Determining Causal Effects on Small-Cluster Data

Nathan E. Liittschwager
Department of Mathematics
University of California, San Diego
La Jolla, CA, 92093
nliittsc@ucsd.edu

June 8, 2019

Abstract

In statistical applications, the assumption of independent and identically distributed data greatly simplifies the process of analysis. However, when data may be ordered into hierarchical clusters in which the clusters are independent, but units within each cluster are correlated, the assumption of independence and identical distribution need not hold. While the statistical literature provides methods of analysis for such situations, analysis is greatly complicated when trying to determine *causal* effects, which are necessarily much stronger than simple statistical association. In this paper we investigate the methodology of determining causal effects in small-cluster data in a real data application, as well as present a simulation study to investigate the consequences of a misspecified causal model.

(Note: This paper is a working draft and results/conclusions may change. The author humbly apologizes for typos, bad notation, and unclear mathematics. All mistakes are entirely his.)

1 Introduction

"Correlation is not causation" is a mantra often repeated in introductory statistics classes. Professors in statistics and the sciences often take great pains to caution beginning students against the pitfalls of erroneous causal conclusions in their data, which may beg the question: When can causation be concluded from data? Intuition and scientific practice appeals to the notion of the randomized experiment, of treatment and control units, and well-defined interventions. Yet this intuition may be misleading when the data is observational, as is often the case in reality. However, determining causal inferences from observational data is not an intractable problem, and under a set of assumptions and careful data modeling, causal inferences may be extracted. However, when the data are not i.i.d. - independent and identically distributed - the problem of causal inference in observational data becomes much harder.

This paper was inspired by a real-world problem of determining the causal effect of pre-natal exposure to a particular drug on the presence of minor malformations in new-born infants, in which some infants have a twin. The difficulty of the problem comes from the fact that the drug exposure was necessarily on the pregnant *mother*, but the outcome of interest was on her

newly born *children*, which is further complicated by the fact that some of the children had a twin. In this case, the data may be viewed as hierarchical, in which children are clustered according to their mother. While many of the mothers in our data set bore only one child, twins born from one mother are necessarily correlated with each other, yet received the same treatment. Our work is in modeling this data and unearthing the causal effect of the drug exposure.

The paper will be organized as follows. We first provide a theoretical summary of the assumptions and theorems that allow an investigator to make causal inferences from observational data. In particular, we work in the *Potential Outcomes* framework that was developed in the Rubin-Neyman Causal Model, and will be more expository in nature. We then present our data as a motivating example of the need for causal methodology, and give an overview of the statistical methodology used in the analysis of this data. Finally, we present the results of our real-data analysis as well assess model fit by means of a simulation study that mimics postulated data-generation mechanism.

2 The Potential-Outcomes Framework

The following discussion is a summary of the results that may be found in (Imbens, G., Rubin, D., 2015). We provide the summary primarily to introduce the reader to our notation, assumptions regarding the data-generating process, as well as justify methodological choices. The main benefit of operating within this framework is it allows us to encode the language of causal theory into a probabilistic framework, though we ignore the measure-theoretic details.

2.1 Notation and Definitions.

To avoid some confusion later down the road, we will initially use our own notation. Suppose we have a finite population of N units, a real valued response Y, and a dichotomous treatment variable $z = 0, 1$ at a particular point in time. For the ith person, if $z = 1$ we say they are *treated* and if $z = 0$, we say they are a *control* or *untreated*. Then for the ith person, we say $Y(i, 1)$ and $Y(i, 0)$ are *potential outcomes* - the value of the response Y under the treatment $z = 1$ and $z = 0$, respectively, for the ith person. We say that z has a *causal effect* for the ith person if

$$Y(i, 1) - Y(i, 0) \neq 0. \tag{1}$$

Note that presently $Y(i, z)$ is not a random variable if z is fixed - rather it is a characteristic of that person, much like hair or eye color, but under the different treatment conditions of $z = 0$ or $z = 1$. We usually say that $Y(i, 0)$ and $Y(i, 1)$ are *counterfactuals* for each other, or *what would have happened* under the conditions induced by $z = 0$ or $z = 1$. The issue here becomes surprisingly philosophical, but we can form a mental model for this by perhaps imagining $Y(i, 0)$ and $Y(i, 1)$ as (say) the starting salaries of the ith person had they pursued a master's degree ($z = 1$) or had they not ($z = 0$). If $Y(i, 1) > Y(i, 0)$, then we can say that the master's degree had a causal effect for person i. We acknowledge that this example is overly simple.

In the finite population case, we define the *average causal* or the *treatment effect* as:

$$\tau_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y(i,1) - \frac{1}{N}\sum_{i=1}^{N} Y(i,0). \tag{2}$$

There is a glaring issue in this definition: it requires that we observe *both* $Y(i,1)$ and $Y(i,0)$ for each person $i = 1, 2, ..., N$ in the population. Since a treatment is defined as occurring at a particular point in time, it is impossible to solve this problem - we cannot go back in time to administer a different treatment, even if we had access to all N individuals. Moreover, if we suppose that the population is infinite, we cannot necessarily even write down the average treatment effect of the individuals in the population. This problem is called the *Fundamental Problem of Causal Inference* (Holland, Paul W., 1986). However, probability theory provides some solutions.

In the event that we are sampling from a hypothetical super population, the individual i is the result of a random sample, and hence the potential outcomes $Y(i,0)$ and $Y(i,1)$ become realizations of a proper random variable. We'll write these random potential outcomes as $Y(0)$ and $Y(1)$ respectively, and say that for a random sample of n people, we have the independent and identically distributed random variables $Y_i(0)$ and $Y_i(1)$ for $i = 1, 2, ..., n$. In this case, (2) becomes

$$\tau_{ATE} = E[Y(1)] - E[Y(0)] \tag{3}$$

which then gives the expected value of the treatment effect on the population. However, for each individual i, we may not necessarily know which potential outcome we are observing, $Y_i(0)$ or $Y_i(1)$. Moreover, in observational data, the sample will not represent the population as a whole, but rather a specific subset of the population. For example, election polls conducted in major cities only represent those individuals who happen to live in a major city. When dealing with imperfect samples, each individual i comes with a set of covariates X_i which may distinguish them from other individuals in the population. These covariates X_i introduce what is often called a *confounding relationship*, though the exact definition of a "confounder" is still debated (VanderWeele, 2013).

2.2 Assumptions

In order to make meaningful causal inferences, we work with several main assumptions. They are *Consistency*, *Conditional Exchangeability*, and *Positivity* (Hernan and Robins 2019).

Assumption 1: Consistency. By *Consistency*, we *do not* mean the statistical property of an estimator converging in probability. Rather, suppose the outcome of interest is the random variable Y and $(Y_i)_{i=1}^n$ is a sequence of i.i.d. realizations. In order to tie each Y_i to potential outcomes $Y_i(0)$ and $Y_i(1)$, we suppose that each individual i is randomly assigned to control or treatment conditions as the result of a random variable Z which we call the *treatment assignment mechanism*. Let $(Z_i)_{i=1}^n$ be a sequence of i.i.d. binary random variables where $Z_i = 1$ assigns the ith individual to treatment and $Z_i = 0$ assigns the ith individual to control. Then the assumption of *Consistency* is

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0). \tag{4}$$

That is, if $Z_i = 1$, we observe $Y_i(1)$ and if $Z_i = 0$, we observe $Y_i(0)$. Therefore, it is meaningful to take expectations of Y_i.

Assumption 2: Conditional Exchangeability. Suppose we are in observational data setting, and for each individual i in the random sample, they come with a set of covariates X_i. Then, under *Conditional Exchangeability*, we suppose that $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i | X_i$. That is, we hope that within levels of X_i, the treatment assignment mechanism is independent of the potential outcomes.

To gain intuition on what this means, suppose X_i is a covariate that indicates how ill a patient is, and Z_i is the result of a doctor choosing to treat or not. If X_i indicates the patient is in severe condition, it is more likely that $Z_i = 1$, since the doctor would be more inclined to treat the patient. But, patients with worse levels of X_i will also have worse levels of $Y_i(1)$ and $Y_i(0)$. So X_i introduces *confounding* in that it is *not independent* of Z_i and $Y_i(z)$, which will introduce statistical correlation between $Y_i(z)$ and Z_i, $z = 0, 1$. Essentially, without conditioning on X_i, the average treatment effect τ_{ATE} may actually be biased. In our example, since more ill patients receive treatment, but also have worse outcomes, it may appear that treatment does nothing at all, or even has negative effects.

Assumption 3: Positivity. By *Positivity*, we assume that for every person i in the sample, $0 < Pr(Z_i = 1|X_i) < 1$. In other words, every person has a non-zero probability of being given control *or* treatment. In the case that $Pr(Z_i = 1|X_i) = 0$ or 1, then by *Consistency*, we cannot observe one of $Y_i(1)$ or $Y_i(0)$ within certain levels of X_i. The problem of causal inference becomes unsolvable in this case.

2.3 Statistical Consequences

Under the 3 assumptions given above, the problem of causal inference fundamentally becomes a "missing data" problem which may be solved by statistics. Given a random sample, we observe Y_i after randomly assigning individuals to treatment and control conditions. Now, by our three assumptions,

$$\begin{aligned} &E[Y_i|Z_i = 1, X_i] - E[Y_i|Z_i = 0, X_i] \\ &\quad = E[Y_i(1)|Z_i = 1, X_i] + E[Y_i(0)|Z_i = 0, X_i] \\ &\quad = E[Y_i(1)|X_i] - E[Y_i(0)|X_i] \end{aligned}$$

where we used *Consistency* in the first equality, and *Conditional Exchangeability* in the second equality. Note that *Positivity* is required to make it conditional expectations meaningful.

Now, by the Double Expectation Theorem, taking the expected value of the conditional expectations above gives

$$\begin{aligned} &E\big\{E[Y_i|Z_i = 1, X_i] - E[Y_i|Z_i = 0, X_i]\big\} \\ &\quad = E\big\{E[Y_i(1)|X_i] - E[Y_i(0)|X_i]\big\} \\ &\quad = E[Y_i(1)] - E[Y_i(0)] \\ &\quad = \tau_{ATE}. \end{aligned}$$

Therefore, by finding the treatment effect within levels of X_i, we can infer the population average treatment effect by taking the expectation over X_i. One way of doing this is by postulating a model to generate the data. In other words, we postulate a function f that generates the conditional expectation

$$E[Y_i|Z_i, X_i] = f(X_i, Z_i). \tag{5}$$

We know from statistical theory that the best linear approximation of f is the model

$$E[Y_i|Z_i, X_i] = \beta_0 + \beta_1 X_i + \theta Z_i, \tag{6}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is a vector of coefficients and θ is the *average change* in Y_i when $Z_i = 1$. This lends θ its interpretation of the *average treatment effect.* In essence, under *Consistency*, *Conditional Exchangeability*, and *Positivity*, linear regression has a causal interpretation, and (6) may be estimated with Ordinary Least Squares. When $Y_i \sim \text{Bernoulli}(p_i)$, then (6) may be modeled with a logit link function and maximum-likelihood.

Aside. What is the difference between "regular" linear regression and a "causal" linear regression? Largely, the assumptions on one's data. When doing *descriptive* inference or analysis, we are concerned with modeling the statistical associations between Y_i and some observed covariates X_i. However, if X_i is vector valued, say $X_i = (X_{i1}, X_{i2})'$, a statistical model for association is still legitimate if (say) X_{i2} is unobserved and thus not included in the model. In *causal inference*, we must observe both X_{i1} and X_{i2}, since without both, we have no guarantees that Conditional Exchangeability holds. This requirement is sometimes called the *No Unmeasured Confounders* assumption. Moreover, in the case that Y_i is correlated with X_i, there exists no statistical test for whether Y_i "caused" X_i, or whether X_i "caused" Y_i, since statistical correlation is simply a measure between two random variables, and not a measure of causal direction. Generally, domain expertise is needed to determine the direction of causality, and this often depicted with directed acyclic graphs, also called DAGs (Judea Pearl, 2009).

In general, causal inference requires heroic assumptions on one's data, and should not be taken lightly. The consolation is that in the case of a randomized experiment, where individuals are randomly sampled from a population, then randomly given treatment or control with equal probability, all three of our required assumptions are satisfied. In observational data, we can only hope that if the most important confounders are observed (those with large effects), then the bias introduced into the estimate of θ is relatively small, maintaining the legitimacy of causal inferences. Naturally, the problem of unmeasured confounders becomes more important as the true treatment effect θ shrinks to 0, as even small amounts of bias may flip the sign on the estimate $\hat{\theta}$.

2.4 The Propensity Score

Rosenbaum and Rubin provided a useful tool for the purposes of causal inference in their seminal paper *The Central Role of the Propensity Score in Observational Studies of Causal Effects*. In the case where X_i is vector valued and high dimensional, there may be non-overlapping regions in the distribution of X_i among certain groups in the sample. Such lack of overlap may violate one of the causal assumptions and make inference impossible. On the

other hand, the high dimension of X_i may make comparisons of treatment effects within levels of X_i computationally expensive and hard to interpret. What Rubin and Rosenbaum showed was that it is sufficient to condition not just on X_i, but also on a *function* of X_i, which they call a *balancing score.* We will sketch what this is, and its implications for casual inference. For a complete exposition on the theory, please see the original paper (Rosenbaum and Rubin, 1983).

Let $e(X_i) = Pr(Z_i = 1|X_i)$. We call this a *propensity score* as it is the ith individual's propensity towards receiving treatment. Rosenbaum and Rubin define a *balancing score* as any function b of X_i such that

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i|b(X_i). \tag{7}$$

That is, Conditional Exchangeability holds given $b(X_i)$. They note a trivial balancing score is $b(X_i) = X_i$. In particular, Rosenbaum and Rubin prove that $e(X_i)$ is a balancing score, and that all of the assumptions of causal inference hold with respect to $e(X_i)$. Therefore,

$$E[Y_i|Z_i = 1, e(X_i)] - E[Y_i|Z_i = 0, e(X_i)] \tag{8}$$
$$= E[Y_i(1)|e(X_i)] - E[Y_i(0)|e(X_i)]. \tag{9}$$

Therefore, taking the expectation of (8) with respect to $e(X_i)$ gives

$$E\big\{E[Y_i|Z_i = 1, e(X_i)] - E[Y_i|Z_i = 0, e(X_i)]\big\} \tag{10}$$
$$= E\{E[Y_i(1)|e(X_i)] - E[Y_i(0)|e(X_i)]\big\} \tag{11}$$
$$= \tau_{ATE}. \tag{12}$$

This result, with its corresponding lemmas in (Rosenbaum and Rubin, 1983) is sometimes called the *Propensity Score Theorem.* The main implication of this result is that the propensity score serves as a data-reduction method that may reduce the high dimensional distribution of X_i to that of a single continuous function. The propensity score is usually estimated with a logistic regression to produce $\hat{e}(X_i)$. From there, it may be used in a variety of ways, from non-parametric weighting, to regression adjustment, or a weighted regression (Imbens and Rubin, 2015).

3 Motivating Example

Working with our notions of *exchangeability*, we can take a covariate X as a *confounder* if X is required for the conditional independence of Y and Z. That is, $Y \perp\!\!\!\perp Z|X$ and $Y \not\perp\!\!\!\perp Z$ otherwise. In other words, a model of causal inference should condition on all random variables X that are predictive of both treatment assignment and the outcome (Dorie et al., 2016). Matters are somewhat complicated in the case that the data exhibits a hierarchical structure - in which individual response variables may be naturally organized into a structure that induces correlation among the responses. For example, patients may be organized into hospitals, and Y_{hk} may be a quantitative score of health of the kth patient in the hth hospital, for $k = 1, 2, ..., n_h$, $h = 1, 2, ..., H$. In this case, the health-scores of patients may be i.i.d. within a particular hospital, but not i.i.d. across hospitals, since the hospitals themselves are subjected to varying levels of funding, resources, skill of medical staff, etc. Matters are further complicated when trying to estimate treatment effects at the cluster- or

individual-level, though the problem is not intractable.

In light of the above, our real-world data provides a motivating example for the difficulties in addressing the fundamental problem of causal inference.

3.1 The Data

(Note: Due to the sensitive nature of the data, some information cannot be divulged. In particular, the treatment of interest was exposure to a proprietary drug, which may not be named in this paper.)

Our data consist of 261 infants born to a cohort of 248 mothers who voluntarily participated in a study measuring the effect of prenatal exposure to a proprietary drug on the occurrence of *minor malformations* in the infant. A minor malformation is defined as "a minor anomaly is a structural defect that deviates from the normal standard and has no major surgical, medical, or cosmetic importance" (AAP News and Journals). Since we have 262 infants born to 248 mothers, some mothers gave birth to multiple infants in the form of twins. There were 13 pairs of twins, hence 26 children total who were a part of a twin pairing. The rest of the infants were born as singletons. The data is naturally hierarchical, and we denote a particular mother with h for $h = 1, 2, ..., 248$. The kth child of the hth mother is denoted with h, k. No mother had more than two children.

The binary response of interest is a binary random variable Y on the infant level. That is, for the hth mother and the kth child, we have $Y_{hk} = 0$ if there were less than 3 minor malformations and $Y_{hk} = 1$ if there were three or more minor malformations. Among all children, there were 63 positive instances of 3+ minor malformations, and 198 negative instances. Similarly, the treatment indicator $Z_h = 0$ if the mother was not exposed to the drug, and $Z_h = 1$ if the mother was exposed to the drug. Note that given $z = Z_h$, all children from the hth mother has treatment level z. Among all children, 172 children were exposed to the drug in prenatal, and 89 were unexposed.

The data contained 41 other covariates. A significant number of these covariates were related to whether the mother presented an autoimmune disorder or not. In particular, there were binary indicators for Rheumatoid Arthritis and Chrohn's Disease, and 'Other' autoimmune diseases. The exposure of interest is a drug related to treating autoimmune disorders, hence nearly every mother in the data also presented some form of autoimmune disorder. 134 mothers had Rheumatoid Arthritis imputed, and various indicators for disease severity recorded. There were binary indicators for other psychiatric history and a history of birth defects. There were also indicators for history of smoking, alcohol use, psychiatric drug use, parity, prenatal care, and how they were referred to the study.

With only 261 individuals, 45 covariates is fairly high dimension, and much of the challenge is in coming up with a good causal model.

3.2 The models

There are two distinct levels present in our data. There is the clustering level (the mothers) and the unit level which are grouped by a cluster (the infants). The unique situation we find in our data is that the treatment is on the cluster-level, and determines the treatment

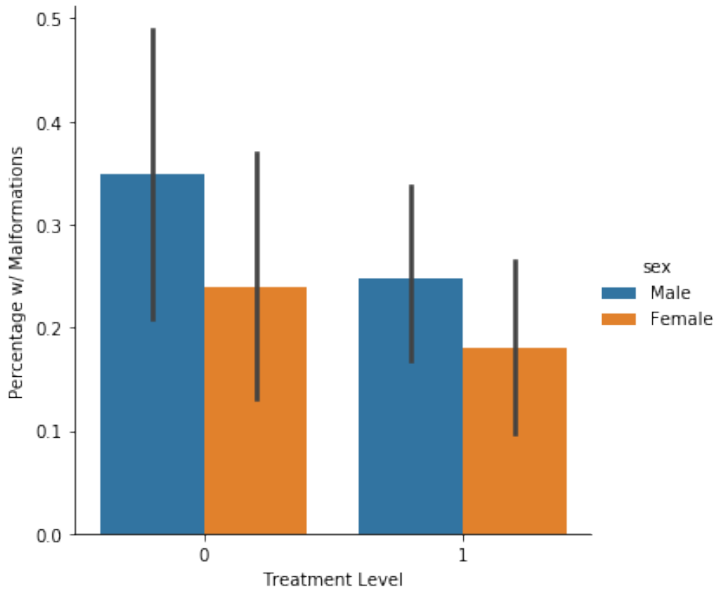

Figure 1: Proportion of infants who suffered minor malformations, divided by sex and drug exposure, with associated error bars. From left to right, the proportions are 0.248, 0.181, 0.349, 0.239.

status at the individual level. However, the response of interest is also at the individual level. However, the majority of potential confounders may be found at the clustering level - in fact, the sex of the infant is the only covariate at the unit level. However, domain expertise hypothesizes the infant's sex has a confounding effect with the response of interest, and thus should be included in a causal model.

We now present a possible statistical model to generate the data. Here "clusters" refer to our hypothetical mothers, as in the real data. We assume that the treatment assignment is generated at the cluster level, in a way that mimics the mother's pregnancy status and other confounders. Let V be a cluster covariate, and suppose there is a vector of binary indicators $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$ which denote the pregnancy type. Here if $X_1 = 1$, then the mother is having a single male, if $X_2 = 1$, then the mother is having a single female. If $X_3 = 1$, the mother is having twins of mixed sex. X_4 and X_5 similarly denote whether the mother is having twin males, or twin females, respectively. Note that for each person, only one of the X_i may be 1, and the rest are zero. Assuming V and $\mathbf{X}$ are confounders, the treatment assignment of the hth mother (cluster) is assumed to follow

$$\text{logit}Pr(Z_h = 1|V_h, X_h) = \alpha_0 + \alpha_1 V_h + \mathbf{X'}_h \boldsymbol{\alpha}_2 \tag{13}$$

where $\boldsymbol{\alpha}_2 = (\alpha_{2_1}, \alpha_{2_2}, \alpha_{2_3}, \alpha_{2_4}, \alpha_{2_5})'$ is a vector of coefficients that each correspond to the pregnancy type.

We have $Y_{hk} \sim \text{Bernoulli}(p_{hk})$, where $p_{hk} = Pr(Y_{hk} = 1|V_h, \mathbf{X}^*_{hk})$, where $\mathbf{X}^*_{hk}$ is the result of some transformation of $\mathbf{X}_h$. For example, since $\mathbf{X}_h$ indicates the type of pregnancy in the hth mother, the individual outcomes may depend not on the pregnancy type, but on

the sex of the offspring, and whether they are a twin or not. Then, $\mathbf{X}^*_{hk} = (\text{sex}_{hk}, \text{twin}_{hk})'$ which gives rise to the following model:

$$\text{logit}Pr(Y_{hk} = 1|V_h, \mathbf{X}^*{}_{hk}) = \beta_0 + \beta_1 V_h + \beta_2 \text{sex}_{hk} + \beta_3 \text{twin}_{hk} + \theta Z_h. \tag{14}$$

Here sex_{hk} and twin_{hk} are just binary indicators.

3.2.1 Justification

1. The treatment clearly occurs at the mother or cluster level, as the drug exposure was delivered as a treatment to the *mother's* disease while she was pregnant. Therefore, if $Z_h = z$, then $Z_{hk} = z$ for each child k.

2. The outcome may be seen as a binomial random variable at the mother level, but we suffer from a missing data problem there are only 13 pairs of twins (thus 26 children who make up a twin), but 31 positive indicators that a child is part of a twin. Hence there are at least 5 missing twins in the data and their outcome with respect to minor malformations is unknown. It's possible that the children expired shortly after birth, or were otherwise unrecorded. This complicates matters in that our data has some missing elements. If all the children were present, one could simply record the number of occurrences of birth defects with each mother, and model the data as $Y_h \sim \text{Binomial}(n_h, p_h)$. But since the data are missing, we have to use the nominal outcomes. That is, $Y_{hk} \sim \text{Bernoulli}(p_{hk})$, and try to estimate p_{hk}.

4 Methodology

We use three main methods to estimate the causal effects.

4.1 Estimation

4.1.1 Multivariate Generalized Estimating Equations

Generalized Estimating Equations (GEE) is a generalization of maximum-likelihood estimation, in which the data may be grouped into clusters as we have seen, and instead of specifying a specific probability distribution, we merely specify a link, variance, and "working correlation" matrix. Hence, GEE are often called "semi-parametric". We give a brief summary. For a good overview of GEE, see (Agresti, 2013) and (Pawitan, 2001).

Generalized Estimating Equations are closely linked with the idea of Maximum Likelihood and the exponential family of models. Given a sequence of observations $(y_i)_{i=1}^n$, for a sample size n, each observation y_i contributes to the log-likelihood of an assumed exponential family model. However, in GEE, we use a "quasi-likelihood", in which the only assumptions we make are the form of mean-function, the variance function, and the link function. That is, given an observation y_i, we assume a mean and variance model:

$$E[y_i] = \mu_i(\beta) = f(x_i'\beta) \tag{15}$$

$$\text{var}(y_i) = \phi v_i(\beta) \tag{16}$$

for known functions u_i and v_i, where ϕ is an unknown dispersion parameter, and β is an unknown regression parameter. Then, the best estimate of β is the solution of

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} V_i^{-1}(y_i - \mu_i) = 0.$$

In Zeger and Liang's seminal paper (1986, Biometrics), they showed that the above equation may actually be generalized. Suppose we have our data organized into $h = 1, 2, ..., H$ clusters, with units $k = 1, 2, ..., n_h$. Then we can take μ_i to be μ_{hk} to denote the mean function of the kth unit in the hth cluster. Now, given a mean model μ_{hk} for the hth cluster and the kth unit in the cluster, regression parameters β_j, and a variance structure R_h, we solve the *estimating equations*

$$\sum \frac{\partial \mu_{hk}}{\beta_j} R_h^{-1}\big(y_h - \mu_h(\beta)\big) = 0. \tag{17}$$

Here μ is a model for the conditional mean given in (6). In our case, μ corresponds to a regression with a logit-link function, but we make no distributional assumptions on y_{hk}. Typical Logistic Regression requires i.i.d. data, which is not present in our situation. Distributional concerns aside, GEE has the advantage of providing consistent estimates of $\boldsymbol{\beta}$, even when the variance structure R_h is misspecified. Hence it is often referred to as a "working correlation". Because our data most nearly satisfies an independent correlation structure, our default working correlation structure is "independence", which is known to produce efficient estimates even when misspecified. Due to the correlated nature of our data, all regression estimates are obtained via solving the GEE with a logit link. GEE estimates may be obtained in standard statistical software in R or Python.

4.1.2 Regression Adjustment

As shown in Section 2.4, it is sufficient to condition on a function of the covariates $\mathbf{X}$. In particular, the propensity score is of use. Suppose a propensity score $(\mathbf{X}_i)$ is obtained, for each i in the sample. Then an estimate of the treatment effect θ may be obtained by fitting the model

$$E[Y_i|Z_i, e(\mathbf{X}_i)] = \beta_0 + \beta_1 e(\mathbf{X}_i) + \theta Z_i. \tag{18}$$

This generalizes to the situation presented in (14) by obtaining an estimator of the propensity score, $\hat{e}(\mathbf{X}_i)$. It's also possible to use the Logit of the propensity score to recover the continuous linear predictor η_i:

$$\eta_i = \text{logit}\big\{e(\mathbf{X}_i)\big\}$$

then (16) becomes

$$E[Y_i|Z_i, \eta_i] = \beta_0^* + \beta_1^* \eta_i + \theta Z_i. \tag{19}$$

This is sometimes desirable, as η_i may have a nicer distribution closer to normality than the propensity score $e(X_i)$, which may be heavily skewed towards 0 or 1. This again generalizes quite easily to (14).

4.1.3 IPW Regression

Yet another option for estimating causal effects is the weighted regression model. Upon obtaining propensity scores $e_i = e(\mathbf{X}_i)$, we obtain *stabilized probability weights* by computing the following for each unit i in the sample:

$$w_i = \frac{Z_i Pr(Z=1)}{e_i} + \frac{(1-Z_i)Pr(Z=0)}{1-e_i}, \tag{20}$$

where $Pr(Z=1)$ and $Pr(Z=0)$ denote the marginal probability of treatment within the sample. The sequence of weights $(w_i)_{i=1}^n$ is then used to weight a regression model of the following form:

$$E[Y_i|Z_i] = \beta_0 + \theta Z_i. \tag{21}$$

This is essentially the regression analogue of the non-parametric Inverse Probability Weighting Estimator. Or a Horvitz-Thompson Estimator. In our case, we use a weighted GEE solution. While estimators may be directly weighted with the inverse of the propensity score, using the stabilized form in (18) can help mitigate the issues that arise when propensity scores are very close to 0 or 1, resulting in extremely large weights (Austin and Stuart, 2015).

4.2 Estimating Propensity Scores

In section 3.2.1, we noted the difficulty of estimation of the true propensity score model induced by some children being missing from the data. While we might believe that model (13) is what generated the propensity scores, the vector $\mathbf{X}_h$ has some missing values, and since we have no indicators on whether the twins are identical or merely fraternal, there is no way to impute the values without consulting the mothers themselves. Hence, instead of (13), we need to estimate

$$\text{logit}Pr(Z_{hk}=1|\mathbf{V}_h, \mathbf{X}^*{}_{hk}) = \alpha_0 + \mathbf{V}_h'\boldsymbol{\alpha}_1' + \alpha_2 \text{sex}_{hk} + \alpha_3 \text{twin}_{hk} \tag{22}$$

where $\mathbf{V}_h$ is a vector of confounders for the hth mother. Such a model is not exactly correct since it essentially models treatment assignment independently for each child, but with only 13 pairs of twins present in the data, and the rest singletons, the model is not far from the truth either. The data is clustered, but *approximately* independent. We hope that our estimates would be fairly efficient, and the estimated propensity scores $\hat{e}_{hk}$ not far from the truth of e_h.

4.3 Confounder Screening

We performed a univariate screening prior to estimating any causal effects or propensity scores in order to reduce the dimensionality of our data and thereby reduce the variance of our estimate. Exploratory analysis preceding this screening may be found in the Appendix in Figure 4. Screening is done in two ways. In the first case, for a given covariate X_{hk}, we estimate both

$$\text{logit}Pr(Z_{hk}=1|X_{hk}) = \alpha_0 + \alpha_1 X_{hk} \tag{23}$$

and

$$\text{logit}Pr(Y_{hk} = 1|X_{hk}) = \beta_0 + \beta_1 X_{hk}. \tag{24}$$

We select X_{hk} as a potential confounder if and only if $|\exp(\alpha_1) - 1| \geq 0.50$ *and* $|\exp(\beta_1) - 1| \geq 0.50$. This metric corresponds to the odds ratio being for X_{hk} being above 1.5 or below 0.50 in both treatment and outcome. In this case, X_{hk} has a statistical association with the treatment and outcome and thus may induce confounding.

In the second screening, we fit two outcome models:

$$\text{logit}Pr(Y_{hk} = 1|Z_{hk}) = \beta_0 + \theta_1 Z_{hk} \tag{25}$$

and

$$\text{logit}Pr(Y_{hk} = 1|X_{hk}, Z_{hk}) = \beta_0 + \beta_1 X_{hk} + \theta_2 Z_{hk} \tag{26}$$

where Z_{hk} is the realized binary treatment indicator from the sample. Thus we obtain two estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. We then select X_{hk} as a confounder if and only if

$$\left| \frac{\exp(\hat{\theta}_1) - \exp(\hat{\theta}_2)}{\exp(\hat{\theta}_1)} \right| \geq 0.10. \tag{27}$$

4.4 Results

Table of estimates of θ		1st Screening Method			2nd Screening Method		
Model Type		Estimate	95% CI (upper)	95% CI (lower)	Estimate	95% CI (lower)	95% CI (upper)
Not Including Sex	Multivariate	-0.590	-1.21	0.03	-0.556	-1.20	0.09
	Reg. Adjust.	-0.548	-1.18	0.09	-0.556	-1.19	0.08
	IPW	-0.569	-1.20	0.06	-0.466	-1.11	0.19
Including Sex	Multivariate	-0.666	-1.30	-0.03	-0.608	-1.25	0.03
	Reg. Adjust.	-0.636	-1.28	0.013	-0.608	-1.26	0.05
	IPW	-0.633	-1.27	0.01	-0.605	-1.24	0.03

Table 1: Results from the different estimating procedures. All confidence intervals are from the Robust Sandwich Estimator. Estimates are in log-odds.

In the first method, we obtain the 8 potential confounders, which include indicators for in-vitro-fertilization, maternal age, history of birth defects, referral source, and infant sex, as well as some indicators for pre-natal care. In the second method, we obtain only 4 potential confounders, 3 of which are indicators for presence of rheumatoid arthritis and its severity level. The 4th is a covariate for year of enrollment into the study. Curiously, the indicator for twins did not make it past screening. The indicators for rheumatoid arthritis are all redundant for each other, and speculated to be post-treatment effects (since the drug

exposure was intended to treat auto-immune disorders). Since it's not appropriate to control for post-treatment effects, only the indicator for enrollment year makes it into the model. Thus, we have two sets of confounders from which we can use to build a model to estimate treatment effects. In the first screening method, we have 8 potential confounders, in the second screening method, we have only one potential confounder: the year of enrollment into the study.

Using the two sets of confounders, we estimate the treatment effect in the form of a log-odds ratio via the estimation methods in section 4.1. Since the *infant's sex* is thought to be a potential confounder by the medical researchers of Rady's Children's Hospital from whom we obtained the data, we estimate treatment effect in two ways for each set of potential confounders. First, we use the confounders without the infant sex included, and estimate the treatment effect with each estimation method in section 4.1. Then we add in infant sex to each set of confounders, and measure the treatment effect again. This procedure obtains 12 estimands for the treatment effect, which are displayed in Table 1. Overall, simple estimations of the treatment effect via an ordinary multivariate GEE model had the largest changes with respect to model choice. Regression Adjustment stayed fairly similar under different choices. Weighted regression proved sensitive to modeling choices as well - in particular to method of screening.

We find that across modeling and screening choices, the treatment effects are all estimated to be fairly similar. With respect to the first screening method, the three models estimate an average treatment effect (on the log-odds scale) of -0.569 when sex was not included in the model. The average estimate decreased to -0.645 when infant sex was included into the there models. With respect to the second screening method, the three models had an average treatment effect estimate of -0.526 when sex was not included into the model. However, note that the IPW model had a fairly different estimate of -0.466 when compared to the other two. On the other hand, when infant sex was included into the models, we estimated an average treatment effect of -0.607, and each model performed similarly. The confidence intervals were estimated with the Robust Sandwich Estimator and all performed similarly, with lower bounds between -1.30 and -1.18, and upper bounds between 0.03 to 0.19. To understand the clinical significance of these bounds, note that a treatment effect of -1.30 on the log odds scale produces an odds ratio of approximately 0.27. This corresponds to those having received treatment being 83% less likely to see a minor malformation, and would be an extremely large effect size. On the other hand, a log odds closer to the average upper bound of 0.10 corresponds to an odds ratio of 1.10, which implies a 10% increase in the odds of experiencing a minor malformation in the treatment group.

Of course, the estimated treatment effects near the center of the confidence intervals are the most likely outcome given this data, which are close to -0.60 on the log odds scale. This corresponds to an odds ratio of roughly 0.55, which implies that the exposed are 45% less likely to have a minor malformation. This result indicates that treatment has a mild protective effect, but more data is likely needed, as these confidence intervals are fairly wide in size, an thus present a great deal of uncertainty.

5 Simulation

Since our models are only approximately correct, to explore the consequences of our modeling choices, we performed a large simulation study. Our interest is in assessing the bias, power, and efficiency of each of 3 estimators given in section 4. We also observe confidence interval coverage.

5.1 Simulation Design

We assume the models (13) and (14) for generating the propensity score and the outcome. To reiterate, we assume that for the hth mother

$$\text{logit}Pr(Z_h = 1|V_h, X_h) = \gamma_0 + \gamma_1 V_h + \mathbf{X}_h'\boldsymbol{\alpha} \tag{28}$$

gives the group-level propensity score, e_h. Here $\mathbf{X}_h = (X_{1h}, X_{2h}, X_{3h}, X_{4h}, X_{5h})'$, where X_{jh} indicates the type of pregnancy the hth mother has. For the hth mother, $X_{1h} = 1$ if she is pregant with a single boy, $X_{2h} = 2$ if she is pregnant with a single girl, $X_{3h} = 1$ if she is pregnant with twin boys, $X_{4h} = 1$ if she is pregnant with twin girls and $X_{5h} = 1$ if she is pregnant with mixed sex twins. Note only one of the X_{jh} may equal 1. Then we assume the probability of the outcomes for the kth infant with the hth mother is generated by

$$\text{logit}Pr(Y_{hk} = 1|V_h, \text{sex}_{hk}, \text{twin}_{hk}) = \beta_0 + \beta_1 V_h + \beta_2 \text{sex}_{hk} + \beta_3 \text{twin}_{hk} + \theta Z_h. \tag{29}$$

We let $V_h \sim N(0, \frac{1}{4})$ and we generate $\mathbf{X}_h$ by drawing a uniform random variable U and selecting cutoffs so that with 0.95 probability, the hth mother is having a singleton child (boy or girl, with equal probability) and with 0.025 probability she is having twins of mixed sex, and with 0.0125 probability she is having twin boys, or twin girls. We fix $\gamma_0 = -1$, $\gamma_1 = -0.75$. In order to get a sense of how our treatment effect estimates vary with changes in the effect size of the propensity score model, we set $\boldsymbol{\alpha}_1 = (0.80, 0.55, 1.33, 0.250, 1)'$ and then cycle the entries in $\boldsymbol{\alpha}_1$ in order to create new $\boldsymbol{\alpha}_j$, $j = 2, 3, 4$ so that $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_j$. That is, we generate

$$\boldsymbol{\alpha}_1 = (0.80, 0.55, 1.33, 0.250, 1)' \tag{30}$$
$$\boldsymbol{\alpha}_2 = (0.55, 1.33, 0.250, 1, 0.80)' \tag{31}$$
$$\boldsymbol{\alpha}_3 = (1.33, 0.250, 1, 0.80, 0.55)' \tag{32}$$
$$\boldsymbol{\alpha}_4 = (0.250, 1, 0.80, 0.55, 1.33)' \tag{33}$$

and fit a propensity score model as in (28) using $\gamma_0 = -1$, $\gamma_1 = -0.75$ and each of the $\boldsymbol{\alpha}_i$'s, $i = 1, 2, 3, 4$. The effect sizes in the vector $\boldsymbol{\alpha}$ are admittedly arbitrary, but range between small (0.250) to very large (1.33), so cycling them should produce a diverse set of datasets.

For the outcomes, we let the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ take on 4 different sets of values to observe the their role in the estimation of treatment effects when the outcome model was specified correctly and misspecified. Keeping β_0 and β_1 fixed, we simulated the response

level data using the following vectors of coefficients.

$$\begin{aligned}\boldsymbol{\beta}_1 &= (-0.50, -2, 0, 1)' \\ \boldsymbol{\beta}_2 &= (-0.50, -2, 1, 0)' \\ \boldsymbol{\beta}_3 &= (-0.50, -2, 1, 1)' \\ \boldsymbol{\beta}_4 &= (-0.50, -2, 0, 0)'.\end{aligned}$$

Here $\boldsymbol{\beta}_4$ corresponds to a null-effect of the sex and category of twin status, since the coefficient of sex and the category of twin status are 0. We are interested in the case where a model is fit with extra variables, so we include these terms. We also let the treatment effect θ take on 0, 0.50, 1.0 and 1.5. Naturally $\theta = 0$ corresponds to a null treatment effect.

Using each θ, $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_j$, we generate approximately 400 individual data points to form a data set. We repeat this 300 times. Since we have 4 of each θ, $\boldsymbol{\alpha}_i$, and $\boldsymbol{\beta}_j$, we end up with a total of $4 * 4 * 4 * 300 = 19200$ total data sets. On each of these data sets, we estimate the treatment effect θ using a multivariate model, a regression adjusted model, and an IPW regression model. For each of the 300 data sets with a fixed θ, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$, we record the bias of the estimate $\hat{\theta}$ from each model, and average the bias over the 300 data sets by computing $n^{-1} \sum_{i=1}^{n} (\hat{\theta}_i) - \theta$, where $n = 300$. We also record the variance, and estimate the power (or type I error), as well as the confidence interval coverage.

From now on, references to $\boldsymbol{\beta}_j$ refer to the vectors present above.

5.2 Simulation Study Results

The majority of the results of the simulation study may be found in the Appendix. We provide a brief summary here. All regression models were fit by solving the Generalized Estimating Equation with an independent working correlation matrix in Python using the StatsModels package.

Under the null hypothesis, $\theta = 0$, and correct model specification (inclusion of the simulated infant sex indicator) all methods performed similarly with respect to bias, variance, MSE and coverage, for each of the $\boldsymbol{\alpha}_i$'s and the $\boldsymbol{\beta}_j$'s However, each method presented confidence intervals slightly too wide, with coverage proportions between 0.96 to 0.97. Type-I error rates were between 0.03 and 0.04. Bias was nearly perfect in each of the models.

When the model was misspecified, each model performed similarly, but bias increased the most with $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$, with the largest amount of bias occurring in the latter case in the misspecified benchmark model. The bias positive and still relatively small, ranging from 0 to 0.05. The misspecified models slightly undercovered with $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ but were slightly too wide with $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_4$.

When $\theta = 0.50$, the treatment effect is moderate, on the scale that we actually observe in our real data. When the model was correctly specified, all approaches had low levels of average bias, in the neighborhood of -0.03 to 0.01, with multivariate GEE performing the best in terms of bias, power, and confidence interval coverage. This is likely because GEE is closest to the true data generating process, so it might have had an advantage.

When the model was misspecified, bias ranged from very small, such as 0.002 in the multivariate GEE with $\boldsymbol{\beta}_1$ to extremely large relative to $\theta = 0.50$, such as the bias of -0.208 in the IPW regression model under $\boldsymbol{\beta}_3$. This indicates a sudden sensitivity to modeling choices, under a permutation of the entries in $\boldsymbol{\alpha}$. This seems to make sense, as the propensity scores are directly utilized in an IPW model, where it has been shown they can cause issue if they are too close to 0 or 1.

When $\theta = 1.0$, the treatment effect is demonstrably large, and all modeling choices saw power converging to 1, and performed almost exactly the same, though it should be noted that with $\boldsymbol{\beta}_3$, each of the each model's confidence interval coverage dropped to nearly 80%, and had large biases of around -0.200. While relatively small compared to the effect size of $\theta = 1$, this should not be ignored.

With $\theta = 1.5$, the treatment effect is enormous, and so the power for each model reached 1.0, even when misspecified. Each model appeared to perform similarly, with good confidence interval coverage, and low bias, except in the problematic case of $\boldsymbol{\beta}_j$, $j = 2, 3$. Here, each model performed dreadfully, with multivariate GEE tending to be slightly better in terms of bias and confidence interval coverage.

6 Discussion

Upon consideration of the results of our simulation, multivariate models and regression adjustment models tend to perform well, with the IPW models tending to be more sensitive to modeling choices. It should be noted that the multivariate GEE proved to be fairly robust to modeling choices, even when misspecified, at least relative to the other two methods of regression adjustment and IPW. The average bias in the simulation was fairly low, and confidence intervals fairly appropriate. With respect to our real data, this gives us some willingness to believe that our estimate of the treatment effect is fairly robust to modeling choices, especially since the treatment effect estimate was not particularly sensitive to choices of model (except IPW, which did not perform well in our simulation). Those in our cohort who were exposed to the drug of interest may see a mild protective effect against minor malformations. That said, the real data set presented a unique challenge and situation, and one must always take time to consider what assumptions are being involved in the modeling process. E.g., an assumption of independence was incorrect for our real data, but is approximately correct when one considers that nearly 90% of the data were independent samples.

Undeniably, when it comes to the art and craft of causal inference, model specification and its related assumptions are a difficult aspect of the data analysis to get correct. In order to correctly specify a causal model, all relevant confounders must be accounted for. However, due to the bias-variance trade-off, one cannot simply fit a causal model on each covariate present in the data set. Doing so needlessly increases the variance of the treatment effect estimates (potentially rending them inactionable, if the resulting confidence intervals are too wide to be of use), and may open so-called "backdoor paths" which may actually *induce* confounding into the estimator.

As a work-around, one needs to rely on statistical association in order to prioritize

confounders by the strength of their statistical association with the outcome and treatment. Moreover, domain expertise is required in order to determine the exact role of the potential confounder in the data. It is not appropriate, for example, to control confounders that are considered "post-treatment effects" (Pearl, 2009).

This thesis has served as a cursory summary of some of the theory and methods of causal inference. We applied these methods to a real data set, as well as verified our results via simulation. In general, we feel that fake-data simulation is a useful tool to sanity check data modeling choices when there is no clear theory to guide practitioners, as in our situation of approximately independent data. We believe our simulation would be improved by considering a wider variety of choices for the $\boldsymbol{\alpha}_i$'s and $\boldsymbol{\beta}_j$'s, and by perhaps manipulating the proportion of twins present in the data set.

We also believe it may be worth exploring the drug's effects on the response of a *major malformation*, in conjunction with the *minor malformations* and contrasting the two subpopulations. This idea lends itself well to a larger replication study.

Acknowledgements

I would like to thank Professor Ronghui 'Lily' Xu for serving as my thesis advisor. None of this work would be possible without her patient and clear guidance. Thank you for taking a chance on me.

References

[1] Holland, Paul W. (1986). "Statistics and Causal Inference". J. Amer. Statist. Assoc. 81 (396): 945–960. doi:10.1080/01621459.1986.10478354. JSTOR 2289064.

[2] VanderWeele, Tyler J., 2013, *On the Definition of a Confounder*, NIH Public Access, Tyler J. VanderWeele

[3] Hernán MA, Robins JM (2019). Causal Inference. Boca Raton: Chapman Hall/CRC, forthcoming.

[4] Imbens, G., Rubin, D. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139025751

[5] Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. Stat Med. 2013;32(19):3373–3387. doi:10.1002/sim.5786

[6] Dorie V, Harada M, Carnegie NB, Hill J. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. Stat Med. 2016;35(20):3453–3470. doi:10.1002/sim.6973

[7] Judea Pearl. 2009. Causality: Models, Reasoning and Inference (2nd ed.). Cambridge University Press, New York, NY, USA.

[8] PAUL R. ROSENBAUM, DONALD B. RUBIN, The central role of the propensity score in observational studies for causal effects, Biometrika, Volume 70, Issue 1, April 1983, Pages 41–55, https://doi.org/10.1093/biomet/70.1.41

[9] AAP News and Journals, *https://neoreviews.aappublications.org/content/4/4/e99?download=true*, accessed February 2019.

[10] Agresti, A. (2013) Categorical Data Analysis. 3rd Edition, John Wiley Sons Inc., Hoboken.

[11] Austin, P. C., and Stuart, E. A. (2015) Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statist. Med., 34: 3661– 3679. doi: 10.1002/sim.6607.

[12] Pawitan, Yudi. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford: Clarendon Press, 2001. Print.

[13] Zeger, Scott L., and Kung-Yee Liang. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." Biometrics, vol. 42, no. 1, 1986, pp. 121–130. JSTOR, www.jstor.org/stable/2531248.

Appendix

The astute reader may notice that with 4 choices for $\boldsymbol{\alpha}$ and 4 choices for $\boldsymbol{\beta}$ in the simulation study, we should have 16 tables of the style given below, but only 8 are present. 8 Tables have been omitted, mostly as a desire to save paper and considerably shorten the length of this thesis. Moreover, the results in the remaining tables do not vary greatly from those already presented. The inquisitive reader may email the author at the email given in the title in order to see the remaining tables.

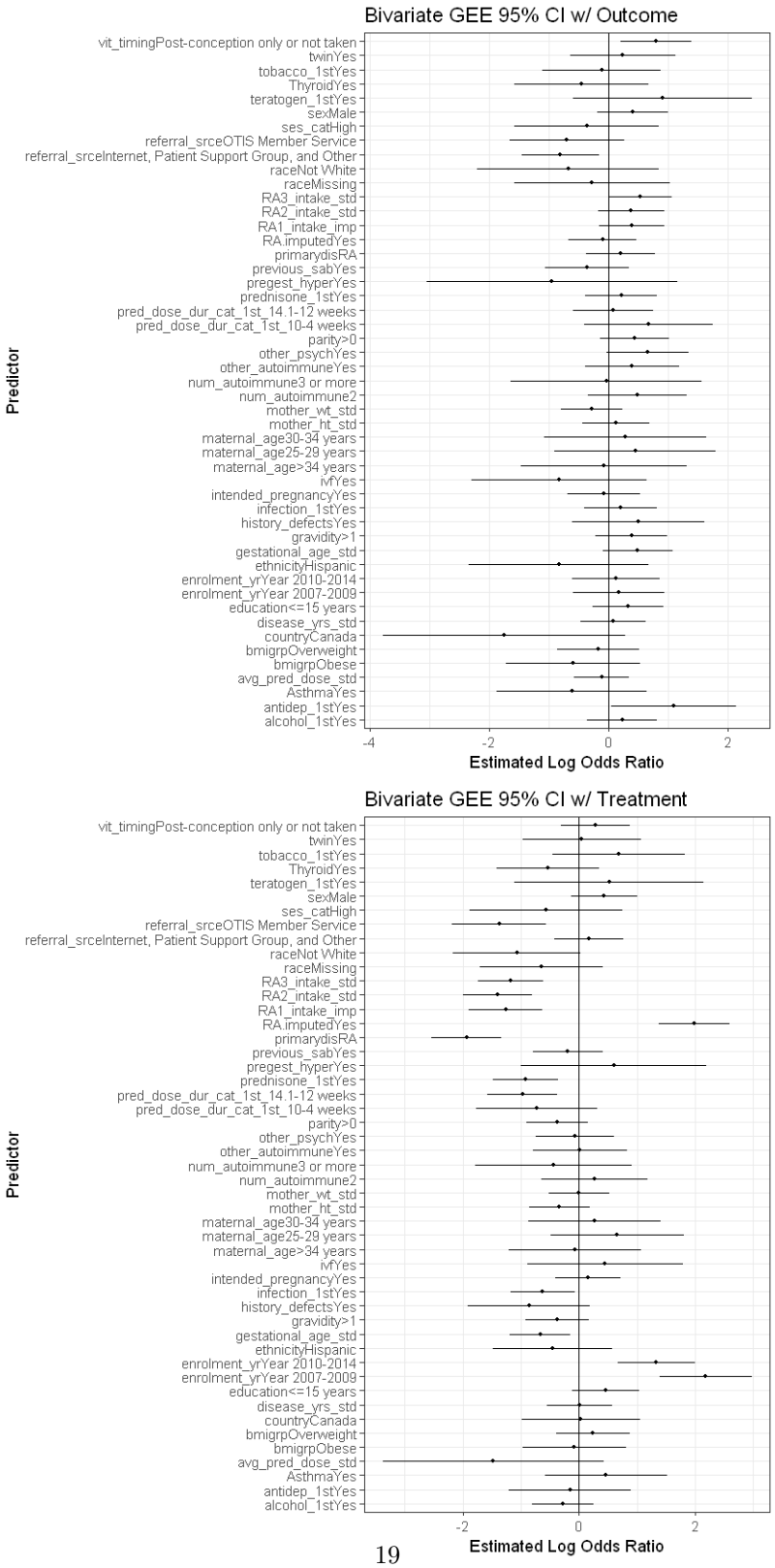

Figure 2: Forest plots for the Univarite Screening. Each predictor was fit by itself with an intercept on the outcome (top) and the treatment (bottom). Points give the point estimate, and bands give the 95% confidence intervals.

$\theta = 0$		Individual Sex Included					Individual Sex Not Included				
α_1	Model Type	Bias	Variance	MSE	Coverage	Type I	Bias	Variance	MSE	Coverage	Type I
$\boldsymbol{\beta}_1$	Multivariate	0.001	0.049	0.049	0.95	0.05	-0.0	0.049	0.049	0.953	0.047
	Reg. Adj.	0.001	0.046	0.046	0.95	0.05	-0.0	0.047	0.047	0.953	0.047
	IPW	-0.004	0.042	0.042	0.96	0.04	0.001	0.041	0.041	0.96	0.04
$\boldsymbol{\beta}_2$	Multivariate	0.017	0.04	0.04	0.953	0.047	0.085	0.038	0.046	0.933	0.067
	Reg. Adj.	0.016	0.036	0.037	0.957	0.043	0.083	0.037	0.044	0.937	0.063
	IPW	0.018	0.032	0.032	0.97	0.03	0.081	0.035	0.041	0.943	0.057
$\boldsymbol{\beta}_3$	Multivariate	-0.004	0.045	0.045	0.957	0.043	0.053	0.043	0.045	0.94	0.06
	Reg. Adj	-0.004	0.041	0.042	0.957	0.043	0.052	0.041	0.044	0.94	0.06
	IPW	-0.006	0.035	0.035	0.963	0.037	0.05	0.037	0.039	0.957	0.043
$\boldsymbol{\beta}_4$	Multivariate	-0.0	0.042	0.042	0.963	0.037	-0.002	0.042	0.042	0.967	0.033
	Reg. Adj.	-0.0	0.039	0.039	0.97	0.03	-0.002	0.04	0.04	0.967	0.033
	IPW	-0.0	0.037	0.037	0.97	0.03	-0.002	0.038	0.038	0.97	0.03

Table 2: Table of simulation results, under $\theta = 0$ and $\boldsymbol{\alpha}_1$ as described in the Section 5. Under these conditions, all models perform fairly well, with low bias, variance and MSE across the board.

$\theta = 0.50$		Individual Sex Included					Individual Sex Not Included				
α_1	Model Type	Bias	Variance	MSE	Coverage	Power	Bias	Variance	MSE	Coverage	Power
$\boldsymbol{\beta}_1$	Multivariate	0.016	0.047	0.047	0.943	0.663	0.013	0.046	0.046	0.943	0.663
	Reg.Adj.	-0.006	0.043	0.043	0.953	0.64	0.001	0.044	0.044	0.95	0.653
	IPW	-0.03	0.041	0.042	0.95	0.61	-0.025	0.04	0.041	0.953	0.62
$\boldsymbol{\beta}_2$	Multivariate	0.023	0.052	0.052	0.923	0.667	0.061	0.048	0.052	0.92	0.77
	Reg.Adj.	-0.0	0.048	0.048	0.93	0.663	0.05	0.046	0.048	0.923	0.75
	IPW	-0.028	0.042	0.043	0.943	0.633	0.033	0.044	0.045	0.937	0.737
$\boldsymbol{\beta}_3$	Multivariate	0.007	0.053	0.053	0.94	0.613	0.042	0.051	0.053	0.94	0.7
	Reg.Adj.	-0.016	0.048	0.049	0.943	0.593	0.031	0.049	0.05	0.95	0.683
	IPW	-0.058	0.042	0.045	0.947	0.543	0.004	0.045	0.045	0.95	0.657
$\boldsymbol{\beta}_4$	Multivariate	0.022	0.044	0.045	0.95	0.7	0.02	0.043	0.044	0.95	0.693
	Reg.Adj.	0.003	0.04	0.04	0.96	0.683	0.007	0.04	0.04	0.96	0.687
	IPW	-0.008	0.039	0.039	0.95	0.673	-0.009	0.039	0.039	0.957	0.673

Table 3: Simulation results with $\theta = 0.50$ and $\boldsymbol{\alpha}_1$. All models perform similarly as when $\theta = 0$.

$\theta = 1.0$		Individual Sex Included					Individual Sex Not Included				
α_1		Bias	Variance	MSE	Coverage	Power	Bias	Variance	MSE	Coverage	Power
$\boldsymbol{\beta}_1$	Multivariate	0.001	0.045	0.045	0.97	1.0	-0.003	0.045	0.045	0.973	1.0
	Reg.Adj.	-0.039	0.041	0.043	0.967	1.0	-0.025	0.043	0.044	0.97	1.0
	IPW	-0.084	0.039	0.046	0.957	1.0	-0.08	0.039	0.045	0.96	1.0
$\boldsymbol{\beta}_2$	Multivariate	0.007	0.051	0.051	0.947	0.99	0.011	0.046	0.047	0.957	0.993
	Reg.Adj.	-0.04	0.045	0.047	0.943	0.987	-0.009	0.044	0.045	0.947	0.993
	IPW	-0.096	0.041	0.05	0.94	0.983	-0.04	0.042	0.044	0.947	0.993
$\boldsymbol{\beta}_3$	Multivariate	0.02	0.048	0.048	0.957	0.99	0.023	0.047	0.047	0.97	0.993
	Reg.Adj.	-0.026	0.045	0.046	0.96	0.987	0.005	0.045	0.045	0.967	0.993
	IPW	-0.1	0.038	0.048	0.933	0.983	-0.04	0.041	0.043	0.957	0.99
$\boldsymbol{\beta}_4$	Multivariate	0.024	0.052	0.053	0.917	1.0	0.02	0.051	0.051	0.91	1.0
	Reg.Adj.	-0.013	0.048	0.048	0.93	1.0	-0.004	0.049	0.049	0.923	1.0
	IPW	-0.038	0.048	0.049	0.933	1.0	-0.038	0.047	0.049	0.933	1.0

Table 4: Simulation results when $\theta = 1.0$ and with $\boldsymbol{\alpha}_1$. Power is converging to 1 now, and each model does well, though the correctly specified multivariate model is among the best.

$\theta = 1.5$		Individual Sex Included					Individual Sex Not Included				
α_1		Bias	Variance	MSE	Coverage	Power	Bias	Variance	MSE	Coverage	Power
β_1	Multivariate	0.023	0.055	0.055	0.953	1.0	0.019	0.055	0.055	0.947	1.0
	Reg.Adj.	-0.033	0.051	0.052	0.94	1.0	-0.014	0.053	0.053	0.94	1.0
	IPW	-0.096	0.047	0.057	0.913	1.0	-0.091	0.047	0.055	0.93	1.0
β_2	Multivariate	0.028	0.07	0.071	0.943	1.0	0.013	0.065	0.065	0.937	1.0
	Reg. Adj.	-0.039	0.064	0.066	0.947	1.0	-0.016	0.062	0.062	0.943	1.0
	IPW	-0.118	0.059	0.073	0.92	1.0	-0.059	0.06	0.063	0.95	1.0
β_3	Multivariate	0.008	0.06	0.06	0.953	1.0	-0.005	0.059	0.059	0.937	1.0
	Reg. Adj.	-0.052	0.057	0.059	0.927	1.0	-0.03	0.056	0.057	0.933	1.0
	IPW	-0.158	0.051	0.075	0.883	1.0	-0.092	0.054	0.062	0.923	1.0
β_4	Multivariate	0.042	0.047	0.049	0.95	1.0	0.04	0.047	0.049	0.95	1.0
	Reg. Adj.	-0.011	0.046	0.046	0.95	1.0	0.005	0.048	0.048	0.94	1.0
	IPW	-0.051	0.042	0.045	0.95	1.0	-0.049	0.043	0.045	0.947	1.0

Table 5: Simulation results with $\theta = 1.5$ and $\boldsymbol{\alpha}_1$. Power is 1 now for each method, as the effect size of 1.5 on the log-odds scale is enormous. Bias is slightly larger, but still very small in comparison with the effect size of θ.

$\theta = 0$		Individual Sex Sex Included					Individual Not Included				
α_2	Model Type	Bias	Variance	MSE	Coverage	Power	Bias	Variance	MSE	Coverage	Power
$\boldsymbol{\beta}_1$	Multivariate	0.008	0.045	0.045	0.95	0.05	0.009	0.044	0.044	0.943	0.057
	Reg. Adj.	0.009	0.039	0.039	0.95	0.05	0.011	0.039	0.039	0.95	0.05
	IPW	0.024	0.039	0.039	0.953	0.047	0.008	0.038	0.038	0.953	0.047
$\boldsymbol{\beta}_2$	Multivariate	0.016	0.048	0.048	0.943	0.057	-0.164	0.045	0.072	0.86	0.14
	Reg. Adj.	0.011	0.039	0.039	0.96	0.04	-0.158	0.043	0.068	0.863	0.137
	IPW	0.007	0.039	0.039	0.957	0.043	-0.154	0.04	0.064	0.867	0.133
$\boldsymbol{\beta}_3$	Multivariate	-0.008	0.043	0.043	0.953	0.047	-0.186	0.039	0.074	0.863	0.137
	Reg. Adj.	-0.011	0.035	0.035	0.973	0.027	-0.176	0.036	0.067	0.87	0.13
	IPW	0.001	0.034	0.034	0.977	0.023	-0.171	0.034	0.063	0.88	0.12
$\boldsymbol{\beta}_4$	Multivariate	-0.008	0.053	0.053	0.927	0.073	-0.004	0.049	0.049	0.93	0.07
	Reg. Adj.	-0.009	0.048	0.048	0.93	0.07	-0.003	0.047	0.047	0.937	0.063
	IPW	-0.008	0.048	0.048	0.927	0.073	-0.004	0.044	0.044	0.947	0.053

Table 6: Simulation results with $\theta = 0$ and $\boldsymbol{\alpha}_2$. Similar results as with $\boldsymbol{\alpha}_1$.

$\theta = 0.50$		Individual Sex Included					Individual Sex Not Included				
α_2	Model Type	Bias	Variance	MSE	Coverage	Power	Bias	Variance	MSE	Coverage	Power
$\boldsymbol{\beta}_1$	Multivariate	0.001	0.038	0.038	0.973	0.663	0.002	0.036	0.036	0.98	0.68
	Reg. Adj.	-0.032	0.032	0.033	0.977	0.64	-0.02	0.032	0.032	0.987	0.66
	IPW	-0.022	0.032	0.033	0.987	0.65	-0.036	0.03	0.032	0.98	0.643
$\boldsymbol{\beta}_2$	Multivariate	0.007	0.048	0.048	0.953	0.64	-0.194	0.041	0.079	0.827	0.357
	Reg. Adj.	-0.045	0.04	0.042	0.953	0.597	-0.198	0.04	0.079	0.823	0.347
	IPW	-0.054	0.038	0.041	0.963	0.58	-0.208	0.037	0.08	0.813	0.333
$\boldsymbol{\beta}_3$	Multivariate	0.01	0.053	0.053	0.92	0.673	-0.198	0.048	0.087	0.813	0.303
	Reg. Adj.	-0.051	0.042	0.045	0.937	0.617	-0.21	0.043	0.087	0.803	0.287
	IPW.	-0.048	0.041	0.043	0.94	0.63	-0.218	0.041	0.089	0.793	0.277
$\boldsymbol{\beta}_4$	Multivariate	0.003	0.046	0.046	0.947	0.687	0.002	0.043	0.043	0.95	0.7
	Reg. Adj.	-0.022	0.041	0.042	0.943	0.66	-0.01	0.041	0.041	0.95	0.69
	IPW	-0.027	0.041	0.041	0.95	0.653	-0.027	0.038	0.039	0.95	0.663

Table 7: Simulation results with $\theta = 0.50$ and $\boldsymbol{\alpha}_2$. While initially identical to the case of $\boldsymbol{\alpha}_1$, note the large bias in the misspecified models, reaching nearly -0.20. This is nearly half the size of θ, indicating that the model choices are sensitive to misspecification.

$\theta = 1.0$		Individual Sex Included					Individual Sex Not included				
α_2	Model Type	Bias1	Variance1	MSE1	Coverage1	Power1	Bias2	Variance2	MSE2	Coverage2	Power2
$\boldsymbol{\beta}_1$	Multivariate	0.007	0.049	0.049	0.95	0.997	0.004	0.046	0.046	0.953	0.997
	Reg. Adj.	-0.058	0.042	0.046	0.947	0.997	-0.041	0.044	0.045	0.943	0.997
	IPW.	-0.055	0.043	0.046	0.947	0.997	-0.071	0.04	0.046	0.947	0.997
$\boldsymbol{\beta}_2$	Multivariate	0.02	0.054	0.054	0.95	0.993	-0.215	0.046	0.092	0.83	0.957
	Reg. Adj.	-0.083	0.042	0.049	0.93	0.99	-0.232	0.043	0.097	0.813	0.957
	IPW.	-0.097	0.041	0.051	0.93	0.987	-0.256	0.041	0.107	0.787	0.953
$\boldsymbol{\beta}_3$	Multivariate	0.018	0.055	0.056	0.943	0.99	-0.219	0.048	0.096	0.813	0.96
	Reg. Adj.	-0.087	0.044	0.052	0.953	0.99	-0.245	0.045	0.105	0.783	0.96
	IPW.	-0.095	0.043	0.052	0.947	0.99	-0.267	0.042	0.113	0.757	0.957
$\boldsymbol{\beta}_4$	Multivariate	0.033	0.048	0.049	0.937	0.997	0.031	0.047	0.048	0.93	1.0
	Reg. Adj.	-0.017	0.044	0.044	0.94	0.997	0.006	0.045	0.045	0.94	1.0
	IPW.	-0.03	0.043	0.044	0.94	0.997	-0.029	0.042	0.043	0.93	1.0

Table 8: Simulation results when $\theta = 1.0$ and $\boldsymbol{\alpha}_2$. Note the models perform well, except in the case of misspecification and $\boldsymbol{\beta}_j$, $j = 2, 3$.

$\theta = 1.5$		Individual Sex Included					Individual Sex Not Included				
α_2		Bias1	Variance1	MSE1	Coverage1	Power1	Bias2	Variance2	MSE2	Coverage2	Power2
$\boldsymbol{\beta}_1$	Multivariate	0.023	0.061	0.061	0.947	1.0	0.022	0.056	0.057	0.937	1.0
	Reg. Adj.	-0.072	0.052	0.057	0.917	1.0	-0.043	0.051	0.053	0.94	1.0
	IPW	-0.075	0.051	0.057	0.927	1.0	-0.088	0.048	0.056	0.93	1.0
$\boldsymbol{\beta}_2$	Multivariate	0.022	0.067	0.067	0.967	1.0	-0.229	0.057	0.11	0.807	1.0
	Reg. Adj.	-0.115	0.058	0.071	0.9	1.0	-0.255	0.054	0.119	0.767	1.0
	IPW.	-0.14	0.053	0.073	0.893	1.0	-0.294	0.054	0.14	0.72	1.0
$\boldsymbol{\beta}_3$	Multivariate	0.014	0.063	0.064	0.953	1.0	-0.244	0.055	0.114	0.79	1.0
	Reg. Adj.	-0.128	0.052	0.068	0.91	1.0	-0.29	0.049	0.133	0.727	1.0
	IPW	-0.148	0.048	0.07	0.91	1.0	-0.32	0.048	0.151	0.67	1.0
$\boldsymbol{\beta}_4$	Multivariate	0.028	0.05	0.051	0.957	1.0	0.023	0.048	0.049	0.953	1.0
	Reg. Adj.	-0.043	0.045	0.047	0.93	1.0	-0.01	0.046	0.046	0.943	1.0
	IPW	-0.064	0.046	0.05	0.933	1.0	-0.063	0.044	0.048	0.93	1.0

Table 9: Simulation results when $\theta = 1.5$ and $\boldsymbol{\alpha}_2$. When misspecified, bias gets as large as -0.32 in the IPW model.