i(0)|101110)+\pi_{101100}\mathrm{E}(Y_i(0)|101100))}{\pi_{101111}+\pi_{101110}+\pi_{101100}} = \frac{q_{1110|0}+q_{1100|0}-q_{1110|1}-q_{1100|1}}{\pi_{101111}+\pi_{101110}+\pi_{101100}}$
which is point identified. Thus to bound the CSACE, we only need to bound $\pi_{101111}\mathrm{E}(Y_i(1) \mid 101111) + \pi_{101110}\mathrm{E}(Y_i(1) \mid 101110) + \pi_{101100}\mathrm{E}(Y_i(1) \mid 101100)$, which defines a linear programming problem:

$$\min / \max \quad (\pi_{101111}\mathrm{E}(Y_i(1) \mid 101111) + \pi_{101110}\mathrm{E}(Y_i(1) \mid 101110) + \pi_{101100}\mathrm{E}(Y_i(1) \mid 101100)) \mid \pi_{101100} \tag{2.4.30}$$

Subject to:

$$q_{1111|1} - q_{1111|0} = \pi_{101111}\mathrm{E}(Y_i(1) \mid 101111) + \pi_{101110}\mathrm{E}(Y_i(1) \mid 101110) + \pi_{101010}\mathrm{E}(Y_i(1) \mid 101010) \tag{2.4.31}$$

$$q_{1101|1} - q_{1101|0} = \pi_{101100}\mathrm{E}(Y_i(1) \mid 101100) + \pi_{101000}\mathrm{E}(Y_i(1) \mid 101000) \tag{2.4.32}$$

$$\mathrm{E}(Y_i(1) \mid 101111) \leq \mathrm{E}(Y_i(1) \mid 101110) \leq \mathrm{E}(Y_i(1) \mid 101100) \tag{2.4.33}$$

$$\mathrm{E}(Y_i(1) \mid 101010) \leq \mathrm{E}(Y_i(1) \mid 101000) \tag{2.4.34}$$

$$\mathrm{E}(Y_i(1) \mid 101110) \leq \mathrm{E}(Y_i(1) \mid 101010) \leq \mathrm{E}(Y_i(1) \mid 101100) \leq \mathrm{E}(Y_i(1) \mid 101000) \tag{2.4.35}$$

$$0 \leq \mathrm{E}(Y_i(1) \mid 101111), \mathrm{E}(Y_i(1) \mid 101110), \mathrm{E}(Y_i(1) \mid 101100), \mathrm{E}(Y_i(1) \mid 101010), \mathrm{E}(Y_i(1) \mid 101000) \leq 1 \tag{2.4.36}$$

where constraints (2.4.33)-(2.4.35) are imposed by assumptions (IV-7) - (IV-9).

The above linear programming problem has a solution if and only if $\frac{q_{1101|1}-q_{1101|0}}{p_{101|1}-p_{101|0}} \geq \frac{q_{1111|1}-q_{1111|0}}{p_{111|1}-p_{111|0}}$.

For each possible value of π_{101100}, we can solve the above linear programming problem; then, combining this result with the bound for π_{101100}, let $L = p_{110|0} + p_{100|0} + p_{111|0} - p_{111|1} - p_{110|1} - p_{100|1}$, $U = \min\{p_{101|1} - p_{101|0}, p_{100|0} - p_{100|1}\}$, then $\pi_{101100} \in I$, where $I = [\max\{0, L\}, \quad U]$, we obtain

$$\begin{aligned}\max CSACE = {} & \frac{q_{1111|1} - q_{1111|0}}{p_{111|1} - p_{111|0}} - \frac{q_{1110|0} - q_{1110|1} + q_{1100|0} - q_{1100|1}}{p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1}} \\ & + \frac{(q_{1101|1} - q_{1101|0})(p_{111|1} - p_{111|0}) - (q_{1111|1} - q_{1111|0})(p_{101|1} - p_{101|0})}{(p_{101|1} - p_{101|0})(p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1})(p_{111|1} - p_{111|0})} \cdot U\end{aligned}$$

If $L \geq 0$

$$\begin{aligned}\min CSACE = \max\{ & \frac{q_{1111|1} - q_{1111|0} + q_{1101|1} - q_{1101|0} - p_{111|1} + p_{111|0} - p_{101|1} + p_{101|0} + p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1}}{p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1}}, \\ & \frac{q_{1111|1} - q_{1111|0}}{p_{111|1} - p_{111|0}}\} - \frac{q_{1110|0} - q_{1110|1} + q_{1100|0} - q_{1100|1}}{p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1}}\end{aligned}$$

If $L < 0$

$$\begin{aligned}\min CSACE = \max\{ & \frac{(q_{1111|1} - q_{1111|0})(p_{101|1} - p_{101|0}) + (q_{1101|1} - q_{1101|0})(p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1} - p_{111|1} + p_{111|0})}{(p_{101|1} - p_{101|0})(p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1})}, \\ & 0\} - \frac{q_{1110|0} - q_{1110|1} + q_{1100|0} - q_{1100|1}}{p_{110|0} - p_{110|1} + p_{100|0} - p_{100|1}}\end{aligned}$$

2.5 Checking the plausibility of ranked average score with two stage survival assumptions and exclusion restriction assumptions

From the observable data, it cannot be determined whether our ranked average score with two stage survival information assumptions for randomized experiments setup or IV settings hold, also it cannot be determined whether the exclusion restriction assumed in the IV settings hold. However, there are some necessary conditions that the probability distribution of the observable data must satisfy when these assumptions are valid. If these conditions are violated, then we know our assumptions do not hold.

For randomized experiments with perfect compliance, from the derivation of the bound for SACE in section 2.3, we know that the linear programming problem (2.3.14)-(2.3.20) under the ranked average score with two stage survival information assumptions as well as the constraints imposed by the observable"infinite sample" probability distribution has a solution if and only if

$$\frac{q_{110|1}}{p_{10|1}} \geq \frac{q_{111|1}}{p_{11|1}} \tag{2.5.1}$$

This constraint says that the probability of the worse non-mortality outcome among the patients that are randomly assigned to treatment and that survive to the first time point but die before the second time point is equal to or larger than

the probability of the worse non-mortality outcome among the patients that are randomly assigned to treatment and that survive at least to the second time point. This is a direct result from our ranked average score with two stage survival assumptions (Assumptions 4-6) which say that $E(Y_i(1) \mid 1111) \leq E(Y_i(1) \mid 1110) \leq E(Y_i(1) \mid 1010) \leq E(Y_i(1) \mid 1100) \leq E(Y_i(1) \mid 1000)$. The first three expectations are for subjects who can survive at least to the second time point under treatment and the last two expectations are for subjects who die before the second time point.

For the IV setting of Section 2.4, based on the calculations in section 2.4.2, the corresponding necessary conditions that the probability distribution of the data must satisfy under Assumptions (IV-1)-(IV-9) are as follows: ,

$$q_{1101|1} - q_{1101|0} \geq 0,\ p_{101|1} - p_{101|0} \geq 0,\ q_{1111|1} - q_{1111|0} \geq 0,\ p_{111|1} - p_{111|0} \geq 0 \quad (2.5.2)$$

$$\frac{q_{1101|1} - q_{1101|0}}{p_{101|1} - p_{101|0}} \geq \frac{q_{1111|1} - q_{1111|0}}{p_{111|1} - p_{111|0}} \quad (2.5.3)$$

Pearl (1995) provides a necessary condition on the joint probability distribution of the outcome, treatment and IV when the exclusion restriction holds. Extending Pearl's result to our case where exclusion restrictions are assumed on both survival at the first time point and the second time point as well as a non-mortality outcome which may be censored, a necessary condition is that the following inequalities hold:

$$p_{00d|z_1} + q_{010d|z_2} + q_{110d|z_3} + q_{011d|z_4} + q_{111d|z_5} \leq 1 \quad (2.5.4)$$

where $d \in \{0, 1\}$, $z_i \in \{0, 1\}$ for $i = 1, 2, 3, 4, 5$

The above constraints to check the plausibility of our assumptions are for "infinite sample" data. In practice, we can estimate the confidence with which the true observable population distribution satisfies the above constraints using a simple bootstrap procedure (Efron and Tibshirani, 1998). We bootstrap from the empirical distribution of the observed data and then count the percentage of the bootstrapped data sets for which the empirical distribution satisfies the constraints as an estimate of the confidence. Efron and Tibshirani (1998) provide some refinements on this simple bootstrap procedure that improve the accuracy of the estimated confidence.

2.6 Confidence Intervals for Bounds

In sections 2.3 and 2.4, the bounds we obtained are "infinite sample" bounds where we assume that the joint distributions of $(D_i, S_{1i}, S_{2i}, Y_i)$ or $(Z_i, D_i, S_{1i}, S_{2i}, Y_i)$ is known. However, in practice, all these probabilities need to be estimated from the observed data. To account for the sampling uncertainty, we would like to construct confidence intervals for the bounds. The simplest way to construct confidence interval is through the Bonferroni method, where if we want an overall level of $1-\alpha$, we can obtain first the individual $1-\frac{\alpha}{2}$ confidence interval for the upper bound and lower bound (e.g., via the bootstrap), then combine the results to derive the simultaneous confidence interval. The disadvantage of the Bonferroni method is it's conservative; the way to form it ignores the joint distribution of the upper bound and lower bound. Horowitz and Manski (2000) proposed a method to obtain the

confidence interval taking into account the joint distribution of the lower and upper bound. The Horowitz and Manski confidence interval adds the same length to the upper and lower bounds in the confidence interval. Beran (1988) proposed the B method which also takes into account the joint distribution of upper and lower bounds without the restriction on the form of the confidence interval of the Horowitz and Manski confidence interval. A description of the above confidence interval approaches for bounds can be found in Cheng and Small (2006). Because of the nice properties of B method, we will use it to construct the confidence interval for the ARDSNet study.

We did a simulation study to examine the finite sample coverage of the B method 95% confidence interval for data like the ARDSNet study (See Section 2.7). We simulated 2000 samples based on the observed empirical distribution of the ARDSNet data (Table 2.6). Then for each simulated data set, we bootstrapped 2000 data sets to obtain the 95% B method confidence interval. We counted the proportion of the two thousand bootstrap CIs that cover the bound of the empirical distribution of the ARDSNet data and did the analysis using both the two stage and one stage assumptions. For the ranked average score with two stage survival information assumptions, the coverage probability of the B method is estimated to be 95.65%, and for the ranked average score with one stage survival information assumptions, the coverage probability of the B method is estimated to be 95.75%. Thus the finite sample coverage of the B method for studies like the ARDSNet study seems to be

good.

2.7 Application to ARDSNet Study

The ARDSNet study described in the introduction involved 861 patients with lung injury and acute respiratory distress syndrome who were randomized to receive mechanical ventilation with either lower tidal volumes or traditional tidal volumes. The non-mortality outcome variable we are interested in is whether patients were able to breathe without assistance by day 28 which is a measurement that reflects the quality of life for patients after treatment. We use Y_i to represent this binary quality of life measurement, with Y_i being 1 indicating that the i^{th} patient were not able to breathe without assistance by day 28. Naturally, the first survival time point is day 28 after the treatment. If the patient died before day 28, then the non-mortality outcome could not be measured, thus will be undefined. The second time point survival indicator is whether the patient was eventually discharged home with unassisted breathing or not. We view the patients who received mechanical ventilation with lower tidal volume as the treatment group, and the patients who received mechanical ventilation with traditional tidal volume as the control group. Let D_i equal 1 if the i^{th} patient is randomized to treatment group, 0 if randomized to control group. Further details on the data are described in appendix.

Table 2.6 presents the observed strata of $(D_i, S_{1i}, S_{2i}, Y_i)$. Among the survivors in the lower tidal volume group, the proportion of patients that cannot breathe

Table 2.6: Observed data for ARDSNet Study

Number of Patients	D_i	S_{1i}	S_{2i}	Y_i
258	1	1	1	0
29	1	1	1	1
10	1	1	0	0
26	1	1	0	1
109	1	0	0	–
211	0	1	1	0
34	0	1	1	1
7	0	1	0	0
25	0	1	0	1
152	0	0	0	–

Table 2.7: The estimated bounds and 95% B method CIs of the SACE for ARDSNet study using ranked average score with two stage survival assumptions and one stage survival assumptions.

SACE	Two-stage survival assumptions	One-stage survival assumptions
Estimated bounds	[−12.99%, −4.02%]	[−17.38%, −4.27%]
95% confidence interval	[−20.11%, 1.99%]	[−27.57%, 2.18%]

without assistance by day 28 is 17.03% (which is 55/323); among the survivors in the traditional tidal volume group, the proportion of patients that cannot breathe without assistance by day 28 is 21.30% (which is 59/277). The difference of those two proportions −4.27% which is a direct comparison of the QOL among survivors in the lower tidal volume and survivors in the traditional tidal volume is likely an upward biased estimate for the SACE due to the informativeness of censoring by death.

The empirical distribution of $(D_i, S_{1i}, S_{2i}, Y_i)$ satisfies the constraint (2.5.1). Using the bootstrap procedure, all of the 2000 bootstrapped datasets satisfy the constraint (2.5.1), thus we are very confident that our set of two stage assumptions is plausible in the sense that it does not violate the constraint (2.5.1).

Table 2.7 compares the estimated bounds of the SACE as well as the 95% confidence intervals obtained through our proposed ranked average score with two stage survival information assumptions to the ranked average score with one stage survival information assumptions. According to the result of our two stage analysis, among the patients with lung injury and the acute respiratory distress syndrome who would

survive under both ventilation tidal volumes, the lower tidal volume would help reduce the probability of breathing with assistance by day 28 by an amount between 4.02% to 12.99%. This bound for the SACE is substantially shorter, thus more informative, than the bound obtained through the one stage analysis which estimates the reduction to be between $[4.27\%, 17.38\%]$. The 95% B method confidence intervals under both sets of assumptions cover 0, meaning that there is not strong evidence that ventilation with lower tidal benefits patients in terms of the quality of life outcome of breathing without assistance by day 28.

2.8 Conclusions and Discussions

The effect of treatment on a non-mortality outcome among always survivors is of interest in many clinical studies. The previous literature on bounding the SACE uses only the survival information before the measurement of the non-mortality outcome; however, in many cases, the survival information after the measurement of non-mortality outcome is informative. We proposed a set of ranked average score with two stage survival information assumptions which are plausibly satisfied in many quality of life studies and developed a two-step linear programming approach to obtain the closed form of the bounds of the SACE under our assumptions. Our method works not only for randomized trials with perfect compliance, but also can be extended to randomized trials with noncompliance or observational studies with a valid IV to obtain bounds on the complier survivor average causal effect.

We applied our method to the ARDSNet study. Making use of the post QOL measurement survival information (patients' status when discharged home) in addition to the pre-QOL survival information (survival status at day 28) helps substantially shorten the bound on the SACE – the effect of lower tidal volume on being able to breathe without assistance by day 28.

The SACE and CSACE are principal strata effects, causal effects on a subgroup of patients defined by the values that post-randomization variables would take under both treatment and control (Frangakis and Rubin, 2002). We have shown that bounds on these principal strata effects can be sharpened by using the further outcome information of survival after the non-mortality outcome is measured. In a different context, Mealli and Pacini (2013) showed that using further outcomes can narrow bounds on principal strata effects. Mealli and Pacini consider an outcome that is not affected by censoring by death in a randomized trial with noncompliance, and study bounds on the intention to treat effects for the compliers, always takers and never takers. Mealli and Pacini consider settings in which the exclusion restriction may not be satisfied and they show that a secondary outcome for which the exclusion restriction is satisfied can be used to narrow the bounds. For randomized trials with noncompliance in which there is censoring by death and the exclusion restriction may not be satisfied, it would be of future research interest to consider combining the post-quality of life measurement survival information we have studied with the secondary outcomes Mealli and Pacini studied to narrow the

bounds on the CSACE.

So far, we have assumed that we are in the context of a randomized trial or an observational study with a valid IV. Our method can also be naturally extended to the cases in which conditional on some discrete covariates there is ignorability such that the subjects are randomized or the IV is valid conditional on the covariates. We can stratify the subjects into subsets defined by each level of covariates, and apply our method to obtain the bound of SACE within each subgroup. Then we can obtain the overall bound of SACE combining the proportions of each subgroup. See (Freiman and Small, 2013) for more details on this topic. How to deal with the case in which the covariates are continuous requires further research.

In this study, we focus on studies where the non-mortality outcome is measured at a fixed time for all subject. However, there are cases where the non-mortality outcome might be measured at different time for different subjects which complicates the analysis. For instance, IVH may happen at any time in the first several days of life of babies. How to handle the situation in which the non-mortality outcome could be measured at continuous time period is a topic we are working on.

Chapter 3

IV with Nonignorable Missing Covariates

3.1 Introduction

3.1.1 Effect of type of delivery NICUs on premature infants

Premature infants are infants born before a gestational age of 37 complete weeks. Compared to term infants, premature infants have less time to develop, so that they are at higher risk of death and complications and often in need of advanced care, ideally in a neonatal intensive care unit (NICU) (Profit et al., 2010; Doyle et al., 2004; Boyle et al., 1983). There are two types of NICUs - a high level NICU is a NICU that has the capacity for sustained mechanical assisted ventilation and that delivers on average of at least 50 premature babies per year, whereas a low level NICU is

a unit that does not meet these requirements. There is literature that shows that delivery at high level vs. low level NICUs is associated with a reduction in neonatal mortality after controlling for measured confounders (Phibbs et al., 2007; Chung et al., 2010; Rogowski et al., 2004). However, there are unmeasured confounders such as fetal heart tracing test results and severity of conditions that could bias these results. The aim of this chapter is to use the instrumental variable method along with a novel method of controlling for nonignorable missing covariates to obtain unbiased inferences about the effect on neonatal mortality of premature babies being delivered in a high level NICU vs. a low level NICU. Understanding how effective high level NICUs are compared to low level NICUs is important for both individual mothers deciding whether to travel a distance to go to a high level NICU rather than going to a local low level NICU, and also for public policy decisions about premature infant care. In the 1970s, a system of perinatal regionalization was built in most states in which most infants at risk of complications such as very premature infants would be sent to regional high level NICUs (Lasswell et al., 2010). This regionalization system has weakened in recent years with more very premature infants being born in low level NICUs (Lasswell et al., 2010; Howell et al., 2002; Richardson et al., 1995; Yeast et al.,1998). If high level NICUs are truly providing considerably better care for premature babies, then it is valuable to invest resources in strengthening the perinatal regionalization system, while if high level NICUs are providing at best marginal improvements in care, then strengthening the perinatal

regionalization should probably not be a priority. Additionally, if only certain types of premature babies benefit from high level NICUs (e.g., only those below a certain gestational age), then resources would be best spent on increasing the rate of high level NICU delivery for those types of babies. To address this, we will estimate the effect of high level NICU delivery for babies with different characteristics, such as different gestational ages.

The ideal way to assess the effectiveness of high level NICUs vs. low level NICUs would be to randomize pregnant women to deliver at different level NICUs, but such a study is not ethical or practical. We instead consider an observational study. We have compiled data on all babies born prematurely in Pennsylvania between 1995-2005 by linking birth certificates to death certificates as well as maternal and newborn hospital records. More than 98% of the birth certificates could be linked to the hospital records (Lorch et al. (2012) for more details). We will use the 189,991 records that could be linked in our analysis. The measured confounders we will consider are gestational age, the month of pregnancy that prenatal care started (precare), and mother's education level. If these measured confounders are the only confounding variables, i.e., the only variables that are related to both level of NICU delivered at and mortality, then we could use propensity score/matching/regression methods to control for the confounders. Unfortunately, some key confounders are unmeasured such as the results of tests like fetal heart tracing which are related to both how strongly a doctor encourages a woman to deliver at a high level NICU

and a baby's risk of mortality. To control for such unmeasured confounders, we will consider the instrumental variable (IV) method.

3.1.2 Instrumental variable approach

The IV method is widely used in observational studies (Angrist and Krueger, 1991; Baiocchi et al., 2010). An instrumental variable (IV) is a variable that is (i) associated with the treatment, (ii) has no direct effect on the outcome, and (iii) is independent of unmeasured confounders conditional on measured confounders. The relationships between the IV, treatment(D), outcome(Y), measured confounders ($\boldsymbol{X}$) and unmeasured confounders(UC) are shown in the directed acyclic graph in Figure 1. The basic idea of the IV method is to extract variation in the treatment that is free of the unmeasured confounders and use this confounder free variation to estimate the causal effect of the treatment on the outcome. The beauty of the IV method is that although treatment is not randomly assigned in observational studies, the method still allows consistent estimation of the causal effect of a treatment.

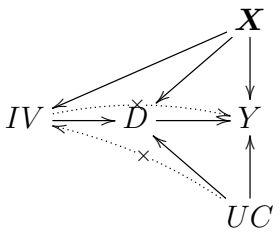

Figure 1. This directed acyclic graph shows the assumptions for a valid IV. D denotes the treatment, Y the outcome, $\boldsymbol{X}$ measured confounders and UC unmeasured confounders.

The key assumptions for an IV are (i) the IV affects D; (ii) the IV does not have a direct effect on Y; (iii) the IV is independent of the unmeasured confounders UC given the measured confounders.

The instrumental variable we consider is whether or not the excess travel time that a mother lives from the nearest high level NICU compared to the nearest low level NICU is less than or equal to 10 minutes; a mother is said to live "near" to a high level NICU if the excess travel time is $\leq$ 10 minutes and "far" otherwise. Excess travel time satisfies the first two characteristics of an IV: (i) association with treatment: previous studies suggest that women tend to deliver at NICUs near their residential zip code (Lorch et al., 2012; Phibbs et al., 1993) and (ii) no direct effect: most women have time to deliver at both the nearest high level or other delivery NICU so the marginal travel time to either facility should not directly affect outcomes (Lorch et al., 2012). The third assumption needed for excess travel time to be an IV, that it is independent of unmeasured confounders conditional on measured confounders, is plausible in that most women do not expect to have a premature delivery and hence do not choose where to live based on distance to a high level NICU. However, because high level NICUs tend to be in certain types of places (e.g. in cities) and people living in places with high level NICUs have different characteristics from people living far away from high level NICUs, for the third IV assumption to hold, we need to condition on these characteristics that may affect the risk of neonatal death in these pregnancies. The measured characteristics

we are able to condition on are the month of pregnancy that prenatal care started (precare), mother's education, and gestational age of the baby. We only have a small number of measured characteristics; for settings where there are a large numbers of measured characteristics, it is worth considering Lasso methods to control for the characteristics as in Imai and Ratkovic (2012). In previous work (Guo et al., 2014; Lorch et al., 2012), we used excess travel time as an IV to estimate the effect of high level vs. low level NICUs, but we did not account for the potential nonignorable missingness of certain measured characteristics. We will develop a method for accounting for nonignorable missing covariates.

3.1.3 Nonignorable missing covariates

Among the measured confounders, the gestational ages are completely recorded but some subjects' precare and education level are missing . We are concerned that the missingness is related with the outcome (death) and the risks of mother and infant. The information for mother is usually filled out partly by mother, and partly by the nurse or doctor. If the baby died, the mother may not want to fill out the questionnaire due to her grief, or nurses may not bother the mother to fill out a questionnaire out of caring for the mother's grief. When the mother or infant is at high risk of complications, nurses and doctors focus on this emergency and may ignore recording mother's information. Consequently, missingness is only plausibly ignorable if we condition on the outcome (death) and mother's/infant's

risk of complications. The outcome is fully observed but the mother's/infant's risk of complication is not fully observed. The measured variable gestational age is a strong predictor of risk but other predictors of risk that are known to the doctor but not recorded in the data include the results of fetal heart tracing and the doctor's knowledge about the severity of mother's and baby's condition. These unmeasured confounders may be related to the compliance status of the mother. The compliance status of the mother refers to whether the mother would deliver at a high level NICU if she lived near to one (excess travel time $\leq$ 10 minutes) and whether she would deliver at a high level NICU if she lived far from one (see section 3.2.2 for further discussion). If the mother would always deliver at a high level NICU regardless of whether she lives near to one, her compliance status is always taker. If the mother would only deliver at a high level NICU if she lives near one, her compliance status is complier. If a doctor knows that a baby/mother is at higher risk of complications based on fetal heart tracing or other knowledge, then the doctor is more likely to recommend the mother to deliver at a high level NICU regardless of how near she lives to the high level NICU and the mother is more likely to be an always taker. Thus, compliance status is related to unmeasured risk and consequently, the missingness of observed variables is likely to be related to compliance status. Compliance status is only partially observed, e.g., under the assumptions in section 3.2.2, if a mother lives far from a high level NICU, but still delivers at a high level NICU, she is an always taker, but if she lives near a high level NICU and delivers

at a high level NICU, she might be an always taker or complier.

Previous literature on IV with missing data has considered missing outcomes (Frangakis and Rubin, 1999; Levy, O'Malley and Normand, 2004; Mealli et al., 2004; Peng, Little and Raghunathan, 2004; Chen, Geng and Zhou, 2009; Small and Cheng, 2009). In this literature, it has been argued that ignorability of the missing outcome may only be plausible after conditioning on the covariates *and* the partially observed compliance status (see (3.2.1)). Methods have been developed for estimating causal effects under this "latent ignorability". For missing covariates rather than missing outcomes, the only work on IV estimation that we are aware of is Peng, Little and Raghunathan (2004), which assumes missingness of covariates is ignorable conditional on observed data, but not allowed to depend on compliance behavior. In this chapter, we develop a method for estimation of the causal effect when the missingness of covariates may depend on the fully observed data as well as the partially observed compliance behavior.

Generally, if missingness depends only on observed variables, even on observed outcome, methods like multiple imputation under the assumption that the data is missing at random (MAR) can provide reasonably good estimates (Schafer, 1997). However, if the missingness of covariates also depends on partially observed compliance status, multiple imputation methods based on MAR assumptions may fail to provide valid inference. In this chapter, we will provide a model which allows for missingness to depend on partially observed compliance status and we use the

EM algorithm to obtain the MLE estimates. We also provide a sensitivity analysis which allows for missingness to depend on further unobserved confounders besides compliance status.

Many other observational studies face similar issues of unmeasured confounding and missing data as ours, and the methods we develop in this paper may be useful for them. For example, for studying the comparative effectiveness of two types of drugs, data collected as part of routine health care practice is often used. Such data may not contain measurements of important prognostic variables that guide treatment decisions such as lab values (e.g., cholesterol), clinical variables (e.g., weight, blood pressure), aspects of lifestyle (e.g., smoking status, eating habits) and measures of cognitive and physical functioning (Walker, 1996; Brookhart and Schneeweiss, 2007). To control for such unmeasured confounders, instrumental variable methods have been used, for example, the prescribing preference of a patient's physician for one type of drug vs. the other has been used as an IV (Korn and Baumrind, 1998; Brookhart et al., 2006). For prescribing preference to be a valid IV, it is often necessary to condition on patient characteristics that differ between different physicians to account for the possibility that certain physicians tend to see sicker patients and these physicians may be more likely to prefer one type of drug than physicians who tend to see less sick patients (Korn and Baumrind, 1998). However, there is often missing data on some of these patient characteristics we would like to condition on, in particular because the data is collected as part of routine practice

rather than as part of a research study. For example, even if lab tests are always measured when a lab test is actually administered, since doctors will only order a lab test for some patients, there will be missing data. The missingness of lab values might be related to the treatment decision and outcome, and be nonignorable. For example, the decision to order a lab test is likely related to patient symptoms and/or disease severity, and we would expect that the probability of a lab test being ordered depends on what the value of the test would be, if measured, with unusual values being more likely to be measured (Roy and Hennessy, 2011). Thus, comparative effectiveness studies of drugs may need to consider instrumental variable methods with nonignorable missing covariates as in our study.

3.2 Notation and Assumptions

3.2.1 Notation

We use the potential outcome approach to define causal effects. Let Z_i represent the binary IV of infant i; 1 if excess travel time is less than 10 minutes, which encourages delivery in a high level NICU; 0 if excess travel time is more than 10 minutes, which does not provide encouragement of delivery in a high level NICU. In our data, 56.4% of subjects have excess travel time less than 10 minutes. We use $\boldsymbol{Z}$ to denote the vector of IVs for all infants. Let $D_i(\boldsymbol{z})$ be the potential binary treatment variable that would be observed for subject i under IV assignment $\boldsymbol{z}$. Let

$D_i(\boldsymbol{z})$ be 1 if baby i would be delivered at a high level NICU under the vector of $\boldsymbol{z}$ and 0 if the baby would be delivered at a low level NICU. We also let $Y_i(\boldsymbol{z})$ denote the potential binary outcome, neonatal death indicator, that would be observed for infant i under IV assignment $\boldsymbol{z}$, with $Y_i(\boldsymbol{z})$ being 1 indicating that the newborn would die in the hospital (neonatal death). We use $\boldsymbol{X}_i$ to denote the covariate values for i^{th} subject. The covariates in our study are discrete: infant's gestational weeks, the month of pregnancy that prenatal care started and mother's education, namely 8th grade or less, some high school, high school graduate, some college, college graduate, and more than college. For simplicity, we include the intercept in $\boldsymbol{X}_i$. Finally, we let $R_i^x(\boldsymbol{z})$ be the binary response indicator of covariate x under IV $\boldsymbol{z}$, that is, $R_i^x(\boldsymbol{z}) = 1$ if covariate x would be observed for infant i under IV assignment $\boldsymbol{z}$, and $R_i^x(\boldsymbol{z}) = 0$ if covariate x would be missing. There is a $R_i^x(\boldsymbol{z})$ for each covariate. In the above set of notations, $D_i(\boldsymbol{z}), Y_i(\boldsymbol{z})$ and $R_i^x(\boldsymbol{z})$ are all potential outcomes of an infant. For each infant, depending on the value of $\boldsymbol{z}$, one scenario is factual (observed), the other ones are counter factual (not observed). We use D_i, Y_i and R_i^x to denote observed treatment received, observed death outcome of infant, and the observed response indicator for covariate x.

3.2.2 Assumptions

We assume the following assumptions hold in our study. The first 5 assumptions are the same as Angrist, Imbens and Rubin (1996).

Assumption 1. Stable unit treatment value assumption (SUVTA) , meaning that a subject's potential outcomes cannot be affected by other individuals' status.

SUVTA allows us to write $D_i(\boldsymbol{z})$ as $D_i(z_i)$, $Y_i(\boldsymbol{z})$ as $Y_i(z_i)$ and $R_i^x(\boldsymbol{z})$=$R_i^x(z_i)$. This assumption is plausibly satisfied for our data since whether a mother delivers at a high level NICU and her baby's outcome is unlikely to be affected by other mothers' choice of living near to a high level NICU or not.

Based on subjects' compliance behavior, we can partition the population into four groups:

$$U_i = \begin{cases} n, & \text{if } D_i(1) = D_i(0) = 0 \\ c, & \text{if } D_i(1) = 1,\ D_i(0) = 0 \\ a, & \text{if } D_i(1) = D_i(0) = 1 \\ d, & \text{if } D_i(1) = 0,\ D_i(0) = 1 \end{cases} \tag{3.2.1}$$

where n, c, a, and d represent never taker, complier, always taker and defier, respectively. Because $D_i(1)$ and $D_i(0)$ are never observed jointly, the compliance behavior of a subject is unknown. The parameter of interest in our study is the complier average causal effect (CACE), $\mathrm{E}(Y_i(1) - Y_i(0) \mid U_i = c, \boldsymbol{X_i} = \boldsymbol{x})$.

Assumption 2. Nonzero average causal effect of Z on D. The average causal effect of Z on D, $\mathrm{E}[D_i(1) - D_i(0)]$, is not equal to zero.

The excess travel time should affect whether mother delivers at a high level or

low level NICU due to near NICUs being more convenient, thus assumption 2 is plausible.

Assumption 3. Independence of the instrument from unmeasured confounders: conditional on $\boldsymbol{X}$, the random vector (Y(0), Y(1), D(0), D(1)) is independent of Z.

This assumption is plausible for our study because premature delivery is unexpected for women, so people do not choose where to live based on the closeness to high level NICU, especially after controlling for measured socioeconomic variable such as mother's education level.

Assumption 4. Monotonicity: $D(1) \geq D(0)$.

If a mother is willing to travel to deliver at a high level NICU when living 10 or more minutes further to a high level NICU than a low level NICU, she is probably also willing to travel to deliver at a high level NICU when living less than 10 minutes further to a high level NICU than a low level NICU.

Assumption 5. Exclusion restrictions among never-takers and always-takers: $Y_i(1) = Y_i(0)$ if $U_i = n$, and $Y_i(1) = Y_i(0)$ if $U_i = a$.

This means that the IV only affects the outcome through treatment and has no direct effect. In our study, this is plausible because most women have enough time to make it to either the nearest high level or low level NICU so that marginal travel

time should not directly affect outcomes.

Assumption 6. Nonignorable missingness assumption (missingness ignorable conditional on compliance class, outcome and fully observed covariates): suppose the first k covariates of $\boldsymbol{X}$ are fully observed, and the last m-k covariates have missing values, then,

$$P(R_i^{X_{i,j}}(z) \mid Y_i(z), U_i, \boldsymbol{X}_i) = P(R_i^{X_{i,j}}(z) \mid Y_i(z), U_i, X_{i,1}, ..., X_{i,k}), \forall j = k+1, ..., m.$$

This is saying that the missingness of covariates precare and mother's education depends only on neonatal death information, compliance status of infant, gestational age (fully recorded) as well as the delivery level of NICU. It's a plausible assumption for our data given the discussion in section 3.1.3.

Assumption 7. Exclusion restriction on missing indicator among never-takers and always-takers. $R_i^{X_{i,j}}(1) = R_i^{X_{i,j}}(0)$ if $U_i = n$, and $R_i^{X_{i,j}}(1) = R_i^{X_{i,j}}(0)$ if $U_i = a$.

These are analogous assumptions to Frangakis and Rubin (1999). This means that the IV has no effect on missingness for never takers and always takers. We think this assumption is plausible for our data for the following reasons. We think that the missingness of covariates if affected by death and the baby's risk of death and complications as captured by gestational age and compliance class. Since for always takers and never takers, death is not affected by their level of the IV z (this is assumption 5), and additionally the gestational age and compliance class are not affected by the level of the IV, the missingness of covariates for always takers and

never takers should not be affected by the level of the IV.

3.3 Model and Estimation

We use a general location model (Olkin and Tate, 1961; Little and Rubin, 2002) for a mixture of continuous and categorical covariate variables, which could be easily adjusted for cases where covariates variables are all categorical or all continuous. We consider logistic models for (i) treatment assignment given covariates, (ii) outcome in each compliance class/ treatment assignment combination given covariates, and (iii) missingness in each compliance class/treatment assignment combination given covariates, and we use a multinomial logistic model for compliance class.

Model for covariate : Suppose that in the m covariates, the first p are categorical and the remaining m-p are continuous. We assign probability $W_{x_1,...,x_p}$ to each combination of possible values of those p categorical covariates variables, where $W_{x_1,...,x_p}$ are unknown parameters, and sum up to 1.

- $(X_{i,1}, ..., X_{i,p})$ are iid distributed with

$$P((X_{i,1}, ..., X_{i,p}) = (x_1, ..., x_p)) = W_{x_1,...,x_p}, \quad \text{where } \sum W_{x_1,...,x_p} = 1 \quad (3.3.1)$$

- Conditional on $(X_{i,1}, ..., X_{i,p}) = (x_1, ..., x_p)$, we assume that the continuous covariates random variables $(X_{i,p+1}, ..., X_{i,m})$ is multivariate normal with unknown mean vector $\boldsymbol{\mu}_{x_1,...,x_p}$ which may depend on the values of $(x_1, ..., x_p)$,

and with unknown common positive definite covariance matrix Σ in order to reduce the number of parameters.

$$X_{i,p+1},...,X_{i,m} \mid (X_{i,1},...,X_{i,p}) = (x_1,...,x_p) \sim_{iid} \mathrm{N}_{m-p}(\boldsymbol{\mu}_{x_1,...,x_p}, \Sigma) \quad (3.3.2)$$

Model for IV :

$$P(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}) = \frac{\exp(\alpha^T \boldsymbol{x})}{1 + \exp(\alpha^T \boldsymbol{x})} \quad (3.3.3)$$

Model for compliance class :

$$P(U_i = n \mid \boldsymbol{X}_i = \boldsymbol{x}) = \frac{1}{1 + \exp(\delta_a^T \boldsymbol{x}) + \exp(\delta_c^T \boldsymbol{x})} \quad (3.3.4)$$

$$P(U_i = c \mid \boldsymbol{X}_i = \boldsymbol{x}) = \frac{\exp(\delta_c^T \boldsymbol{x})}{1 + \exp(\delta_a^T \boldsymbol{x}) + \exp(\delta_c^T \boldsymbol{x})} \quad (3.3.5)$$

$$P(U_i = a \mid \boldsymbol{X}_i = \boldsymbol{x}) = \frac{\exp(\delta_a^T \boldsymbol{x})}{1 + \exp(\delta_a^T \boldsymbol{x}) + \exp(\delta_c^T \boldsymbol{x})} \quad (3.3.6)$$

Model for outcome :

$$P(Y_i(z) = 1 \mid U_i = u, \boldsymbol{X}_i = \boldsymbol{x}) = \frac{\exp(\beta_{uz}^T \boldsymbol{x})}{1 + \exp(\beta_{uz}^T \boldsymbol{x})} \quad (3.3.7)$$

According to assumption 4, $\beta_{a0} = \beta_{a1}$, and $\beta_{n0} = \beta_{n1}$. The quantity of interest is the average treatment effect for compliers of each covariate level, which is estimated by $\mathrm{E}(Y(1) - Y(0) \mid U = c, \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1+\exp(\beta_{c0}^T \boldsymbol{x})} - \frac{1}{1+\exp(\beta_{c1}^T \boldsymbol{x})}$.

Model for missingness indicators :

$$P(R_i^{X_{i,j}}(z) = 1 \mid Y_i(z) = y, U_i = u, \boldsymbol{X}_{i,1,...,k} = \boldsymbol{x}_{1,...,k}) = \frac{\exp(\theta_{j,u}^T \boldsymbol{x}_{1,...,k} + \gamma_{j,u} I_{y=1} + \eta_{j,u} I_{z=1})}{1 + \exp(\theta_{j,u}^T \boldsymbol{x}_{1,...,k} + \gamma_{j,u} I_{y=1} + \eta_{j,u} I_{z=1})} \quad (3.3.8)$$

where $j = k+1,...,m$. Based on assumption 7, $\eta_{j,a} = \eta_{j,n} = 0, \forall j = k+1,...,m$

Under the model (3.3.1)-(3.3.8), we seek to maximize the likelihood of the joint distribution of X, Z, U, Y, R. If we know the compliance classes and the missing covariates for each subject, we can get the MLE of parameters involved in those models easily. Based on this idea, we are going to use EM algorithm.

3.3.1 EM algorithm

For simplicity, we are going to present the EM algorithm for the case where all the covariates are categorical and that there are 4 covariates (including intercept) with only the first two completely observed, which is the case of our data. The EM algorithm can be easily extended to other scenarios. The first covariate is the intercept, and we further assume that the other three covariates are ordered categorical with q_2, q_3, q_4 levels respectively. For a nominal categorical variable, we can use indicator functions for each category, which the following algorithm could be easily adjusted for.

Let $N_{r_3,r_4,x_2,x_3,x_4,u,z,y}$ be the number of cases where $R_i^{X_3} = r_3, R_i^{X_4} = r_4, X_{i,2} = x_2, X_{i,3} = x_3, X_{i,4} = x_4, Z_i = z, Y_i = y, U_i = u$. Notice that $X_{i,1} = 1, \forall i$. Those numbers are only partially observed, however, if they are known, the complete data log likelihood is,

$$\begin{aligned} l_c \quad = \quad & \sum_{r_3,r_4,x_2,x_3,x_4,u,z,y} N_{r_3,r_4,x_2,x_3,x_4,u,z,y} \cdot (\log(W_{x_2,x_3,x_4}) + \log(P(Z_i = z \mid X_i = (1, x_2, x_3, x_4))) \\ & + \log(P(U_i = u \mid X_i = (1, x_2, x_3, x_4))) + \log(P(Y_i = y \mid Z_i = z, U_i = u, X_i = (1, x_2, x_3, x_4))) \\ & + \log(P(R_i^{X_{i,3}} = r_3 \mid Z_i = z, Y_i = y, U_i = u, X_{i,2} = x_2)) \\ & + \log(P(R_i^{X_{i,4}} = r_4 \mid Z_i = z, Y_i = y, U_i = u, X_{i,2} = x_2))) \end{aligned}$$

Once we know N, the MLE estimates of the logistic models in (3.3.3)-(3.3.8) are standard, and the MLE for $W_{x_2,x_3,x_4} \propto N_{,,x_2,x_3,x_4,,,}$, where $N_{,,x_2,x_3,x_4,,,}$ is defined to be $\sum_{r_3,r_4,u,z,y} N_{r_3,r_4,x_2,x_3,x_4,u,z,y}$.

In the E-step, conditional on observed data and parameters' estimates obtained through the previous step, we can get the expected values for $N_{r_3,r_4,x_2,x_3,x_4,u,z,y}$.

From the observed data, we can get the following counts:

1. $NN_{x_2,x_3,x_4,d,z,y}$ which denotes the number of cases that $X_{i,3}, X_{i,4}$ are both observed and that $X_{i,2} = x_2, X_{i,3} = x_3, X_{i,4} = x_4, D_i = d, Z_i = z, Y_i = y$

2. $N3_{x_2,x_4,d,z,y}$ which denotes the number of cases that only $X_{i,3}$ are unobserved and that $X_{i,2} = x_2, X_{i,4} = x_4, D_i = d, Z_i = z, Y_i = y$

3. $N4_{x_2,x_3,d,z,y}$ which denotes the number of cases that only $X_{i,4}$ are unobserved and that $X_{i,2} = x_2, X_{i,3} = x_3, D_i = d, Z_i = z, Y_i = y$

4. $NB_{x_2,d,z,y}$ which denotes the number of cases that $X_{i,3}, X_{i,4}$ are both missing and that $X_{i,2} = x_2, D_i = d, Z_i = z, Y_i = y$

Further, let $P_{r_3,r_4,x_2,x_3,x_4,u,z,y}$ be the probability of a subject having case where $R_i^{X_3} = r_3, R_i^{X_4} = r_4, X_{i,2} = x_2, X_{i,3} = x_3, X_{i,4} = x_4, Z_i = z, Y_i = y, U_i = u$ which are calculated based on models (3.3.1)-(3.3.8). Then we can get the expected values for each $N_{r_3,r_4,x_2,x_3,x_4,u,z,y}$, for example,

$$\mathrm{E}N_{1,1,x_2,x_3,x_4,a,1,y} = NN_{x_2,x_3,x_4,1,1,y} \frac{P_{1,1,x_2,x_3,x_4,a,1,y}}{P_{1,1,x_2,x_3,x_4,a,1,y} + P_{1,1,x_2,x_3,x_4,c,1,y}}$$

To save space, all the formulas to update each $N_{r_3,r_4,x_2,x_3,x_4,u,z,y}$ are given in appendix. By iteratively finding the E step estimate of N and maximizing the expected value of the complete data log likelihood in the M step until the algorithm converges, we obtain estimates of the parameters in models (3.3.1)-(3.3.8).

3.4 Simulation

In this section, we conduct simulation studies to estimate the complier average causal effect in the simplest context where there is only one covariate, the values of which could only be 0,1. We consider the following three scenarios under assumptions 1-7: 1) covariate is missing completely at random; 2) covariate is missing at random, meaning that the missingness does not depend upon the unobserved data, for example, does not depend on latent compliance status ; 3) missing mechanism for covariate is nonignorable: the missingness of covariate can depend on not only the observed outcome Y, treatment assignment Z, but also latent compliance status U.

In each scenario, we are going to apply the following three estimation methods and compare their results: 1) complete case analysis, which provides unbiased estimates when the missing mechanism of the data is missing completely at random. 2) the estimates using multiple imputation by chained equations (conducted by MICE, see Van Buuren and Groothuis-Oudshoorn, 2011) which gives valid estimates when data are missing at random. 3) Our method, which is designed to deal

with nonignorable missingness of covariates.

In the single covariate case, the models described in section 3 can be represented simply by the following set of parameters: W_u, which is $P(U_i = u)$; M_u, which is $P(X_i = 1 \mid U_i = u)$; ξ_x, which represents $P(Z_i = 1 \mid X_i = x)$; θ_{zux}, which denotes $P(Y_i(z) = 1 \mid U_i = u, X_i = x)$ and ρ_{yzu}, which are parameters for missingness indicators $P(R_i(z) = 1 \mid Y_i = y, U_i = u)$, where $R_i = 0$ if covariate for i^{th} subject is missing. $\theta_{1c1} - \theta_{0c1}$ and $\theta_{1c0} - \theta_{0c0}$ are corresponding compliers' average causal effect for subjects with X being 1 and 0 respectively.

In all three scenarios, the parameters other than the ones in the missingness model are arbitrarily chosen and fixed as follows:

$$W_n = 0.2,\ W_a = 0.375,\ M_n = 0.5,\ M_a = 0.25,\ M_c = 0.8,\ \xi_1 = 0.4,\ \xi_0 = 0.6$$

$$\theta_{1n1} = 0.5,\ \theta_{1n0} = 0.3,\ \theta_{0a1} = 0.8,\ \theta_{0a0} = 0.7$$

$$\theta_{1c1} = 0.7,\ \theta_{1c0} = 0.45,\ \theta_{0c1} = 0.45,\ \theta_{0c0} = 0.3$$

The missingness parameters in each scenario are described below, the values for $\rho's$ are chosen to generate 12% missingness for covariate (the same missing rate as in the NICU study), and satisfy the exclusion restriction for missing indicator, which implies that $\rho_{y0a} = \rho_{y1a},$ and $\rho_{y0n} = \rho_{y1n}$. In the first case, the missingness parameters $\rho's$ are the same for all possible outcomes, IV levels as well as compliance classes, thus the covariate is missing completely at random; in the second case, the missing rates are different for different outcomes and IV levels, however won't be

affected by partially observed compliance status, so that the missingness won't depend on unobserved data, which is a case of missing at random; in the last case, besides outcome and IV, the compliance status also plays a role in deciding the probability of missingness, and the values of $\rho's$ are chosen so that even the largest effect of compliance status on missingness is still moderate ($\rho_{11a} - \rho_{11n} = 0.25$) and realistic.

1. Missing Completely at Random

$$\rho_{11n} = \rho_{01n} = \rho_{10a} = \rho_{00a} = \rho_{11c} = \rho_{01c} = \rho_{00c} = \rho_{10c} = 0.88$$

2. Missing at Random

$$\rho_{11n} = \rho_{10n} = \rho_{10c} = 0.88,\ \rho_{10a} = \rho_{11a} = \rho_{11c} = 0.78$$

$$\rho_{01n} = \rho_{00n} = \rho_{00c} = 0.94,\ \rho_{00a} = \rho_{01a} = \rho_{01c} = 0.97$$

3. Nonignorable Missingness

$$\rho_{11n} = \rho_{10n} = 0.75,\ \rho_{01n} = \rho_{00n} = 0.8,\ \rho_{10a} = \rho_{11a} = 1$$

$$\rho_{00a} = \rho_{01a} = 0.95,\ \rho_{11c} = 0.8,\ \rho_{01c} = 0.9,\ \rho_{00c} = 0.83,\ \rho_{10c} = 0.97$$

We simulated 500 data sets for each scenario described above with each simulated dataset containing 5000 subjects. Under the above setup, the CACE for subjects with covariate being 1 is 0.25, whereas the CACE for subjects with covariate 0 is 0.15. Table 3.1 shows the means and standard deviations for the estimates of CACE

across 500 simulated datasets using the EM algorithm based on our nonignorable missingness assumption, the complete-case estimates and multiple imputation estimates using MICE for each missingness mechanism. The corresponding bias in percentage is given in parentheses.

Table 3.1: Simulation Results under MCAR, MAR and Nonignorable Missing Mechanism

MCAR	CACE	EM(NI)		Complete-Case		MICE	
		Mean	SD	Mean	SD	Mean	SD
	$\theta_{1n1} - \theta_{0n1}$=0.250	0.250 (0.00%)	0.027	0.249 (0.40%)	0.028	0.248 (0.80%)	0.028
	$\theta_{1n0} - \theta_{0n0}$=0.150	0.149 (0.67%)	0.095	0.148 (1.33%)	0.096	0.154 (2.67%)	0.095
MAR	CACE	EM(NI)		Complete-Case		MICE	
		Mean	SD	Mean	SD	Mean	SD
	$\theta_{1n1} - \theta_{0n1}$=0.250	0.250 (0.00%)	0.027	0.221 (11.60%)	0.029	0.246 (1.60%)	0.028
	$\theta_{1n0} - \theta_{0n0}$=0.150	0.147 (2.00%)	0.097	0.113 (24.67%)	0.096	0.160 (6.67%)	0.097
Nonignorable	CACE	EM(NI)		Complete-Case		MICE	
		Mean	SD	Mean	SD	Mean	SD
	$\theta_{1n1} - \theta_{0n1}$=0.250	0.250(0.00%)	0.027	0.188 (24.80%)	0.029	0.234(6.40%)	0.029
	$\theta_{1n0} - \theta_{0n0}$=0.150	0.148(1.33%)	0.093	0.089(40.60%)	0.096	0.221(47.33%)	0.084

From Table 3.1 we see that when data is missing completely at random, all three methods provide unbiased estimates. In the second scenario when the missingness depends on observed data, we can no longer obtain unbiased estimates from complete-case analysis, whereas both our EM algorithm for nonignorable missing-

ness and MICE designed for data missing at random still provide reasonable estimates as we expected. However, when the missingness of covariates depends not only on the observed outcome, but also on the partially observed compliance status, simply using the complete cases or assuming missing at random to impute missing covariates based on the observed data gives us biased estimates of CACE. The complete-case analysis provides biased estimates due to the fact that it is actually estimating $\mathrm{E}(Y_i(1) - Y_i(0) \mid U_i = c, R_i = 1)$, which is generally different from $\mathrm{E}(Y_i(1) - Y_i(0) \mid U_i = c)$ when the data is not missing completely at random. Imputation based on missingness at random is actually imputing X as if the missing mechanisms for compliers and always takers assigned to treatment are the same, and that for compliers and never takers assigned to control are the same. When this is not the case, the imputation estimates are biased.

From our simulation study, we can see that even if the missingness rate of a covariate is low (12%), and the compliance class has only a moderate effect on the missingness, it's still important and necessary to model the effect of compliance class on missingness in the analysis, otherwise, the results could be significantly biased.

3.5 Application to NICU study

The data describes 189,991 babies born prematurely in Pennsylvania between 1995-2005. These premature babies are the ones whose gestational ages are between 23

and 37 weeks. The outcome variable we are interested in is neonatal death of babies, which refers to death during the initial birth hospitalization; we use Y_i to represent the outcome of i^{th} baby in the data set, with Y_i being 1 indicating the death of baby i. We view infants that are delivered in a high level NICU as the treatment group, whereas the ones that are delivered in a low level NICU are the control group. Let D_i equal 1 if the i^{th} baby is delivered in a high level NICU, 0 if in a low level NICU. The instrumental variable we consider is whether or not the mother's excess travel time that a mother lives from the nearest high level NICU compared to the nearest low level NICU is less than or equal to 10 minutes. As we discussed in section 3.2.2, mother's excess time is a plausible IV in our study which satisfies the IV assumptions 1-7 in section 3.2.2. We use Z_i to denote the IV value for the i^{th} baby, with Z_i being 1 indicating that the excess travel time is less than 10 minutes. The measured confounders $\boldsymbol{X}_i$ for baby i are baby's gestational age, the month of pregnancy that prenatal care started and mother's education. We also include an intercept in $\boldsymbol{X}_i$.

In this data set, all variables mentioned above are fully observed except the month of pregnancy that prenatal care started and mother's education level. The missing rates for those two covariates are 10.3% and 2.3% respectively. We did Chi-Square tests of independence to test whether the missingness of those two covariates depends on outcome Y. The p-values are both below 10^{-15}, strong evidence that missingness depends on the outcome. We also did logistic regression to test whether

the missing indicators also depend on the observed risk characteristic of gestational age given the outcome of neonatal death. The results show that gestational age has a significant negative association with the missingness of those two covariates even conditional on outcome (p-values are both below 10^{-15}). Since we have strong evidence that the missingness depends on observed risk characteristics, we believe that the missingness should also depend on unobserved risk characteristics which are reflected in compliance status.

Table 3.2 describes the estimated proportions of each compliance class - always takers, compliers and never takers - for some typical combinations of covariates. There is a clear trend that as the gestational ages gets larger, the proportion of always takers gets smaller, and the proportions of the other two compliance classes gets larger. A reasonable explanation for this phenomenon is that the gestational age is a strong predictor for the risk of complications as well as death– the smaller the baby is, the higher risk the baby and mother have. For babies or mothers at higher risk of complications or death, doctors are more likely to encourage them to go to a high level NICU no matter the mother lives near one or not, i.e. those mothers are more likely to be always takers. Notice that from the fit of our model, there is a substantial proportion of never takers. Although it may be surprising that people would choose to bypass a high level NICU for a low level NICU (i.e., be a never taker). Choice of hospital is driven by a number of factors, including where a patient's physician practices; the general view of the hospital by a specific

community of patients; and what family or friends believe about a hospital. There are families who choose to deliver at smaller hospitals regardless of where they live, and their illness severity. This may be because some families are suspicious of academic hospitals, which make up the majority of high level NICUs, and would rather travel to deliver at a community hospital even if the hospital has fewer resources to care for them.

Table 3.3 shows the estimates of parameters in outcome model for compliers, which are the parameters to estimate the CACE, $\mathrm{E}(Y(1) - Y(0) \mid U = c, X = x) = \frac{1}{1+\exp(\beta_{c0}^T x)} - \frac{1}{1+\exp(\beta_{c1}^T x)}$. The standard errors for the corresponding parameters are provided in parentheses; the standard errors are estimated through bootstrap using 1000 re-samples. From the estimates for the outcome model, we see that larger gestational age and higher mother's education level are related to low death rate, and that for the mothers who started prenatal care late, the baby is at more risk of death.

Table 3.4 shows the estimated CACE of delivering at high level NICU vs. low level NICU for various combinations of the measured covariates. High level NICUs substantially reduce the probability of death for very premature babies. For example, for an infant of gestational age 24 weeks, whose mother started prenatal care in the second month of pregnancy and has a high school education, being delivered in a high level NICU will reduce the probability of death by 0.296, with a 95% confidence interval of -0.429 to -0.137. The effect of high level NICUs is less for less

Table 3.2: Percentages of Always Takers, Compliers and Never Takers in %

Gestational age	Precare	Mother's education	Percentage of always takers		Percentage of compliers		Percentage of never takers	
			Estimate	95%CI	Estimate	95%CI	Estimate	95%CI
24	2	High School	87.2	[86.3, 88.0]	4.4	[3.6, 5.1]	8.4	[7.8, 9.0]
24	4	High School	87.4	[86.5, 88.3]	4.8	[3.9, 5.6]	7.8	[7.3, 8.5]
24	2	College	92.0	[91.5, 92.6]	2.7	[2.2, 3.2]	5.3	[4.8, 5.7]
24	4	College	92.2	[91.5, 92.7]	2.9	[2.4, 3.5]	4.9	[4.5, 5.3]
30	2	High School	59.4	[57.9, 60.2]	20.0	[18.4, 21.4]	21.0	[20.2, 21.8]
30	4	High School	58.9	[57.7, 60.3]	21.6	[19.9, 23.1]	19.5	[18.8, 20.3]
30	2	College	71.1	[70.1, 72.0]	14.0	[12.8, 15.1]	15.0	[14.3, 15.6]
30	4	College	70.9	[69.8, 72.1]	15.1	[13.8, 16.4]	13.9	[13.3, 14.6]
37	2	High School	17.4	[17.1, 17.7]	54.2	[53.6, 54.9]	28.4	[27.9, 28.8]
37	4	High School	17.0	[16.5, 17.4]	57.3	[56.6, 58.0]	25.8	[25.2, 26.3]
37	2	College	26.5	[26.0, 26.9]	48.0	[47.2, 48.8]	25.6	[25.1, 26.0]
37	4	College	25.9	[25.2, 26.6]	50.8	[49.9, 51.8]	23.3	[22.7, 23.9]

premature babies; when the baby's gestational age is about 37 weeks, the high level NICU has almost no effect on mortality. This is plausible since a 37 baby is almost mature and is at less risk, and consequently, the type of delivery NICU may not matter much.

Using our method, the estimated CACE weighted by the probability of each combination of the measured covariates is -0.010, with a 95% confidence interval [-0.014, -0.006]; and the estimated CACE weighted by the number of compliers in each combination of the measured covariates is -0.002, with a 95% confidence interval [-0.004, -0.001]. Thus, our analysis shows that high level NICU significantly reduce the probability of death for premature babies.

We compare our analysis to several "baseline" methods commonly used to analyze observational studies that are not designed to allow for unmeasured confounders or nonignorable missingness. The first method we consider is an unadjusted analysis using the observed rates of neonatal death in high level NICUs and low level NICUs to estimate $E(Y \mid D = 1) - E(Y \mid D = 0)$. The estimate is 0.01 with a 95% confidence interval [0.009, 0.011], which shows that high level NICU is associated with a higher probability of death. The second method we consider is a logistic regression model of neonatal death indicator Y on treatment D as well as the measured confounders to get an estimate $\frac{1}{N}\sum_{i=1}^{N}[\hat{E}(Y \mid D = 1, \boldsymbol{X}) - \hat{E}(Y \mid D = 0, \boldsymbol{X})]$ to adjust for covariates. We use mice under MAR assumption to impute the missing values in the data. This adjusted estimate is 0.000, with a 95% confidence interval

[-0.001, 0.001], which provides no evidence of an association between level of NICU and chance of neonatal death. The third method we consider is subclassification on the propensity score following Rosenbaum and Rubin (1984). As suggested in Rosenbaum and Rubin (1984), we divided babies into five subclasses based on the propensity score, and obtain the average treatment effect by weighting each subpopulation's average treatment effect by the proportion of each subclass. This adjusted analysis shows that the high level NICU increases the probability of neonatal death by 0.002, with a 95% confidence interval [0.001, 0.003]. The conclusions of all the three baseline methods contradicts with the result of our method, which found evidence that delivery at a high level NICU increases a premature baby's chance of survival. Unlike the three baseline methods, our method allows for unmeasured confounders and a certain type of nonignorable missingness of covariates.

Table 3.3: Estimates of Outcome Model for Compliers

Parameters	Intercept	Gestational age	Precare	Mother's education
β_{c1}	1.400 (1.617)	-0.153 (0.043)	0.091 (0.063)	-0.522 (0.118)
β_{c0}	9.450 (1.274)	-0.395 (0.042)	0.144 (0.055)	-0.315 (0.113)

3.6 Sensitivity Analysis

In this section, we will assess the sensitivity of our causal conclusions to an unmeasured patient risk characteristic relevant to both the outcome of death and missing-

Table 3.4: CACE with Different Covariate Values

Gestational age	Precare	Mother's education	CACE	95% Confidence interval
24	2	High School	-0.296	[-0.429, -0.137]
24	4	High School	-0.343	[-0.490, -0.162]
24	2	College	-0.192	[-0.349, -0.064]
24	4	College	-0.230	[-0.421, -0.077]
30	2	High School	-0.032	[-0.043, -0.017]
30	4	High School	-0.040	[-0.056, -0.023]
30	2	College	-0.019	[-0.033, -0.008]
30	4	College	-0.024	[-0.043, -0.009]
37	2	High School	0.001	[-0.001, 0.002]
37	4	High School	0.001	[-0.001, 0.002]
37	2	College	0.000	[-0.001, 0.001]
37	4	College	0.000	[-0.002, 0.001]

ness of covariates, for example, results of tests like fetal heart tracing or doctor's knowledge about mother's severity of condition. Following the idea of Rosenbaum and Rubin (1983), we assume that there is an unobserved binary covariate Q which represents the risk not explained by compliance status and gestational age, and that is independent of the observed covariates, the compliance status and the instrument. We want to know after accounting for such an unmeasured covariate, is there still evidence that the high level NICU reduces the probability of death for babies of small gestational age.

The adjusted model is as follows:

$$P(Q = 1) = \pi,$$

the parameter π gives the probability that the unobserved binary risk variable being 1. We assume that the unobserved binary risk variable Q is independent of IV Z, compliance class U and covariates $\boldsymbol{X}$, thus, the models (3.3.1)-(3.3.6) remain the same in our sensitivity analysis. The model of outcome controlling also for Q is:

$$P(Y_i(z) = 1 \mid U_i = u, \boldsymbol{X}_i = \boldsymbol{x}, Q_i = q) = \frac{\exp\left(\beta_{uz}^T \boldsymbol{x} + \xi_{uz} q\right)}{1 + \exp\left(\beta_{uz}^T \boldsymbol{x} + \xi_{uz} q\right)}$$

Again, according to assumption 4, $\beta_{a0} = \beta_{a1}, \beta_{n0} = \beta_{n1}, \xi_{a0} = \xi_{a1}, \text{and } \xi_{n0} = \xi_{n1}$. ξ_{uz} gives the log odds ratio for Y in two subpopulations q=1 and q=0. Finally, the model for missing indicators of covariate j controlling further for Q is:

$$\begin{aligned} & P(R_i^{X_{i,j}}(z) = 1 \mid Y_i(z) = y, U_i = u, \boldsymbol{X}_{i,1,\ldots,k} = \boldsymbol{x}_{1,\ldots,k}, Q_i = q) \\ = & \frac{\exp\left(\theta_{j,u}^T \boldsymbol{x}_{1,\ldots,k} + \gamma_{j,u} I_{y=1} + \eta_{j,u} I_{z=1} + \kappa_{j,u} q\right)}{1 + \exp\left(\theta_{j,u}^T \boldsymbol{x}_{1,\ldots,k} + \gamma_{j,u} I_{y=1} + \eta_{j,u} I_{z=1} + \kappa_{j,u} q\right)} \end{aligned}$$

where $j = k+1, ..., m$. Based on assumption 6, $\eta_{j,a} = \eta_{j,n} = 0, \forall j = k+1, ..., m$. $\kappa_{j,u}$ gives the log odds ratio for R in two subpopulations q=1 and q=0.

For fixed sensitivity parameters $\pi, \xi_{uz}, \kappa_{j,u}$, there exist unique MLEs of the remaining parameters. Our EM algorithm for the original model could be easily extended to obtain those estimates. The average treatment effect for compliers of each covariate level is estimated by $\mathrm{E}(Y(1) - Y(0) \mid U = c, \boldsymbol{X} = \boldsymbol{x}) = \pi \cdot \left(\frac{1}{1+\exp\left(\beta_{c0}^T \boldsymbol{x} + \xi_{c0}\right)} - \frac{1}{1+\exp\left(\beta_{c1}^T \boldsymbol{x} + \xi_{c1}\right)}\right) + (1 - \pi) \cdot \left(\frac{1}{1+\exp\left(\beta_{c0}^T \boldsymbol{x}\right)} - \frac{1}{1+\exp\left(\beta_{c1}^T \boldsymbol{x}\right)}\right)$.

In order to limit the size of the sensitivity analysis, (κ, ξ) is assumed in the sensitivity analysis to be the same across all subclasses defined by IV, compliance class and covariates. And also as in Table 3.4, we estimated CACE for some typical combinations of the measured confounders under each assignment of (π, κ, ξ).

Table 3.5 presents part of the sensitivity analysis results, showing how the unobserved binary covariate Q affects the CACE for patients with prenatal care starting at second month of pregnancy, mother's education being high-school and babies' gestational age being 24 weeks, 30 weeks and 37 weeks respectively. From Table 3.5, we observe that when the odds ratios are doubled, the estimated CACEs do not change much in each assignment of sensitivity parameters; and when the odds ratios are tripled, the estimated CACEs vary more. It's time consuming to conduct bootstrap for each combination of sensitivity parameters to obtain the 95% confidence interval for each scenario, however, due to the fact that we are using the same dataset in outcome analysis in section 3.5 and also in our sensitivity analysis, it is reasonable to assume that the width of the confidence intervals would be similar to the ones shown in Table 3.4 for each scenario. Specifically, if the point estimate and the confidence interval for a parameter in Table 3.4 is a and [b, c] respectively and the point estimate for a corresponding parameter in the sensitivity analysis tables is d, then we estimate the confidence interval for the parameter in the sensitivity analysis to be [d-(a-b), d+(c-a)]. For example, in the first case in Table 3.5, where the gestational age is 24, precare is 2, and mother's education

level is high school, if 10% of patients' unobserved risk covariate is 1, and the unobserved covariate doubles both odds ratios for Y and missingness indicators R, the estimated CACE is -0.289, with approximate 95% confidence interval [–0.422, -0.130]. Consequently, the unobserved covariate Q, would have to more than triple the odds in both the outcome and missing indicator models, before altering the conclusion obtained in section 3.3.5 that high level NICUs reduce the probability of death in babies of small gestational age. To provide some idea about how large an effect an unobserved covariate would have to have to change our conclusions, we compare the effect to that of the observed covariate gestational age, which is a strong predictor for death and risk of complications. According to the fit of our model (see Table 3.3), if gestational age is changed by 2 weeks, then the odds ratios for the outcome death would be altered by a factor of 2.2 and the odds ratios for the response would be altered by a factor of 1.6. Thus based on our sensitivity analysis results, an unobserved covariate with the same effect as changing gestational age by 2 weeks would not change our conclusion that high level NICUs reduce the probability of death in babies of small gestational age. We conclude that even if some confounders, for instance, results of tests like fetal heart tracing, doctor's knowledge about mother's severity of condition, are unmeasured and affect both the outcome and missingness of covariates, they would not change our conclusions unless they had very large effects.

3.7 Summary

We proposed a series of models to estimate the causal effect of a treatment using an instrumental variable when the missingness of covariates may depend on the fully observed outcome, fully observed covariates, IV as well as the partially observed compliance behavior. Simulation studies show that under our nonignorable missingness assumption where the missingness depend on partially observed compliance class, even if the missing rate of covariate is low (12%), and the effect of compliance class on the missingness is only moderate, the commonly used estimation methods, complete case analysis and multiple imputation by chained equations assuming MAR, could provide substantially biased estimates; in contrast, our proposed method, which is designed to deal with nonignorable missingness of covariate, provides unbiased results.

In this chapter, we have developed a maximum likelihood method for instrumental variable estimation with nonignorable missingness of covariates. Further research could consider a Bayesian version of our model which would enable carrying our multiple imputation based on our model.

We applied our method to an observational study of neonatal care that aims to estimate the delivery effect on mortality of premature babies being delivered in a high level NICU vs. a low level NICU. We found that high level NICUs substantially reduce the death risk for babies with small gestational age, which implies that high level NICUs are truly providing considerably better care for babies

with small gestational age. Therefore, it is valuable to invest resources to strengthen the perinatal regionalization system for those babies. For babies that are almost mature, strengthening the perinatal regionalization system should probably not be a priority.

The methods we develop in this chapter may be useful for many other observational studies facing unmeasured confounders as well as nonignorable missing data like ours. One example we described in the introduction is comparative effectiveness studies where it is a concern that the missingness of important lab values might be related with compliance status. For these settings, our simulation study shows that it is important and necessary to model the effect of compliance status on missingness to get valid estimates.

In this study, we focus on cases which contain missing covariates, and the missingness of covariates is nonignorable. However, in practice, many studies face the issue of not only missing covariates but also missing outcomes. In our nonignorable missingness assumption (Assumption 6), we allow the missingness of covariates to depend on the outcome. If there are also missing outcomes, since the covariates are predictors for the outcome, it is likely that the missingness of the outcome is related to the values of covariates which are unobserved for some subjects. If missingness exists in both the covariates and the outcome, identifiability is a major issue to study since the missingness of the covariates and outcome may depend on each other. Additional assumptions beyond what we have considered are needed

for identifiability. Possible assumptions could be developed based on Peng, Little and Raghunathan (2004) where missingness of outcome is allowed to depend on compliance and fully observed data whereas missingness of covariates is allowed to depend on only the fully observed data but not compliance status.

Table 3.5: Effects of Q on the CACE for patients with prenatal care starting at second month of pregnancy, mother's education being high-school and with gestational age being 24 weeks, 30 weeks and 37 weeks respectively

Gestational age	Effect of Q on Y	Effect of Q on R	$P(Q=1):\pi$		
			0.1	0.5	0.9
24	$\exp(\xi)$=2	$\exp(\kappa)$=2	-0.289	-0.283	-0.290
		$\exp(\kappa)=\frac{1}{2}$	-0.297	-0.296	-0.293
	$\exp(\xi)=\frac{1}{2}$	$\exp(\kappa)$=2	-0.293	-0.296	-0.297
		$\exp(\kappa)=\frac{1}{2}$	-0.290	-0.283	-0.289
	$\exp(\xi)$=3	$\exp(\kappa)$=3	-0.273	-0.228	-0.217
		$\exp(\kappa)=\frac{1}{3}$	-0.296	-0.252	-0.225
	$\exp(\xi)=\frac{1}{3}$	$\exp(\kappa)$=3	-0.298	-0.329	-0.379
		$\exp(\kappa)=\frac{1}{3}$	-0.289	-0.300	-0.352
30	$\exp(\xi)$=2	$\exp(\kappa)$=2	-0.032	-0.031	-0.031
		$\exp(\kappa)=\frac{1}{2}$	-0.032	-0.033	-0.032
	$\exp(\xi)=\frac{1}{2}$	$\exp(\kappa)$=2	-0.032	-0.033	-0.032
		$\exp(\kappa)=\frac{1}{2}$	-0.031	-0.031	-0.032
	$\exp(\xi)$=3	$\exp(\kappa)$=3	-0.029	-0.024	-0.022
		$\exp(\kappa)=\frac{1}{3}$	-0.032	-0.026	-0.022
	$\exp(\xi)=\frac{1}{3}$	$\exp(\kappa)$=3	-0.032	-0.038	-0.046
		$\exp(\kappa)=\frac{1}{3}$	-0.032	-0.035	-0.043
37	$\exp(\xi)$=2	$\exp(\kappa)$=2	0.001	0.001	0.001
		$\exp(\kappa)=\frac{1}{2}$	0.001	0.001	0.001
	$\exp(\xi)=\frac{1}{2}$	$\exp(\kappa)$=2	0.001	0.001	0.001
		$\exp(\kappa)=\frac{1}{2}$	0.001	0.001	0.001
	$\exp(\xi)$=3	$\exp(\kappa)$=3	0.001	0.001	0.001
		$\exp(\kappa)=\frac{1}{3}$	0.001	0.001	0.001
	$\exp(\xi)=\frac{1}{3}$	$\exp(\kappa)$=3	0.001	0.001	0.001
		$\exp(\kappa)=\frac{1}{3}$	0.001	0.001	0.001

Chapter 4

Aporetic Conclusions When Testing the Validity of an Instrumental Variable

4.1 Testing untestable assumptions in causal inference with instrumental variables

4.1.1 What is an instrument? What assumptions underlie its use?

An instrument is a haphazard nudge to accept a treatment where the nudge can affect the outcomes only to the extent that it alters the treatment received. The most

basic example is Holland's (1988) randomized encouragement design, in which people are randomized to one of two groups, and members of one group are encouraged to adopt some health promoting behavior, say quit smoking, but the outcome, say an evaluation of lung tissue, might respond to a reduction in cigarettes consumed but not to encouragement to quit that leaves cigarette consumption unchanged. There are two key elements here. First, in the encouragement experiment, people are picked at random for encouragement — selection does not just look haphazard, it is actually randomized — so the comparison of encouraged and unencouraged groups is equitable, not subject to biases of self-selection. Even in the randomized encouragement design, people who change their behavior, quit smoking, are a self-selected part of the encouraged and possibly unencouraged groups, so a comparison of quitters and others could be very biased: quitters may be more self-disciplined in all areas of their lives and may be more concerned with health promotion. The second element is that encouragement works, affects the outcome, only if it changes behavior, the so-called exclusion restriction. Stated informally in words, the instrumental variable (IV) estimate, the Wald estimate, attributes the entire difference in outcomes between the randomized encouraged and unencouraged groups to the greater change in behavior in the encouraged group, thereby avoiding biases of self-selection. If the encouraged group has a mean outcome that is one unit better than the mean in the unencouraged group, and if half of the encouraged quit while none of the unencouraged quit, then the Wald estimator claims the effects of quitting on

those who quit when encouraged is two units, because encouragement only affected half of those who were encouraged. See Angrist, Imbens and Rubin (1996) for an equivalent formal statement.

So there are two key elements in the randomized encouragement design:

(i) encouragement is randomized,

(ii) encouragement affects only those individuals who change their behavior in response to encouragement, the exclusion restriction.

In the encouragement design, (i) is ensured by the use of randomization, and (ii) seems highly plausible because of what we think we know about the relationships that might exist between advice, behavior and lung tissue. Typical applications of the reasoning involving instruments are less compelling, sometimes much less compelling, because (i) is not ensured by actual random assignment, and (ii) is less firmly grounded in other things we think we know. In particular, (i) is typically rendered somewhat plausible by adjusting for visible differences in measured pre-treatment covariates between encouraged and unencouraged groups, but of course this strategy may fail to control a covariate that was not measured. Typically, the encouraged and unencouraged groups are not formed by random assignment, but rather in a way that appears irrelevant and haphazard, but these appearances may deceive. Typically, the exclusion restriction seems plausible to anyone who cannot imagine a way encouragement could affect the outcome without altering the

treatment, but this may simply reflect inadequate imagining. So it is natural to want to test the assumptions that define an instrument.

Instruments are increasingly used in the study of health outcomes; see, for example, McClellan et al. (1994), Lalani et al. (2010) and Lorch et al. (2012). Outside of randomized clinical trials, the treatment a patient receives may reflect a physician's judgment about the best treatment for this patient or else a patient's preference for a particular treatment. Instruments are used in health outcomes research in the hope of finding circumstances in which attributes of the patient do not decide treatment assignment, and instead something haphazard and irrelevant decides the treatment. For instance, if a patient has a heart attack and lives far from a hospital capable of performing coronary bypass surgery, then the heart attack may be treated without bypass surgery just because of where the patient happens to live. Obviously, geography might appear to be haphazard and irrelevant, might appear to satisfy conditions (i) and (ii), yet these appearances may be incorrect; so, testing (i) and (ii) is important. For several recent discussions of instrumental variables in health outcomes, see Baiocchi et al. (2010), Brookhart and Schneeweiss (2007), Cheng et al. (2011), Swanson and Hernán (2013) and Tan (2006).

4.1.2 Untestable assumptions?

The assumptions required for an instrument are often said to be untestable (e.g., Stock 2002, §4.1). Whether this is true or not depends in part on what one

means by untestable. Assumptions might be said to be untestable if they (A) are premises of a theorem that is the basis for an inference, (B) these premises are not self-evident or implied by other premises that are self-evident, (C) these premises cannot be tested against data from the observable distributions specifically mentioned in the statement of the theorem. This is an internally consistent way to use the word untestable, but it is a manner of speaking at considerable tension with typical scientific practice. Typically in science, each new claim to know something is checked for consistency with the other things we think we know. There is no reason to confine this checking for consistency to the short list of premises of a theorem. This checking may involve logical consistency, but more often the question is whether the new knowledge claim and old knowledge claims could plausibly be describing one and the same world, or whether something has to give. We discuss an example in detail in §4.2. In practical work with instruments, it is quite common to hear people announce that IV assumptions are untestable and then to see them do the sorts of checks that test IV assumptions.

Why are IV assumptions often said to be untestable when people often test them? We suspect there is a reason. A test of IV assumptions may lead neither to rejection of the assumptions nor to acceptance but rather to an aporia.

4.1.3 Aporia: mutually inconsistent but individually plausible claims

The Oxford American Dictionary defines the noun aporia as "an irresolvable internal contradiction ... in a text, argument or theory," with aporetic as the adjective. A collection of propositions, $\varpi_1, \ldots, \varpi_L$ is an aporia if each ϖ_ℓ is plausible on its own but they are jointly inconsistent, that is, $\varpi_1 \wedge \ldots \wedge \varpi_L$ is false or implausible; see Rescher (2009). A special case of aporia occurs in mathematical reasoning in a proof by contradiction, in which one proves $\sim \varpi_L$ by showing that $\varpi_1, \ldots, \varpi_{L-1}$ are certainly true and $\varpi_1, \ldots, \varpi_L$ is aporetic in yielding a contradiction. In contrast, in a typical aporia, in the general case, the identity of the culpable proposition or propositions is unknown. In Plato's early dialogues, Socrates would invalidate the views of his opponents by demonstrating that those views were aporetic; see Vlastos (1994, p. 58).

To recognize that one's beliefs contain an aporia is an advance in understanding, albeit an uncomfortable one. From a false premise, one can logically deduce every conclusion, true or false (because, in elementary propositional logic, $A \Rightarrow B$ is true for all B if A is false). To believe $\varpi_1, \ldots, \varpi_L$ individually but fail to recognize them as aporetic is to risk logically deducing false propositions from beliefs one holds (because one believes $\varpi_1, \ldots, \varpi_L$, can deduce the false proposition $A = \varpi_1 \wedge \ldots \wedge \varpi_L$ from one's beliefs, and can deduce any B from A because A is false). To recognize that one's beliefs $\varpi_1, \ldots, \varpi_L$ are aporetic is to recognize that one harbors

at least one false belief, to be motivated to identify that belief, and to be hesitant in deducing consequences from $\varpi_1, \ldots, \varpi_L$. To recognize an aporia is an advance in understanding, and it is certainly better than believing the component propositions without recognizing their aporetic status.

One can escape an aporia $\varpi_1, \ldots, \varpi_L$ by arbitrarily discarding propositions ϖ_ℓ until the remaining propositions are no longer inconsistent. In this process, there is nothing to ensure that one has discarded false propositions and retained true ones. Rather, one has narrowed the scope of one's beliefs to the point that one is committed to sufficiently few beliefs that one is safe from accusations of inconsistency. For instance, one can avoid an aporia in testing the assumptions of IV by defining those assumptions so narrowly that they become untestable.

4.1.4 Outline: an IV study; a test of IV assumptions; two technical innovations

We are currently using an instrument in a study of the possible effects of delivery by cesarean section of extremely premature infants of 23-24 weeks gestational age. Some background is discussed in §4.2.1 and the IV analysis is presented in §4.2.2-§4.2.4. In §4.2.5, the IV assumptions are tested, resulting in an aporia that is discussed in detail. The two appendices present two technical innovations: a new simpler approach to strengthening an instrument, and a sensitivity analysis for an attributable effect closely related to the Wald estimator. A reader who wishes

to reproduce the analysis reported here will need to consult these appendices. We placed this material in the appendix because we wanted to emphasize the conceptual discussion of aporetic conclusions when testing the assumptions underlying instrumental variables.

4.2 Does delivery by cesarean section improve survival of extremely premature neonates?

4.2.1 Background: Studies of cesarean section without an instrumental variable

We are currently engaged in a study of the possible effects of cesarean section on the survival of very premature babies of 23-24 gestational age. For reasons to be described shortly, we tried to find an instrument for delivery by cesarean section and to check its validity by contrast with other trusted information. Some terminology and background are needed.

The gestational age of a full-term baby is 39 weeks or 9 months. Babies born under 37 weeks gestation are considered premature, with infants born younger having more medical problems, requiring more intensive medical care to survive, and having a higher likelihood of long-term neurodevelopment and medical problems. This issue is most prominent for the infants at the limits of viability, that is, those

infants born at 23 and 24 weeks gestation. Babies born between 23 and 24 weeks of gestational age are very premature and face high risks of death and life-long health problems even with special care. A fetus of 23 and 24 weeks of gestational age that is not born alive is defined as a fetal death, whereas an infant who dies after delivery is designated as a neonatal death. There are clinical indicators around a pregnancy at the limits of viability that give the physician information about the likelihood that an infant will survive first the delivery, and then the initial period of time after delivery.

In clinical epidemiology, the phrase "confounding by indication" is often defined as the bias introduced when patients receive medical treatments based on pretreatment indications that the patient would benefit from the treatment. To the extent that such indications for treatment are incompletely recorded, thus incompletely controlled by adjustments for recorded pretreatment differences, they may lead to bias in elementary analyses that rely on adjustments for confounding factors using recorded pretreatment differences. At gestational age 23-24 weeks, delivery by cesarean section is likely to reflect clinical judgement about the clinical stability and likelihood of survival of the infant and the generally unrecorded preferences of the mother. Both of these factors are likely to be incompletely recorded in most large-scale population datasets.

A major use of instrumental variables in medicine is to break up or otherwise avoid confounding by indication, that is, to find some circumstances in which pa-

tients received a medical treatment for reasons other than that the patient was expected to benefit from treatment. In a randomized trial, patients receive treatments for no reason at all, the flip of a fair coin, and instruments are sought in observational studies to recover as best one can some aspects of the randomized situation.

Existing literature suggests that routine or optional use of cesarean delivery for babies of $\geq$ 30 weeks gestational age is not of benefit to the baby. For instance, Werner et al. (2013) concluded:

> In this preterm cohort, cesarean delivery was not protective against poor outcomes and in fact was associated with increased risk of respiratory distress and low Apgar score compared with vaginal delivery. (page 1195)

More than seventy percent of the preterm cohort in Werner et al. (2013) were $\geq$ 30 weeks gestational age, and more than half were $\geq$ 32 weeks, while less than 6% were less than 26 weeks. Werner et al. (2013) compared babies delivered by cesarean section and babies delivered vaginally adjusting for measured covariates using logit regression. For instance, women on Medicaid were more likely to deliver vaginally with an odds ratio of 1.43, while women with third party insurance (e.g., Blue Cross) were more likely to deliver by cesarean section with odds ratio 1.46, and additive adjustments on the logit scale were intended to correct for this. Using similar methods and focusing on premature babies of $\geq$ 32 weeks gestational age,

Malloy (2009) reached similar findings.

In contrast, for very premature infants of 22-25 weeks gestational age, Malloy (2008) concluded: "Cesarean section does seem to provide survival advantages for the most immature infants..." (page 285). As in the other studies, the comparison was of babies delivered by one method or the other with adjustments for measured covariates by logit regression.

With varied emphasis, these studies note the problem of confounding by indication. They note that a direct comparison of babies delivered by cesarean section and babies delivered vaginally could be biased by aspects of the baby and the mother that led to the decision to deliver by one method rather than the other, and this is true even if logit regression is used to adjust for measured covariates. The decision to perform a cesarean section in one case but not in another may reflect indications that were evident to the physicians or mothers involved but not evident in measured covariates. This seems especially likely when a complex choice is made in a thoughtful, deliberate way. For a baby of gestational age 23-24 weeks, these considerations may include a medical judgement about the viability of the baby, and a mother's concern for a baby who may face severe life-long health problems. When studying a survival outcome, one is especially concerned about comparing groups of babies that may have been constructed with the viability of those babies in mind. One might prefer circumstances in which more or fewer babies were delivered by cesarean section for reasons that had nothing to do with the particular situation of

the baby and mother.

The finding that cesarean sections did not benefit more mature preterm babies did not stir up much controversy, but the finding of benefit for very premature babies was more controversial and surprising. We set out to study this using an instrument for cesarean section among babies 23-24 weeks of gestational age.

4.2.2 An instrument: variation among hospitals in the use of cesarean section for older babies

As noted in §4.2.1, confounding by indication occurs when patients receive treatments for good reasons, for instance because a physician believes giving the treatment to this patient will benefit this patient. It turns out that the use of cesarean section varies substantially from one hospital to the next. A mother may deliver by cesarean section not because of anything unique to her but simply because she delivers at a hospital that makes more extensive use of cesarean section.

Our instrument is the predicted c-section rate among babies of 23-24 weeks gestational age at the hospital where the baby was delivered. The rate is predicted using logit regression with four predictors. Three predictors describe the hospital's use of c-sections for older babies, that is: (a) the rate among babies with gestational age 25-32 weeks, (b) the rate among babies with gestational age 33-36 weeks, (c) the rate among babies with gestational age 37+ weeks. The fourth predictor was (d) the malpractice insurance rate in the county in which the hospital was located.

There is evidence that cesarean sections are more common in regions where the risk of malpractice litigation is greater; e.g., Dubay, Kaestner, and Waidmann (1999), Baicker, Buckles, and Chandra (2006) and Yang et al. (2009). The continuous instrument was the predicted probability from the logit regression. The value of this instrument would have been constant within a hospital but for predictor (d) which varied from year to year, so the instrument was constant in a given hospital in a given year, and was describing the proclivity of the hospital to perform c-sections rather than anything about a particular baby or mother.

4.2.3 Matching to strengthen the instrument

Available pretreatment covariates described the mother (e.g., her age), her baby (e.g., birth weight), the mother's Census tract (e.g., median household income), and the hospital. Hospitals vary in their abilities to care for premature infants. In particular, neonatal intensive care units (NICUs) are graded into seven levels of care based on available technology to care for sicker newborn patients. We matched exactly for the level of the NICU; see Table 4.1. We also used logit regression to estimate a hospital's risk-adjusted rates of two complications, thrombosis and wound infection, and matched to balance these variables. These scores were estimated from older babies, $\geq$ 25 weeks gestational age, so the scores make no use of outcomes for the group under study, namely babies of 23-24 weeks gestational age. The literature has suggested these two factors, thrombosis and wound infection, as

measures of the quality of care provided by the obstetrical hospital. In brief, the matching sought to compare similar mothers and babies from similar neighborhoods at similar hospitals.

Matched pairs were formed to be similar in terms of covariates and very different in terms of the instrument. Specifically, each of 1489 pairs contained two babies of 23-24 weeks gestational age, one at a hospital with a high frequency of use of c-sections for older babies, the other with a low frequency of use of c-sections for older babies. So the high and low groups looked similar in measured covariates, but one group went to hospitals that often delivered by c-section for older babies and the other group went to hospitals that used c-sections sparingly. As seen in Tables 4.1-4.3, the 1489 babies in the high group and the 1489 babies in the low group were similar in terms gestational weeks (23 or 24), birth weight, year of birth, mother's age, mother's education, mother's race/ethnicity, mother's health insurance, the technical level of the hospital's neonatal intensive care unit (NICU), pregnancy complications such as hypertension and oligohydramnios, number of prenatal care visits, parity, month that prenatal care started, various aspects of the mother's census tract. In Table 4.1, the three covariates were matched exactly. In Table 4.2, the five covariates had identical marginal distributions but were not exactly matched, a condition known as "fine balance." In Table 4.3, the difference in means for the covariates was never more than a tenth of a standard deviation, while the difference in the instrument was more than three standard deviations. This is

depicted for three continuous covariates and the instrument in Figure 1.

The matching was done in a new but simple way described in Appendix C. The approach taken here is a small extension and slight simplification of the approach taken in Zubizarreta et al. (2013). Described informally, nonoverlapping high and low instrument groups were defined by cutting the instrument in three places, discarding the middle. High and low babies were then selectively matched to push the groups further apart on the instrument, balance the covariates, and produce close individual pairs. The match was the solution to a constrained optimization problem.

4.2.4 Outcomes: c-section and mortality rates

The instrument is intended to manipulate one outcome, whether or not a baby is delivered by cesarean section, with possible effects on another outcome, mortality of the baby. As intended and expected, the instrument did manipulate the rate of cesarean sections; see Table 4.4. Table 4.4 counts pairs, not babies, in the manner that is commonly associated with McNemar's test; see Cox (1970). More than half the babies in both the high and low groups were delivered vaginally, but the 24.6% c-section rate in the low group was increased by more than half to 38.2% in the high group. When the two babies in a pair were delivered in different ways, the odds were $396/194 = 2.04$ to 1 that the high baby had the c-section.

Table 4.5 displays the main outcome, namely total in-hospital mortality. Table

4.5 is examining the possible effects of delivering at a high c-section hospital rather than a low c-section hospital, not yet the effects of c-sections themselves. The point estimate of the odds ratio favoring survival at a hospital with a high c-section rate is $360/185 = 1.95$. In the high group, survival rate was 34.8% and in the low group it was 23.0%, or a difference of $360 - 185 = 175$ survivors. If one believed naively that the matching in Tables 4.1-4.3 and Figure 1 had reproduced a paired randomized experiment that assigned one baby in each pair at random to the high hospital and the other to the low hospital (i.e., if one believed (i) but perhaps not (ii) in §4.1.1), then, using the method in Rosenbaum (2002, §6), one would be 95% confident that $A \geq 132$ babies were caused to survive because of delivery at a high hospital. (This is a one-sided 95% confidence interval derived from the randomization distribution, but if one prefers a two-sided interval, then the one-sided 97.5% interval is $A \geq 124$ babies rather than 132. In a paired randomized experiment, A is an unobserved random variable; see Rosenbaum (2002) or on-line Appendix II.) Moving away from the naive model for treatment assignment (i.e., moving away from (i) in §4.1.1), if an unobserved covariate doubled the odds of delivery at a high hospital and doubled the odds of survival, then the one-sided 95% confidence interval is $A \geq 66$ babies were caused to survive because of delivery at a high hospital. (More precisely, the 95% interval is $A \geq 66$ at $\Gamma = 1.25$ by the method in Rosenbaum (2002), and this amplifies to $(\Lambda, \Delta) = (2, 2)$ by the method in Rosenbaum and Silber (2009).) If an unobserved covariate doubled the odds of delivery at a high hospital and quadrupled

the odds of survival, then the one-sided 95% confidence interval is $A \geq 23$ babies were caused to survive because of delivery at a high hospital (or technically, this is the 95% interval at $\Gamma = 1.25$ which amplifies to $(\Lambda, \Delta) = (2, 4)$). The ostensible effects of delivering at a high rather than low c-section hospital are not sensitive to small departures from random assignment. So far, nothing has been said about the effects of c-sections, only about the effects of delivering at hospitals that do more of them.

In Table 4.4, the high c-section hospitals did $D = 396 - 194 = 202$ more c-sections than did the low c-section hospitals and 175 more babies survived. If the high-versus-low grouping were a valid instrument for delivery by c-section, then the Wald estimator would attribute the additional survivors at high c-section hospitals to the additional c-sections at those hospitals, that is, ignoring sampling variability, 175 additional survivors attributed to 202 additional c-sections. Assuming that the high-versus-low grouping is a valid instrument (that is, assuming both (i) and (ii) in §4.1.1), the Wald estimate of the effect of c-sections on the survival of babies who receive them because they were born at high c-section hospitals is $175/202 = 0.87$, an impressive ratio, not quite one more survivor for one more c-section. There is substantial sampling variability and possible bias in assignment to high or low hospitals, and both must be addressed, the first using a confidence statement, the second using sensitivity analysis. An interesting quantity is A/D where A is the attributable effect in the previous paragraph and D is number of additional c-

sections at high c-section hospitals. The 95% confidence intervals for A/D are $A/D \geq 132/202 = 0.65$ for randomization inference ($\Gamma = 1$), $A/D \geq 66/202 = 0.33$ for an unobserved covariate that doubled the odds of delivering at a high c-section hospital and doubled the odds of survival ($\Gamma = 1.25$), and $A/D \geq 23/202 = 0.11$ for an unobserved covariate that doubled the odds of delivering at a high c-section hospital and quadrupled the odds of survival ($\Gamma = 1.5$). (In on-line Appendix II, it is noted that A/D is the ratio of an unobserved to an observed random variable and a confidence interval for it is discussed.)

The exclusion restriction would be false if high c-section hospitals were more aggressive in many ways in their efforts to save babies of 23-24 weeks gestational age and if some of the reduced mortality were due to other aspects of the care provided at high c-section hospitals. Is the exclusion restriction compatible with other things we think we know?

4.2.5 A test of the exclusion restriction

As discussed in §4.2.1, the literature claims that there is no benefit from cesarean section for older preterm babies, say 30-34 weeks gestational age. Presuming — that is, tentatively and uncritically assuming — that claim to be true, we tested the exclusion restriction by redoing the study for babies of 30-34 weeks gestational age. It is important to realize that the literature is based on direct comparisons of babies delivered by c-section and babies delivered vaginally, whereas we used

an instrument, and there are other differences to be discussed in a moment. So we are really asking whether different methodologies concur in saying c-sections benefit babies at 23-24 weeks gestational age and not at 30-34 weeks gestational age, or whether an aporia has been produced, in which it is not reasonable to believe everyone's methodology, in the literature and our own, is producing correct conclusions about the effects of c-sections.

There were, of course, many more babies born at 30-34 weeks gestational age and the mortality rate was much lower. We matched in a manner similar to that in §4.2.3, but because there were many more babies involved, we made more extensive use of exact matching. This produced 23631 pairs of babies of 30-34 weeks gestational age with covariate balance and instrument separation similar to that seen in Tables 4.1-4.3 and Figure 1 for the younger babies.

As before for babies of 23-24 weeks gestational age, the instrument worked for babies of 30-34 weeks gestational age, with high babies more likely than low babies to be delivered by cesarean section. The mortality results appear in Table 4.6. After noting that the mortality rates are very different in Tables 4.5 and 4.6, one notes also that high babies had lower mortality rates than low babies in both tables, and the odds ratios are somewhat different in magnitude but neither is small, $360/185 = 1.95$ for 23-24 weeks and $1076/672 = 1.60$ for 30-34 weeks. We also looked for a trend, and indeed the odds ratio is larger at 30 weeks gestational age and smaller at 34 weeks. We redid the study again for babies of 25-29 weeks gestational age, finding

mortality results between Tables 4.5 and 4.6.

So the claims in the literature and our results sound plausible and reasonable if taken one at a time, but they cannot all be correct inferences about the effects of cesarean section on mortality. The conclusion is an aporia, individually plausible claims that are mutually incompatible. Of course, many things could have gone wrong, either in the literature or in our study. In our study, the two assumptions required of an instrument might be false. The literature implicitly assumes that if one takes account of observed covariates, say by logit regression, then one has reproduced a randomized experiment (or formally, they implicitly assume ignorable treatment assignment), and that assumption gets people in no end of trouble in observational studies. Are there other possibilities?

Indeed, there is another possibility. The cited literature in §4.2.1 focused on neonatal deaths, excluding fetal deaths, whereas we looked at all deaths. If a woman was pregnant with a baby of 23-24 weeks gestational age and the pregnancy terminated at that time, then we did not distinguish a death moments before birth and a death moments after birth. Remember that a baby of 23-24 weeks gestational age will require substantial medical assistance to remain alive. To our minds, the death of a baby of 23-24 weeks gestational age is a biological event, whereas the classification of that death as before or after birth may be little more than bookkeeping, perhaps an attempt to reduce the emotional pain of an event that is typically distressing for the mother.

Because our findings differ from the literature, we separated fetal and neonatal deaths, as shown in Tables 4.7 and 4.8. Consider what Tables 4.7 and 4.8 would look like if one removed all pairs with at least one fetal death, that is, removed the first row and first column of each table. The remaining babies would be either alive or neonatal deaths, the outcomes studied in the existing literature. Indeed, the resulting tables would then agree with the existing literature, in that c-sections would look beneficial in Table 4.7 but not in Table 4.8. By contrast, including fetal deaths, c-sections look beneficial in both tables. Arguably, a death of a fetus of 23-24 weeks gestational age is a death of an extremely premature baby, a biological event, whereas the classification of that death into a fetal death or a neonatal death is partly a style of practice and a manner of speaking. Arguably, fetal deaths should not be excluded from all deaths, as they were not excluded in Tables 4.5 and 4.6.

The available evidence is aporetic. Each part looks plausible on its own but the parts are mutually inconsistent. Something has to give, but it is less than clear what that something should be. The literature finds a benefit from c-sections at 23-24 weeks gestational age but not at 30-34 weeks gestational age. The literature makes no effort to address unmeasured biases in the selection of individual babies for delivery by cesarean section, though biases at the individual level are at least plausible, perhaps more plausible than not. In contrast, our analysis uses an instrument to avoid selection biases operating at the level of individual babies, using the frequency of c-sections among older babies at a hospital as an instrument

for c-sections among babies of 23-24 weeks gestational age. Hospitals with higher frequencies of c-sections have somewhat lower mortality, and this difference is not sensitive to small biases of selection into high or low c-section hospitals. By virtue of assuming the exclusion restriction, the Wald estimator attributes higher survival to higher rates of c-sections, producing a point estimate of 87%, and that seems implausibly large — that is, 87% of c-sections save babies who would otherwise have died — however, confidence intervals include substantially smaller effects. The exclusion restriction could easily be false here if hospitals that do more c-sections also are more aggressive in other ways in their treatment of extremely premature infants — the exclusion restriction would wrongly attribute the effects of those other efforts to c-sections. Our results would look much more like the existing literature if we followed the literature in ignoring fetal deaths at 23-24 weeks gestational age, counting only neonatal deaths at 23-24 weeks gestational age, but we worry that in many cases the distinction between a fetal death and a neonatal death at 23-24 weeks gestational age is a distinction without much of a difference. The element that seems least ambiguous in all this is that hospitals that do more c-sections have lower total mortality at 23-24 weeks gestational age, a difference that is not easily attributed to small biases in selection of mothers into hospitals, although it could conceivably be explained by moderately large biases. Whether this difference is caused by c-sections or by something else these hospitals are doing is not as clear.

4.3 Summary

We have suggested that the assumptions of the instrumental variable argument are often testable providing an aporia is seen as an acceptable conclusion. An aporia is a collection of individually plausible but mutually incompatible propositions. An aporia is an advance in understanding, albeit an uncomfortable one. In the example, the result of testing the exclusion restriction is a heightened concern that the exclusion restriction may be false, and the IV analysis may be wrong, but also a heightened concern that some of the things we think we know from the literature, some of the things we assumed in testing the exclusion restriction, may themselves be false.

Table 4.1: Three variables were exactly matched in forming 1489 pairs of two babies with gestational ages 23-24 weeks, namely gestational age (23 or 24 weeks), the capability or level of the neonatal intensive care unit (NICU), and the year of birth (1993-2005). The table gives counts of babies, and these are identical in the high and low instrument group defined by the estimated probability of a c-section at a given hospital.

	Instrument Group	
	High	Low
Gestational age in weeks		
23 weeks	726	726
24 weeks	763	763
NICU Level		
1	333	333
2	56	56
3A	126	126
3B	480	480
3C	438	438
3D	15	15
FC	41	41
Year of birth		
1993	30	30
1994	47	47
1995	90	90
1996	89	89
1997	104	104
1998	124	124
1999	133	133
2000	132	132
2001	129	129
2002	166	166
2003	188	188
2004	132	132
2005	125	125

Table 4.2: Five variables were finely balanced in forming 1489 pairs of two babies with gestational ages 23-24 weeks, meaning that these variables had the same marginal distributions in the high and low instrument groups, so the counts are identical. The table gives counts of babies, and these are identical in the high and low instrument group defined by the estimated probability of a c-section at a given hospital.

	Instrument Group	
	High	Low
Mother had hypertension during pregnancy		
Yes	75	75
No	1437	1437
Oligohydramnios		
Yes	52	52
No	1308	1308
Mother's race/ethnicity		
Non-Hispanic White	551	551
Non-Hispanic Black	305	305
Hispanic	478	478
Non-Hispanic Asian/P. Islander	87	87
Other	36	36
Missing	32	32
Mother's education		
8th grade or less	128	128
Some high school	249	249
High school graduate	473	473
Some college	303	303
College graduate	164	164
More than college (MS, PhD)	108	108
Missing	64	64
Mother's health insurance		
Fee for service	116	116
HMO	647	647
Federal/State	662	662
Other	20	20
Uninsured	42	42
Missing	2	2

Table 4.3: Covariates balanced in mean only and forced imbalance in mean in the instrument in forming 1489 pairs of two babies with gestational ages 23-24 weeks. The table gives the mean of each covariate or instrument before and after matching, together with the difference in means divided by the standard deviation before matching (S-Dif). For Yes/No = Y/N variables, 1=Yes, 0=No. RAHR = risk adjusted hospital rate. PROM = premature rupture of membrane.

	Before matching			After matching		
	Mean		S-Dif	Mean		S-Dif
	High	Low		High	Low	
	Hospital Covariates					
Hospital delivery volume (#)	2850	2903	-0.03	2568	2722	-0.09
RAHR of thrombosis	0.00	0.00	0.31	0.00	0.00	0.09
RAHR of wound infection	0.00	0.00	0.18	0.00	0.00	-0.05
	Mother/Baby Covariates					
Birth weight (grams)	591.12	577.25	0.16	587.08	580.31	0.08
Hypertension (Y/N)	0.07	0.04	0.12	0.05	0.05	0.00
Chorioamnionitis (Y/N)	0.28	0.26	0.04	0.27	0.26	0.02
Mother's age (years)	28.15	26.89	0.19	27.69	27.61	0.01
Prenatal care visits (#)	7.04	5.89	0.27	6.52	6.32	0.05
Prenatal care missing (Y/N)	0.09	0.05	0.14	0.07	0.05	0.07
Parity	1.90	1.90	-0.00	1.91	1.77	0.09
Parity missing (Y/N)	0.01	0.01	0.02	0.01	0.01	0.03
Month prenatal care started	2.00	2.16	-0.14	2.00	2.12	-0.10
Month care started missing (Y/N)	0.08	0.04	0.20	0.06	0.04	0.09
Multiple delivery	1.27	1.19	0.15	1.22	1.18	0.09
Congenital (Y/N)	0.15	0.14	0.04	0.15	0.14	0.03
Placentation (Y/N)	0.23	0.20	0.07	0.22	0.20	0.04
Diabetes (Y/N)	0.03	0.03	0.00	0.03	0.03	-0.03
Pre-term labor (Y/N)	0.81	0.74	0.17	0.80	0.76	0.09
PROM (Y/N)	0.35	0.28	0.15	0.33	0.30	0.08
Small for gestation age (Y/N)	0.09	0.12	-0.11	0.09	0.12	-0.09
	Neighborhood Covariates from the Census					
Household median income ($)	45024	41435	0.21	44730	44848	-0.01
Income missing (Y/N)	0.00	0.00	0.03	0.00	0.00	0.00
Below Poverty Level (fraction)	0.11	0.16	-0.10	0.15	0.16	-0.03
	Instrumental variable					
C-sec. predicted prob.	0.38	0.23	2.56	0.40	0.22	3.12

Table 4.4: C-sections in 1489 matched pairs of babies of 23-24 weeks gestational age. The table counts pairs, not babies. As expected, c-section rates are higher in the high c-section group.

	Low Baby			
High Baby	C-section	Other	Total	High Baby Rate
C-section	173	396	569	38.2%
Other	194	726	920	61.8%
Total	367	1122	1489	
Low Baby Rate	24.6%	75.4%		100.0%

Table 4.5: Mortality in 1489 matched pairs of babies of 23-24 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby			
High Baby	Dead	Alive	Total	High Baby Rate
Dead	786	185	971	65.2%
Alive	360	158	518	34.8%
Total	1146	343	1489	
Low Baby Rate	77.0%	23.0%		100.0%

Table 4.6: Mortality in 23631 matched pairs of babies of 30-34 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby			
High Baby	Dead	Alive	Total	High Baby Rate
Dead	108	672	780	3.3%
Alive	1076	21775	22851	96.7%
Total	1184	22447	23631	
Low Baby Rate	5.0%	95.0%		100.0%

Table 4.7: Mortality by type of death in 1489 matched pairs of babies of 23-24 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby				
High Baby	Fetal Death	Neonatal Death	Alive	Total	High Baby Rate
Fetal Death	111	99	47	257	17.2%
Neonatal Death	220	356	138	714	48.0%
Alive	141	219	158	518	34.8%
Total	472	674	343	1489	
Low Baby Rate	31.7%	45.3%	23.0%		100.0%

Table 4.8: Mortality by type of death in 23631 matched pairs of babies of 30-34 weeks gestational age. The table counts pairs, not babies. Mortality rates are higher in the low c-section group.

	Low Baby				
High Baby	Fetal Death	Neonatal Death	Alive	Total	High Baby Rate
Fetal Death	64	6	298	368	1.6%
Neonatal Death	26	12	374	412	1.7%
Alive	692	384	21775	22851	96.7%
Total	782	402	22447	23631	
Low Baby Rate	3.3%	1.7%	95.0%		100.0%

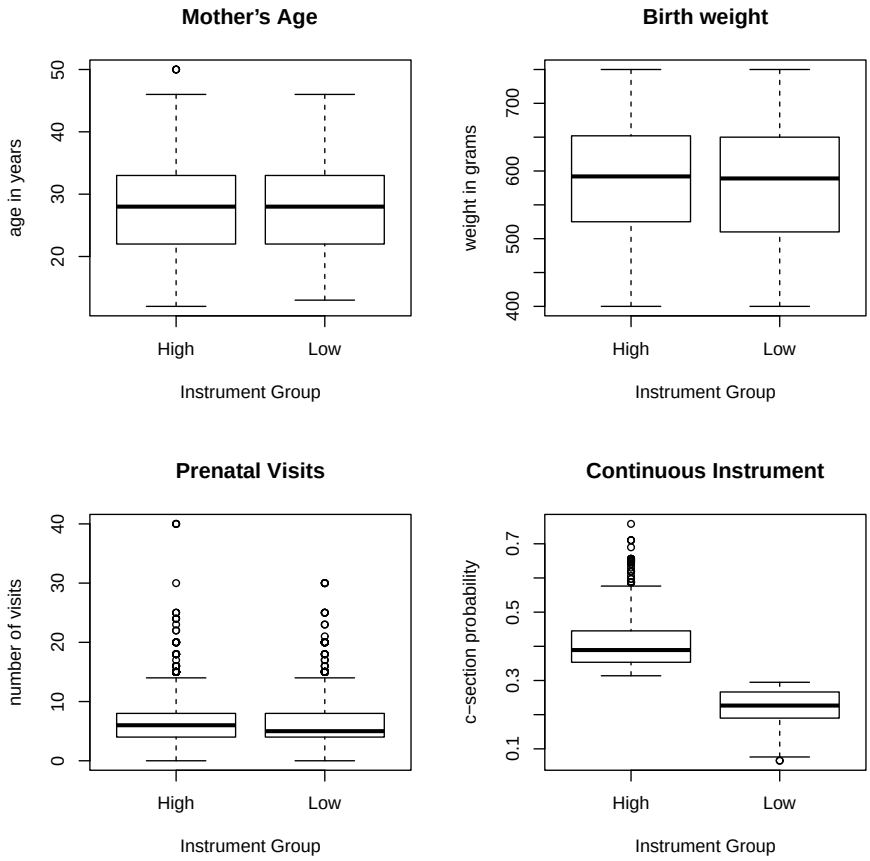

Figure 1: The match was intended to balance covariates and imbalance the instrument, and the boxplots depict this for three continuous covariates – mother's age, birth weight, and number of prenatal visits – and for the continuous instrument – the estimated probability of a c-section at the hospital predicted from c-section rates for older babies.

Bibliography

[1] Abadie A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97, 284-292.

[2] Abadie, A.(2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* 113, pp. 231-263.

[3] Angrist, J. D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review*, 80, 3, pp. 313-336.

[4] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91, pp. 444-455.

[5] Angrist, J.D. and Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics* 106,4, pp. 979-1014.

[6] Baicker, K., Buckles, K.S. and Chandra, A. (2006). Geographic variation in the appropriate use of cesarean delivery. *Health Affairs*, 24, w355-w467.

[7] Baiocchi, M., Small, D., Lorch, S., and Rosenbaum, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association* 105(492), pp. 1285-1296.

[8] Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83, pp. 679-697.

[9] Bertsekas, D. P. (1981). A new algorithm for the assignment problem. *Mathematical Programming*, 21, pp. 152-171.

[10] Boyle M. H., Torrance, G. W., Sinclair, J. C., and Horwood S. P. (1983). Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *The New England Journal of Medicine* 1983 Jun 2; 308(22):1330-7.

[11] Brookhart, M.A. and Schneeweiss, S.(2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *International Journal of Biostatistics* 3, pp. 14.

[12] Brookhart M. A., Wang P. S., Solomon D. H., and Schneeweiss S.(2006) Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 17, pp. 268-275.

[13] Chiba, Y. (2012). The large sample bounds on the principal strata effect with application to a prostate cancer prevention trial. *The International Journal of Biostatistics*, Vol. 8: Iss. 1, Article 12.

[14] Chiba, Y. and VanderWeele, T. J. (2011). A simple method for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology*, 173(7), pp. 745-751.

[15] Chen, H., Geng, Z., and Zhou, X. (2009) Identifiability and estimation of causal effects in randomized trials with noncompliance and completely non-ignorable missing data. *Biometrics* 65, pp. 675-691.

[16] Cheng, J. (2009). Estimation and inference for the causal effect of receiving treatment on a multinomial outcome. *Biometrics*, 65, pp. 96-103.

[17] Cheng, J., Qin, J. and Zhang, B. (2011). Semiparametric estimation and inference for distributional and general treatment effects.*Journal of the Royal Statistical Society: Series B*, 71, pp. 881-904.

[18] Cheng, J. and Small, D. (2006). Bounds on causal effects in three-arm trials with noncompliance. *Journal of the Royal Statistical Society, Series B*, 69, pp. 79-99.

[19] Clarke, P.S. and Windmeijer, F. (2012). Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107, pp. 1638-1652.

[20] Chung, J. H., Phibbs, C. S., Boscardin, W. J., Kominski, G. F., Ortega, A. N., and Needleman, J. (2010). The effect of neonatal intensive care level and hospital volume on mortality of very low birth weight infants. *Medical Care* 2010 Jul;48(7):635-44

[21] Committee on fetus and newborn. (2012) Levels of Neonatal Care *Pediatrics* 130:3, pp. 587-597.

[22] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E. (1959). Smoking and lung cancer. *Journal of the National Cancer Institute*, 22, pp. 173-203.

[23] Cox, D. R. (1970), *Analysis of Binary Data*, London: Metheun.

[24] Cox, D. R., Fitzpatrick, R., Fletcher, A. E., Gore, S. M., Spiegelhalter, D. J., and Jones, D. R. (1992). Quality-of-life assessment: Can we keep it simple? *Journals of the Royal Statistical Society: Series A*, 155, pp. 353-393.

[25] Derigs, U. (1988). Solving nonbipartite matching problems by shortest path techniques. *Annals Operatations Research,* 13, pp. 225-261.

[26] Doyle, L. W. and Victorian Infant Collaborative Study Group (2004). Evaluation of neonatal intensive care for extremely low birth weight infants in Victoria over two decades: II. Efficiency. *Pediatrics* 2004 Mar;113(3 Pt 1):510-4.

[27] Dubay, L., Kaestner, R., and Waidmann, T. (1999). The impact of malpractice fears on cesarean section rates. *Journal of Health Economics*, 18, pp. 491-522.

[28] Egleston, B. L., Scharfstein, D. O., Freeman, E. E., and West, S. K. (2007). Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, 8, pp. 526-545.

[29] Efron, B. and Tibshirani, R. (1998) The problem of regions. *Annals of Statistics*, 26, pp. 1687-1718.

[30] Fisher, R.A. (1935). *The Design of Experiments*, Edinburgh: Oliver & Boyd.

[31] Frangakis, C. E., and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-non treatment-noncompliance and subsequent missing ourcomes. *Biometrika* 86, 2, pp. 365-379.

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

[33] Freiman, M. and Small, D. (Under review). Large sample bounds on the survivor average causal effect when outcomes are censored by death.

[34] Gastwirth, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* 33, pp. 19-34.

[35] Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, 59, pp. 531-541.

[36] Guo, Z., Cheng, J., Lorch, S.A, and Small, D.S. (Under Review). Using an instrumental variable to test for unmeasured confounding.

[37] Hansen, B. B. (2007). Optmatch. *R News*, 7, pp.18-24. `R` package `optmatch`.

[38] Hayden, D., Pauler, D. K., and Schoenfeld, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*, 61, pp. 305-310.

[39] Hernan, M., Robins, J. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17, 360.

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

[41] Holland, P. W. H. (1988), Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18, pp. 449-484.

[42] Horowitz, J. and Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95, pp. 77-84.

[43] Hosman, C. A., Hansen, B. B., and Holland, P. W. H. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Annals of Applied Statistics*, 4, pp. 849-870.

[44] Howell E. M., Richardson D., Ginsburg P., Foot B. (2002). Deregionalization of neonatal intensive care in urban areas. *American Journal of Public Health* 92, pp.119-124.

[45] Hudgens, M. G., Hoering, A., and Self, S. G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine* 22, pp. 2281-2298.

[46] Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statistics & Probability Letters*, 78, pp. 144-149.

[47] Imai, K and Ratkovic M.(2012). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *Annals of Applied Statistics*, 7, 1, pp. 443-470.

[48] Imbens, G. W. and Rosenbaum, P. R. (2004). Robust, accurate confidence intervals with a weak instrument. *Journal of the Royal Statistical Society*, A 168, 109-126.

[49] Kling, J. R. (1999). The Effect of Prison Sentence Length on the Subsequent Employment and Earnings of Criminal Defendants. *Princeton University, Woodrow Wilson School*, Discussion Paper 208.

[50] Korn, E. L. and Baumrind, S.(1998). Clinician preferences and the estimation of causal treatment differences. *Statistical Science* 13, pp. 209-235.

[51] Korte, B. and Vygen, J. (2008), *Combinatorial Optimization: Theory and Algorithms*, New York: Springer.

[52] Lalani, T., Cabell, C. H., Benjamin, D. K. et al. (2010). Analysis of the impact of early surgery on in-hospital mortality of native valve endocarditis: use of propensity score and instrumental variable methods to adjust for treatment-selection bias. *Circulation*, 121, pp. 1005-1013.

[53] Lasswell S. M., Barfield W. D., Rochat R. W., Blackmon L. (2010). Perinatal regionalization for very low-birth-weight and very preterm infants: a meta-analysis. *Journal of the American Medical Association* 304, pp. 992-1000.

[54] Lee, K. G.(September 2013). Intraventricular hemorrhage of the newborn. MedlinePlus. http://www.nlm.nih.gov/medlineplus/ency/article/007301.htm

[55] Levy, D. E., O'Malley, A. J., and Normand, S. T. (2004). Covariate adjustment in clinical trials with non-ignorable missing data and non-compliance. *Statistics in Medicine* 23, pp. 2319-2339.

[56] Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data, 2nd ed. Wiley, New York.

[57] Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14, pp. 570-580.

[58] Lorch, S. A., Baiocchi, M., Ahlberg, C. E, and Small, D. S. (2012). The differential impact of delivery NICU on the outcomes of premature infance. *Pediatrics*,130, pp. 1-9.

[59] Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *American Statistician*, 65, pp. 21-30. `R` package `nbpmatching`.

[60] Malloy, M. H. (2008). Impact of cesarean section on neonatal mortality rates among very preterm infants in the United States, 2000-2003. *Pediatrics*, 122, pp. 285-292.

[61] Marcus, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational Statistics*, 22, pp. 193-201.

[62] Malloy, M. H. (2009), Impact of cesarean section on intermediate and late preterm births: United States 2000-2003. *Birth*, 36, pp. 26-33.

[63] McClellan, M., McNeil, B. J., Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortal-

ity? Analysis using instrumental variables. *Journal of the American Medical Association*, 272, 859–66.

[64] Mealli F., Imbens, G., Ferro S., and Biggeri, A. (2004). Analyzing a randomized trial on breast self examination with noncompliance and missing outcomes. *Biostatistics*, 5, 2, pp. 207-222.

[65] Mealli, F. and Pacini, B. (2013). Using secondary outcomes and covariates to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108, 503, pp. 1120-1131.

[66] Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statistical Science*, 5, pp. 463-480.

[67] Olkin, I. and Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* 32 , pp. 448-465, (correction in 36, pp. 343-344).

[68] 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, pp. 598-607.

[69] Phibbs, C. S., Baker, L. C., Caughey, A. B., Danielsen, B., Schmitt, S. K., and Phibbs, R. H. (2007). Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *The New England Journal of Medicine* 2007 May 24;356(21):2165-75;

[70] Phibbs C. S., Mark D. H., Luft H. S. et al. (1993). Choice of hospital for delivery: a comparison of high-risk and low-risk women. *Health Services Research* 28(2), pp. 201-222.

[71] Profit, J., Lee, D., Zupancic, J. A., Papile, L., Gutierrez, C., Goldie, S. J., Gonzalez-Pier, E., Salomon, J. A. (2010). Clinical benefits, costs, and cost-effectiveness of neonatal intensive care in Mexico. *PLoS Medicine* ,2010 Dec 14;7(12)

[72] Rescher, N. (2009), *Aporetics: Rational Deliberation in the Face of Inconsistency*, Pittsburgh: University of Pittsburgh Press.

[73] Richardson, D. K., Reed, K., Cutler, J. C., et al. (1995). Perinatal regionalization vs hospital competition: the Hartford example. *Pediatrics* 96, pp. 417-423.

[74] Rogowski, J. A., Horbar, J. D., Staiger, D. O., Kenny, M., Carpenter, J., and Geppert, J. (2004). Indirect vs direct hospital quality indicators for very low-birth-weight infants. *THe Journal of the American Medical Association* 2004 Jan 14;291(2):202-9.

[75] Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74, pp. 13-26.

[76] Rosenbaum, P. R. (2002). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97, pp. 183-192.

[77] Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, 21, pp. 57-71.

[78] Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society* Series B (Methodological). Vol. 45, No. 2, pp. 212-218.

[79] 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* Vol 79, pp. 516-524.

[80] Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in observational studies. *Journal of the American Statistical Association*, 104, 1398-1405. (`R` package `sensitivitymv`)

[81] Roy, J. and Hennessy, S. (2011). Bayesian hierarchical pattern mixture models for comparative effectiveness of drugs and drug classes using healthcare data: a case study involving antihypertensive medications. *Statistics in biosciences* 3, pp. 79-93.

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

[83] Rubin, D. B. (2000). Comment on causal inference without counterfactuals. *Journal of the American Statistical Association*, 95, pp. 435-438.

[84] Schafer, J. L. (1997a). Analysis of incomplete multivariate data, Chapman & Hall, London.

[85] Small, D. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102, pp. 1049-1058.

[86] Small, D. and Cheng, J. (2009). Discussion of "Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics* 65, pp. 682-686.

[87] Small, D. S. and Rosenbaum, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103, pp. 924-933.

[88] Stock, J. (2002). Instrumental Variables in Economics and Statistics. *International Encyclopedia of the Social Sciences*, Amsterdam: Elsevier, pp. 7577-7582.

[89] Swanson, S. A. and Hernan, M. A. (2013). How to report instrumental variable analyses. *Epidemiology*, 24, pp. 924-933.

[90] Tan Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101, pp. 1607-1618.

[91] Taylor, L., and Zhou, X. (2009). Relaxing latent ignorability in the ITT analysis of randomized studies with missing data and noncompliance. *Statistica Sinica* 19, pp, 749-764.

[92] The Acute Respiratory Distress Syndrome Network (2000). Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *The New England Journal of Medicine*, 342, 18, pp. 1301-1308.

[93] Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), pp. 1-67.

[94] Vlastos, G. (1994). *Socratic Studies*, New York: Cambridge University Press.

[95] Walker A.(1996). Confounding by indication. *Epidemiology* 7, pp. 335-336.

[96] Weiss, L. (1955). A note on confidence sets for random variables. *Annals of Mathematical Statistics*, 26, pp. 142-144.

[97] Welch, B. L. (1937). On the z-test in randomized blocks. *Biometrika*, 29, pp. 21-52.

[98] Werner, E. F., Han, C. S., Savitz, D. A., Goldshore, M., Lipkind, H. S. (2013). Health outcomes for vaginal compared with cesarean delivery of appropriately grown preterm neonates. *Obstetrics and Gynecology*, 121, pp. 1195-1200.

[99] Yan, W., Hu, Y. and Geng, Z. (2012). Identifiability of causal effects for binary variables with baseline data missing due to death. *Biometrics* 68, pp. 121-128.

[100] Yang, Y. T., Mello, M. M., Subramanian, S. V. and Studdert, D. M. (2009). Relationship between malpractice litigation pressure and rates of cesarean section and vaginal birth after cesarean section. *Medical Care*, 47, pp. 234-242.

[101] Yang, F., Lorch, S. A. and Small, D. S.(2014). Estimation of causal effects using instrumental variables with nonignorable missing covariates: application to effect of type of delivery hospital on premature infants. *The Annals of Applied Statistics*, to appear.

[102] Yang, F., Lorch, S. A. and Small D. S. (2014). Supplement to "Estimation of causal effects using instrumental variables with nonignorable missing covariates: application to effect of type of delivery NICU on premature infants."

[103] Yang, F., and Small, D.S. (2014). Using post-quality of life measurement information in censoring by death problems. Submitted.

[104] Yang, F., Zubizarreta, J., Small, D.S., Lorch, S.A., and Rosenbaum, P.(2014). Aporetic conclusions when testing the validity of an instrumental variable. Submitted.

[105] Yeast J. D., Poskin M., Stockbauer J. W., Shaffer S. (1998). Changing patterns in regionalization of perinatal care and the impact on neonatal mortality. *American Journal of Obstetrics and Gynecology* 178, pp.131-135.

[106] Yu, B. B. and Gastwirth, J. L. (2005). Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics*, 6, pp. 201-209.

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

[108] Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107, 1360-1371. (`R` software `mipmatch` at www-stat.wharton.upenn.edu/~josezubi/)

[109] Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Annals of Applied Statistics*, 7, pp. 25-50.

Appendix A

Two-Stage Censoring by Death

A.1 Bounds of the SACE

Given the value of π_{1100}, the linear programming problem (2.3.14)-(2.3.20) has a solution if and only if the set $\Phi = [\max\{q_{110|1} - \pi_{1000}, \frac{q_{111|1}\pi_{1100}}{\pi_{1111}+\pi_{1110}+\pi_{1010}}\},\ \frac{q_{110|1}\pi_{1100}}{\pi_{1100}+\pi_{1000}}]$ is not empty, which is essentially $\frac{q_{110|1}}{p_{10|1}} \geq \frac{q_{111|1}}{p_{11|1}}$, an inequality that must be satisfied based on assumptions 4-6. If Φ is not empty, let $\underline{T} = \max\{q_{110|1} - \pi_{1000}, \frac{q_{111|1}\pi_{1100}}{\pi_{1111}+\pi_{1110}+\pi_{1010}}\}$, $\overline{T} = \frac{q_{110|1}\pi_{1100}}{\pi_{1100}+\pi_{1000}}$, the solution to the linear programming problem is,

$$\begin{aligned} &\max((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100}) \\ &= \frac{q_{111|1}(\pi_{1111} + \pi_{1110})}{\pi_{1111} + \pi_{1110} + \pi_{1010}} + \overline{T} \end{aligned} \tag{A.1.1}$$

$$\min((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100})$$

$$= \begin{cases} \underline{T} & \text{if } \frac{q_{111|1}\pi_{1100}}{\pi_{1010}} \leq \underline{T} \\ q_{111|1} + (1 - \frac{\pi_{1010}}{\pi_{1100}})\dot{T} & \text{if } \frac{q_{111|1}\pi_{1100}}{\pi_{1010}} \geq \overline{T} \\ q_{111|1} + (1 - \frac{\pi_{1010}}{\pi_{1100}})\ddot{T} & \text{if } \underline{T} < \frac{q_{111|1}\pi_{1100}}{\pi_{1010}} < \overline{T} \end{cases} \tag{A.1.2}$$

where

$$\dot{T} = \begin{cases} \underline{T} & \text{if} \pi_{1010} \leq \pi_{1100} \\ \overline{T} & \text{if} \pi_{1010} > \pi_{1100} \end{cases}$$

$$\ddot{T} = \begin{cases} \underline{T} & \text{if} \pi_{1010} \leq \pi_{1100} \\ \frac{q_{111|1}\pi_{1100}}{\pi_{1010}} & \text{if} \pi_{1010} > \pi_{1100} \end{cases}$$

Thus, given a fixed value of π_{1100}, the bounds for the SACE are given by:

$$\min(SACE \mid \pi_{1100}) = \frac{\min((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100}) - (q_{111|0} + q_{110|0})}{\pi_{1111} + \pi_{1110} + \pi_{1100}}$$

$$\max(SACE \mid \pi_{1100}) = \frac{\max((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100}) - (q_{111|0} + q_{110|0})}{\pi_{1111} + \pi_{1110} + \pi_{1100}}$$

From section 2.3.1, we know that π_{1100} is not point identified, but bounded: $\pi_{1100} \in I$, $I = [\max\{0, p_{11|0} + p_{10|0} - p_{10|1}\}, \quad \min\{p_{10|0}, p_{10|1}\}]$, we have,

$$\min SACE = \min_{\pi_{1100} \in I} [\frac{\min((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100}) - (q_{111|0} + q_{110|0})}{\pi_{1111} + \pi_{1110} + \pi_{1100}}]$$

$$\max SACE = \max_{\pi_{1100} \in I} [\frac{\max((\pi_{1111}\mathrm{E}(Y_i(1) \mid 1111) + \pi_{1110}\mathrm{E}(Y_i(1) \mid 1110) + \pi_{1100}\mathrm{E}(Y_i(1) \mid 1100)) \mid \pi_{1100}) - (q_{111|0} + q_{110|0})}{\pi_{1111} + \pi_{1110} + \pi_{1100}}]$$

One can prove that the expression on the left side of equation (A.1.2) is continuous as a function of π_{1100} and both the functions on the left side of equations (A.1.1) and (A.1.2) are non-decreasing as functions of π_{1100}. Thus, the $\max SACE$ could be

achieved when π_{1100} is $\min\{p_{10|0}, p_{10|1}\}$ which is the right end point of the range for π_{1100}, and the min $SACE$ could be achieved when π_{1100} is $\max\{0, p_{11|0}+p_{10|0}-p_{10|1}\}$ which is the left end point of the range for π_{1100}. Based on this observation, we can obtain the formula for the bound of SACE which is given in (2.3.21) and (2.3.22).

A.2 The ARDSNet data

861 patients were randomized to receive mechanical ventilation with either lower tidal volume or traditional tidal volume. The lower tidal volume group contained 432 patients and the traditional tidal volume group contained 429 patients. We created our variables based on the recorded answers for the study termination form and weaning form.

The first time point (day 28) survival information is obtained through the "ST2DT" variable in the study termination sub-dataset which recorded the date of death. If the date of death for subject i is below day 28, then S_{1i} is 0 and the QOL is not defined; otherwise, S_{1i} is 1.

For the patients who survive to day 28, the QOL that whether patient was able to breathe without assistance by day 28 was well defined. The variable "UNASSIST" in the study termination sub-dataset recorded whether the patient was able to sustain unassisted breathing for $\geq$ 48 hours during the first 28 days after initiation of study procedures. However, even if the patient sustained unassisted breathing for at least 48 hours, the patient could return to assisted breathing before day 28. The variable

"ASSIST" recorded this information. If the patient returned to assisted breathing from unassisted breathing for at least 48 hours, the "ASSIST" was recorded as "Yes". Thus, for patients whose "UNASSIST" was recorded as "No", we view them as the ones who were not able to breathe without assistance by day 28. For patients whose "UNASSIST" was recorded as "Yes", and "ASSIST" was recorded as "No", we view them as the ones who were able to breathe without assistance by day 28; for patients whose "UNASSIST" was recorded as "Yes" and "ASSIST" was recorded as "Yes", each patient could either (a) have had unassisted breathing at some point and then returned to assisted breathing and still be on assisted breathing at day 28 or (b) have had unassisted breathing before day 28, returned to assisted breathing before day 28 and then returned to unassisted breathing before day 28. For these patients, we further use the weaning sub-dataset which recorded in detail about each patients' breathing status to figure out whether the patient was able to breathe without assistance by day 28.

Our second time point survival indicator is whether the patient was eventually discharged home with unassisted breathing. This information was recorded in the variable "STATUS" which described patient status at study termination.

Appendix B

IV with Nonignorable Missing Covariates

B.1 E-step Estimates

The fomulas to update N in E step are as follows, $\forall x_2, x_3, x_4, y$

$$N_{1,1,x_2,x_3,x_4,a,0,y} = NN_{x_2,x_3,x_4,1,0,y}$$

$$N_{1,1,x_2,x_3,x_4,n,1,y} = NN_{x_2,x_3,x_4,0,1,y}$$

$$\mathrm{E}N_{1,1,x_2,x_3,x_4,a,1,y} = NN_{x_2,x_3,x_4,1,1,y}\frac{P_{1,1,x_2,x_3,x_4,a,1,y}}{P_{1,1,x_2,x_3,x_4,a,1,y} + P_{1,1,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{1,1,x_2,x_3,x_4,c,1,y} = NN_{x_2,x_3,x_4,1,1,y}\frac{P_{1,1,x_2,x_3,x_4,c,1,y}}{P_{1,1,x_2,x_3,x_4,a,1,y} + P_{1,1,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{1,1,x_2,x_3,x_4,n,0,y} = NN_{x_2,x_3,x_4,0,0,y}\frac{P_{1,1,x_2,x_3,x_4,n,0,y}}{P_{1,1,x_2,x_3,x_4,n,0,y} + P_{1,1,x_2,x_3,x_4,c,0,y}}$$

$$\mathrm{E}N_{1,1,x_2,x_3,x_4,c,0,y} = NN_{x_2,x_3,x_4,0,0,y}\frac{P_{1,1,x_2,x_3,x_4,c,0,y}}{P_{1,1,x_2,x_3,x_4,n,0,y} + P_{1,1,x_2,x_3,x_4,c,0,y}}$$

$$\mathrm{E}N_{0,1,x_2,x_3,x_4,a,0,y} = N3_{x_2,x_4,1,0,y}\frac{P_{0,1,x_2,x_3,x_4,a,0,y}}{\sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,a,0,y}}$$

$$\mathrm{E}N_{0,1,x_2,x_3,x_4,n,1,y} = N3_{x_2,x_4,0,1,y}\frac{P_{0,1,x_2,x_3,x_4,n,1,y}}{\sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,n,1,y}}$$

$$\mathrm{E}N_{0,1,x_2,x_3,x_4,a,1,y} = N3_{x_2,x_4,1,1,y}\frac{P_{0,1,x_2,x_3,x_4,a,1,y}}{\sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,a,1,y} + \sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{0,1,x_2,x_3,x_4,c,1,y} = N3_{x_2,x_4,1,1,y}\frac{P_{0,1,x_2,x_3,x_4,c,1,y}}{\sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,a,1,y} + \sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{0,1,x_2,x_3,x_4,n,0,y} = N3_{x_2,x_4,0,0,y}\frac{P_{0,1,x_2,x_3,x_4,n,0,y}}{\sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,n,0,y} + \sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,c,0,y}}$$

$$\mathrm{E}N_{0,1,x_2,x_3,x_4,c,0,y} = N3_{x_2,x_4,0,0,y}\frac{P_{0,1,x_2,x_3,x_4,c,0,y}}{\sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,n,0,y} + \sum_{x_3=1}^{x_3=q_3} P_{0,1,x_2,x_3,x_4,c,0,y}}$$

$$\mathrm{E}N_{1,0,x_2,x_3,x_4,a,0,y} = N4_{x_2,x_3,1,0,y}\frac{P_{1,0,x_2,x_3,x_4,a,0,y}}{\sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,a,0,y}}$$

$$\mathrm{E}N_{1,0,x_2,x_3,x_4,n,1,y} = N4_{x_2,x_3,0,1,y}\frac{P_{1,0,x_2,x_3,x_4,n,1,y}}{\sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,n,1,y}}$$

$$\mathrm{E}N_{1,0,x_2,x_3,x_4,a,1,y} = N4_{x_2,x_3,1,1,y}\frac{P_{1,0,x_2,x_3,x_4,a,1,y}}{\sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,a,1,y} + \sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{1,0,x_2,x_3,x_4,c,1,y} = N4_{x_2,x_3,1,1,y}\frac{P_{1,0,x_2,x_3,x_4,c,1,y}}{\sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,a,1,y} + \sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{1,0,x_2,x_3,x_4,n,0,y} = N4_{x_2,x_3,0,0,y}\frac{P_{1,0,x_2,x_3,x_4,n,0,y}}{\sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,n,0,y} + \sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,c,0,y}}$$

$$\mathrm{E}N_{1,0,x_2,x_3,x_4,c,0,y} = N4_{x_2,x_3,0,0,y}\frac{P_{1,0,x_2,x_3,x_4,c,0,y}}{\sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,n,0,y} + \sum_{x_4=1}^{x_4=q_4} P_{1,0,x_2,x_3,x_4,c,0,y}}$$

$$\mathrm{E}N_{0,0,x_2,x_3,x_4,a,0,y} = NB_{x_2,1,0,y}\frac{P_{0,0,x_2,x_3,x_4,a,0,y}}{\sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{1,0,x_2,x_3,x_4,a,0,y}}$$

$$\mathrm{E}N_{0,0,x_2,x_3,x_4,n,1,y} = NB_{x_2,1,0,y}\frac{P_{0,0,x_2,x_3,x_4,n,1,y}}{\sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{1,0,x_2,x_3,x_4,n,1,y}}$$

$$\mathrm{E}N_{0,0,x_2,x_3,x_4,a,1,y} = NB_{x_2,1,1,y}\frac{P_{0,0,x_2,x_3,x_4,a,1,y}}{\sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,a,1,y} + \sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{0,0,x_2,x_3,x_4,c,1,y} = NB_{x_2,1,1,y}\frac{P_{0,0,x_2,x_3,x_4,c,1,y}}{\sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,a,1,y} + \sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,c,1,y}}$$

$$\mathrm{E}N_{0,0,x_2,x_3,x_4,n,0,y} = NB_{x_2,1,1,y}\frac{P_{0,0,x_2,x_3,x_4,n,0,y}}{\sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,n,0,y} + \sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,c,0,y}}$$

$$\mathrm{E}N_{0,0,x_2,x_3,x_4,n,0,y} = NB_{x_2,1,1,y}\frac{P_{0,0,x_2,x_3,x_4,c,0,y}}{\sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,n,0,y} + \sum_{x_4=1}^{x_4=q_4}\sum_{x_3=1}^{x_3=q_3} P_{0,0,x_2,x_3,x_4,c,0,y}}$$

Appendix C

Testing IV Assumptions

C.1 A new bipartite matching algorithm for strengthening an IV

Following Baiocchi et al. (2010) and Zubizarreta et al. (2013), we used matching to strengthen the instrumental variable while balancing observed covariates. However, we changed, simplified, and in some contexts improved, a key element. These two papers both took a single population, discarded part of the population, split the remainder into pairs, where the pairs balance covariates while being far apart on the instrument. Discarding a middle portion, an ambiguous portion, of the population makes the instrument stronger, improving its design sensitivity, making the study less sensitive to bias from nonrandom assignment of encouragement; see Small and Rosenbaum (2008). Traditionally, splitting a single population into pairs is called

by the awkward name "nonbipartite matching" which means "not two parts." The history of the awkward name involves the fact that optimal two-part matching (e.g., treatment versus control matching), so called optimal bipartite matching, was studied and solved first; see Korte and Vygen (2008) for a textbook discussion of both problems with comprehensive references. Baiocchi et al. (2010) used an algorithm and Fortran code for optimal nonbipartite matching created by Derigs (1988), as implemented in Lu et al.'s (2011) R package `nbpmatching`; it minimizes the total distance within pairs formed from a single population, discarding a portion of the population using a technical trick called "sinks". Zubizarreta et al. (2013) used integer programming, specifically Zubizarreta's (2012) `mipmatch` package, to impose additional linear constraints on the nonbipartite match, such as requiring nominal covariates to be perfectly balanced or requiring means of continuous covariates to be close. Also, Zubizarreta et al. (2013) changed the optimized objective function along the lines suggested in Rosenbaum (2012), so as to optimize the number of individuals discarded. A feature of the nonbipartite approach is that individual pairs are far apart on the instrument, but the high baby in one pair may be lower on the instrument than the low baby in some other pair. Depending upon the nature of the instrument and the covariates, that feature may or may not be reasonable. It might be reasonable if the meaning of the instrument changed with the levels of the covariate. In the current study, with an instrument defined in terms of a hospital's rate of use of c-sections in older babies, this feature did not seem reasonable.

We wanted each and every baby in the high group to have a higher value of the instrument than each and every baby in the low group. This change was implemented in a simple way using bipartite matching. We cut the population into three groups based on the value of the instrument, V, where the middle group, $0.29 \leq V \leq 0.31$ contained 10% of the population and was discarded. Write $\{\alpha_1, \ldots, \alpha_h, \ldots, \alpha_H\}$ for the H remaining babies in the high group and $\{\beta_1, \ldots, \beta_\ell, \ldots, \beta_L\}$ for the L remaining babies in the low group, noting that $V_{\alpha_h} > V_{\beta_\ell}$ for every h, ℓ. We then matched babies in the high group to babies in the low group to be close in terms of a covariate distance, $\delta_{h\ell}$, measuring how similar baby α_h and baby β_ℓ were in terms of covariates, and far apart on the instrument, with $\delta_{h\ell} = \infty$ if $V_{\alpha_h} - V_{\beta_\ell} < \omega$ for an $\omega > 0$. The covariate distance combined a robust Mahalanobis distance for covariates with $\delta_{h\ell} = \infty$ for mismatches on the variables in Table 1. Write $a_{h\ell} = 1$ if baby α_h in the high group is paired with baby β_ℓ in the low group, $a_{h\ell} = 0$ otherwise, so that we require $a_{h\ell} \in \{0, 1\}$, $\sum_{i=1}^{H} a_{i\ell} \leq 1$, $\sum_{j=1}^{L} a_{hj} \leq 1$, for each h, ℓ. In principle, one could simply minimize the total distance within matched pairs, $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell}\, \delta_{h\ell}$, subject to $\sum_{h=1}^{H} \sum_{\ell=1}^{L} a_{h\ell} = \min(H, L)$, and this could be done using the optimal assignment algorithm — e.g., Bertsekas' (1981) auction algorithm as made available in the `pairmatch` function of Hansen's (2007) `optmatch` package in `R`. Alternatively, one could make ω larger, as we did, to further strengthen the instrument, discarding some babies to achieve this more stringent objective. This can be done using the same software for the assignment algorithm without constrain-

ing $\sum_{h=1}^{H}\sum_{\ell=1}^{L} a_{h\ell}$ and instead minimizing $\sum_{h=1}^{H}\sum_{\ell=1}^{L} a_{h\ell}\,\delta_{h\ell} - \lambda\sum_{h=1}^{H}\sum_{\ell=1}^{L} a_{h\ell}$ for specified $\lambda > 0$, and this determines an optimal number of babies to discard; see Rosenbaum (2012) for extensive specifics.

As in Zubizarreta (2012) and Zubizarreta et al. (2013), we used integer programming, not the optimal assignment algorithm, to minimize $\sum_{h=1}^{H}\sum_{\ell=1}^{L} a_{h\ell}\,\delta_{h\ell} - \lambda\sum_{h=1}^{H}\sum_{\ell=1}^{L} a_{h\ell}$ but with additional linear constraints. As in these references, these added constraints forced the fine balance in Table 4.2 and the close mean match seen in Table 4.3. Moreover, we added a new constraint to further strengthen the instrument. Setting $\delta_{h\ell} = \infty$ if $V_{\alpha_h} - V_{\beta_\ell} < \omega$ forces each matched pair to differ by $\geq \omega$ in terms of the instrument. The new additional constraint forced the mean difference in the instrument V to differ by a larger number, $\Omega > \omega$, so every pair meets the minimum requirement of ω, but on average a larger difference of Ω is achieved. The new constraint was $\sum_{h=1}^{H}\sum_{\ell=1}^{L} a_{h\ell}\,(V_{\alpha_h} - V_{\beta_\ell} - \Omega) > 0$.

C.2 Confidence intervals and sensitivity analyses for A/D

Section 4.2.4 of the paper reported confidence intervals for ratios of survival effects to differences in the frequency of use of c-sections. These intervals are new but are a direct extension of an existing method. This appendix describes the new method and briefly indicates its justification. There are I matched pairs, $i = 1, \ldots, I$, of two

subjects, $j = 1, 2$, one encouraged, $Z_{ij} = 1$, the other not, $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. In §4.2.3, there are $I = 1489$ pairs of two babies, one at a high c-section hospital, $Z_{ij} = 1$, the other at a low c-section hospital, $Z_{ij} = 0$. Pairs were matched for observed covariates $\mathbf{x}_{ij}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i, but the matching may have failed to control an unobserved covariate u_{ij}, so possibly $u_{i1} \neq u_{i2}$ for many or all i. Baby ij has two potential binary responses (r_{Tij}, r_{Cij}), one r_{Tij} if encouraged with $Z_{ij} = 1$, the other r_{Cij} if unencouraged with $Z_{ij} = 0$. In §4.2.4, $r_{Tij} = 1$ if baby ij would survive at the high c-section hospital in the i^{th} pair, $r_{Tij} = 0$ otherwise, and $r_{Cij} = 1$ if baby ij would survive at the low c-section hospital in the i^{th} pair, $r_{Cij} = 0$, otherwise. Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$ for all babies ij — in words, switching from a low c-section hospital to a high c-section hospital does not change any baby's survival. In a randomized paired experiment with binary response, McNemar's test is the randomization test of Fisher's H_0. Each baby is observed under one treatment, so the effect of the treatment, $r_{Tij} - r_{Cij}$, is not observed for any baby; see Neyman (1923), Welch (1937) and Rubin (1974). An important unobservable quantity in §4.2.4 is the attributable effect $A = \sum_{i=1}^{I} \sum_{j=1}^{2} Z_{ij} (r_{Tij} - r_{Cij})$; it is the unobservable number of babies caused to survive by virtue of delivering at the high c-section hospital. In constructing one-sided confidence intervals for A, we follow Angrist, Imbens and Rubin (1996) in additionally assuming $r_{Tij} \geq r_{Cij}$, so a 23-24 week baby who would survive with the stress of a vaginal delivery, $r_{Cij} = 1$, would

also survive with the reduced stress of a cesarean delivery, $r_{Tij} = 1$. A two-sided interval for A may be constructed from two one-sided intervals. Under Fisher's null hypothesis of no effect, every $r_{Tij} - r_{Cij} = 0$, so $A = 0$ no matter how treatments Z_{ij} are assigned.

Similarly, (d_{Tij}, d_{Cij}) is the binary indicator of delivery by cesarean section or vaginal delivery (1 for c-section, 0 for vaginal delivery) at the high and low c-section hospital. Baby ij is said to be a complier if encouragement shifts the baby's delivery in the encouraged direction, that is, if $1 = d_{Tij} > d_{Cij} = 0$, so this baby would be delivered by c-section at the high c-section hospital in pair i and would be delivered vaginally at the low c-section hospital in pair i. Baby ij is said to be an always taker if $d_{Tij} = d_{Cij} = 1$, a never taker if $d_{Tij} = d_{Cij} = 0$, and a defier if $0 = d_{Tij} < d_{Cij} = 1$, and we follow the usual practice of assuming there are no defiers, $d_{Tij} \geq d_{Cij}$, so a baby who would be delivered by c-section at a low c-section hospital would also be delivered by c-section at a high c-section hospital; see Angrist, et al. (1996) for discussion of this terminology. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}, \mathbf{x}_{ij}, u_{ij}),\ i = 1, \ldots, I,\ j = 1, 2\}$ and $\mathcal{Z}$ for the event that $Z_{i1} + Z_{i2} = 1$ for each i. In a randomized paired encouragement design, encouragement Z_{ij} is assigned by $\Pr(Z_{i1} = 1 \mid \mathcal{F}, \mathcal{Z}) = 1/2$, $Z_{i2} = 1 - Z_{i1}$, and assignments in distinct pairs are independent. A simple model for sensitivity analysis in observational studies has $1/(1+\Gamma) \leq \Pr(Z_{i1} = 1 \mid \mathcal{F}, \mathcal{Z}) \leq \Gamma/(1+\Gamma)$ for specified $\Gamma \geq 1$, $Z_{i2} = 1 - Z_{i1}$, with independent assignments in distinct pairs, so ran-

domization inference corresponds with $\Gamma = 1$; see Rosenbaum (1987; 2002, §4) for discussion of this method of sensitivity analysis, and for other methods, see Cornfield et al. (1959), Rosenbaum and Rubin (1983), Gastwirth (1992), Marcus (1997), Small (2007), Yu and Gastwirth (2005), Hosman et al. (2010), and Liu et al. (2013). Write R_{ij} for the baby ij's observed survival response, $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\; r_{Cij}$, and D_{ij} for the observed delivery, $D_{ij} = Z_{ij}\, d_{Tij} + (1 - Z_{ij})\; d_{Cij}$. Appendix Table 1 renumbers the two babies in a pair so $Z_{i1} = 1$, $Z_{i2} = 0$, and then records the joint distribution of $(R_{i1}, D_{i1}, R_{i2}, D_{i2}) = (r_{Ti1}, d_{Ti1}, r_{Ci2}, d_{Ci2})$.

As $I \rightarrow \infty$ in a randomized encouragement design, for fixed α, $0 < \alpha < 1$, conventionally $\alpha = 0.05$, it is possible to find an observed random variable $\widetilde{A}$ such that $\Pr\left(A \geq \widetilde{A} \,\middle|\, \mathcal{F},\, \mathcal{Z}\right)$ tends to a probability $\geq 1 - \alpha$, so that $A \geq \widetilde{A}$ holds with 95% confidence, that is, the unobserved attributable effect A is at least equal to $\widetilde{A}$ except in at most $100\alpha\%$ of experiments; see Rosenbaum (2002) for specifics and Weiss (1955) for general discussion of confidence sets for unobserved random variables in terms of observed random variables. Moreover, in a sensitivity analysis in an observational study, if the bias in treatment assignment is at most $\Gamma \geq 1$, then there is an observed random variable $\widetilde{A}_\Gamma$ such that $\Pr\left(A \geq \widetilde{A}_\Gamma \,\middle|\, \mathcal{F},\, \mathcal{Z}\right)$ tends to a probability $\geq 1 - \alpha$ as $I \rightarrow \infty$; again, see Rosenbaum (2002).

The exclusion restriction says that encouragement that does not change the delivery (d_{Tij}, d_{Cij}) does not change the response (r_{Tij}, r_{Cij}), that is, $r_{Tij} = r_{Cij}$ whenever $d_{Tij} = d_{Cij}$. Stated informally, the exclusion restriction says that if

high c-section hospitals sometimes save the lives of babies, then they do it by performing c-sections not by doing something else. The exclusion restriction could easily be false: high c-section hospital could be more aggressive in many ways in trying to save the lives of babies of 23-24 weeks gestational age, and c-sections may produce only a part or even none of the survival effect of generally more aggressive treatment. The exclusion restriction places a series of constraints on the relationship between the observed appendix Table C.1 and the unobservable table recording $(r_{Ci1}, d_{Ci1}, r_{Ci2}, d_{Ci2})$. The unobserved table is called the pivot table. Consider, for example, the 44 pairs in the first row and first column of the observed appendix Table C.1. Because the exclusion restriction says $r_{Tij} = r_{Cij}$ whenever $d_{Tij} = d_{Cij}$, those 44 pairs could be in the same place in the pivot table or some could move to the third and fourth row of the first column, but none could move to the second row. Also, pairs in the second row could move to the fourth row. In fact, the only differences that can exist between the observed and pivot tables are the movements just described. Under the exclusion restriction, A is the total number of pairs that are in the first row of the observed table and in the fourth row of the pivot table.

Let $b_{ij} = 1$ if $r_{Tij} > r_{Cij}$ and $d_{Tij} > d_{Cij}$, and $b_{ij} = 0$ otherwise. If $b_{ij} = 1$, then baby ij would survive receiving a c-section at the high c-section hospital in pair i and would die without a c-section at the low c-section hospital in pair i. Using the exclusion restriction, $r_{Tij} - b_{ij}\left(d_{Tij} - d_{Cij}\right) = r_{Cij}$, and the attributable effect is $A = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}\left(r_{Tij} - r_{Cij}\right) = \sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij} b_{ij}\left(d_{Tij} - d_{Cij}\right)$. The mean

difference in survival is:

$$\begin{aligned} T_r &= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}R_{ij} - (1-Z_{ij})\,R_{ij} = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}\left\{r_{Cij} + b_{ij}\left(d_{Tij} - d_{Cij}\right)\right\} - (1-Z_{ij})\,r_{Cij} \\ &= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} (2Z_{ij}-1)\,r_{Cij} + \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}b_{ij}\left(d_{Tij} - d_{Cij}\right) \\ &= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} (2Z_{ij}-1)\,r_{Cij} + \frac{A}{I}. \end{aligned}$$

In a randomized paired encouragement experiment, $\mathrm{E}\left(Z_{ij}=1 \,\middle|\, \mathcal{F},\, \mathcal{Z}\right) = 1/2$ so that

$$\mathrm{E}\left\{ \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} (2Z_{ij}-1)\,r_{Cij} \,\middle|\, \mathcal{F},\, \mathcal{Z} \right\} = 0, \text{ and } \mathrm{E}\left(\frac{A}{I} \,\middle|\, \mathcal{F},\, \mathcal{Z} \right) = \frac{1}{2I}\sum_{i=1}^{I}\sum_{j=1}^{2} (r_{Tij} - r_{Cij}) = \tau$$

so that T_r and A/I are both unbiased for the average effect of encouragement, τ; however, departures from random assignment (i.e., failures of (i) in §4.1.1) can introduce bias. The observable random variable $T_d = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}D_{ij} - (1-Z_{ij})\,D_{ij}$ is the difference between the number of c-sections performed by the high and low c-section hospitals; in Table 4.4 it is $T_d = 396 - 194 = 202$. It is a descriptive, not a causal quantity: it describes what happened, not what would happen. The Wald estimator is T_r/T_d. For the Wald estimate to work, encouragement must increase the frequency of what is encouraged so that T_d converges in probability to a strictly positive quantity $\delta > 0$ as $I \to \infty$, and that is assumed here; therefore, with high probability, high c-section hospitals have done more c-sections among the I pairs than low c-section hospitals for sufficiently large I, and $\Pr\left(T_d \le 0 \,\middle|\, \mathcal{F},\, \mathcal{Z}\right)$ is negligible for large I. The quantity

$$W = \frac{\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}b_{ij}\left(d_{Tij} - d_{Cij}\right)}{\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}D_{ij} - (1-Z_{ij})\,D_{ij}} = \frac{A}{\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}D_{ij} - (1-Z_{ij})\,D_{ij}} = \frac{A/I}{T_d}$$

is the number of babies caused to survive by a c-section in a high c-section hospital as a fraction of the number of additional c-sections performed by high c-section hospitals. Now, W is the ratio of an unobservable random variable A/I, a causal quantity, and an observed random variable T_d, a descriptive quantity, so W is unobservable. The quantity W is directly interpretable on its own; however, it might reasonably be regarded as the intended finite sample estimand of the Wald estimator, in the sense that T_r/T_d and W both converge in probability as $I \to \infty$ to the average effect of c-sections on compliers if encouragement is randomized within pairs; see Angrist et al. (1996) for discussion of this estimand. Given the large sample confidence interval, $A \geq \widetilde{A}_\Gamma$ with $\Pr\left\{ A \geq \widetilde{A}_\Gamma \middle|\ \mathcal{F}, \mathcal{Z} \right\} \geq 1 - \alpha$ for sufficiently large I, and continuing to regard $\Pr\left(T_d \leq 0 \,|\, \mathcal{F}, \mathcal{Z}\right)$ is negligible for large I, we have $\Pr\left\{ A/T_d \geq \widetilde{A}_\Gamma/T_d \middle|\ \mathcal{F}, \mathcal{Z} \right\} = \Pr\left\{ W \geq \widetilde{A}_\Gamma/T_d \middle|\ \mathcal{F}, \mathcal{Z} \right\} \geq 1 - \alpha$ for sufficiently large I. The confidence interval $W \geq \widetilde{A}_\Gamma/T_d$ was reported in §4.2.4.

Table C.1: Mortality R_{ij} and mode of delivery D_{ij} (C = C-section, V = vaginal) in 1489 matched pairs of babies of 23-24 weeks gestational age. For the high baby with $Z_{ij} = 1$, mortality is $R_{ij} = r_{Tij}$ and delivery is $D_{ij} = d_{Tij}$, whereas for the low baby with $Z_{ij} = 0$, mortality is $R_{ij} = r_{Cij}$ and delivery is $D_{ij} = d_{Cij}$. To avoid notational ambiguity, in this table j is changed so the first baby, $j = 1$, is the high baby. The table counts pairs, not babies.

	Low Baby, $Z_{i2} = 0$			
	C-Alive	C-Dead	V-Alive	V-Dead
	$r_{Ci2} = 1$	$r_{Ci2} = 0$	$r_{Ci2} = 1$	$r_{Ci2} = 0$
High Baby, $Z_{i1} = 1$	$d_{Ci2} = 1$	$d_{Ci2} = 1$	$d_{Ci2} = 0$	$d_{Cij} = 0$
C-Alive, $r_{Ti1} = 1$, $d_{Ti1} = 1$	44	54	37	144
C-Dead, $r_{Ti1} = 0$, $d_{Ti1} = 1$	37	38	36	179
V-Alive, $r_{Ti1} = 1$, $d_{Ti1} = 0$	31	35	46	127
V-Dead $r_{Ti1} = 0$, $d_{Ti1} = 0$	47	81	65	488