Nonparametric identification of causal effects in clustered observational studies with differential selection

Ting Ye[1], Ted Westling[2], Lindsay Page[3], and Luke Keele [4]

Abstract

The clustered observational study (COS) design is the observational study counterpart to the clustered randomized trial. In a COS, a treatment is assigned to intact groups, and all units within the group are exposed to the treatment. However, the treatment is non-randomly assigned. COSs are common in both education and health services research. In education, treatments may be given to all students within some schools but withheld from all students in other schools. In health studies, treatments may be applied to clusters such as hospitals or groups of patients treated by the same physician. In this manuscript, we study the identification of causal effects in clustered observational study designs. We focus on the prospect of differential selection of units to clusters, which occurs when the units' cluster selections depend on the clusters' treatment assignments. Extant work on COSs has made an implicit assumption that rules out the presence of differential selection. We derive the identification results for designs with differential selection and that contexts with differential cluster selection require different adjustment sets than standard designs. We outline estimators for designs with and without differential selection. Using a series of simulations, we outline the magnitude of the bias that can occur with differential selection. We then present two empirical applications focusing on the likelihood of differential selection.

Keywords: Causal inference, Identification, Clustered Observational Study

[1]University of Washington, Email: tingye1@uw.edu

[2]University of Massachusetts Amherst, Email: twestling@umass.edu

[3]Brown University, Email: lindsay_page@brown.edu

[4]University of Pennsylvania, Email: luke.keele@gmail.com.

1 Introduction

Many applied analyses focus on whether a treatment given to some set of units causes a hypothesized effect. In some settings, treatments of interest are allocated individually. That is, a treatment is assigned to certain people individually and not to others. However, in other settings, the treatment is allocated to groups or intact clusters of units – e.g., to hospitals or schools – while outcomes of interest are measured at the unit level— e.g., patients or students. The critical feature of such treatment assignment processes is that all or none of the units within a cluster are exposed to the treatment of interest. When group-level treatments are randomly assigned, the study design commonly is referred to as a clustered randomized trial (CRT) (Raudenbush, 1997; Hedges and Hedberg, 2007). In a clustered observational study (COS), treatment is still assigned at cluster level but assignment is non-random (Page et al., 2020). Given non-random assignment in a COS, differences in outcomes may reflect pretreatment differences in treated and control groups rather than actual treatment effects (Hansen et al., 2014). Moreover, the COS design requires specialized forms of statistical adjustment for observed confounders (Keele and Zubizarreta, 2017; Pimentel et al., 2018). Next, we highlight two areas of applied research where the COS design is common.

1.1 Education

COSs are common in educational research, where treatments are often applied to schools. For example, Adelson et al. (2012) study how changes in gifted programs in some schools affect student-level outcomes. Much research has focused on whether Catholic schools are more effective than public schools in fostering student achievement (Coleman et al., 1982; Hoffer et al., 1985; Coleman and Hoffer, 1987). Often a new reading program may be implemented in some schools but not in others (Page et al., 2020). Other interventions broadly seek to change entire school structures in hopes of improving chronically underperforming

schools (Bryk et al., 2015; Mehta et al., 2012; McGuinn, 2006). For example, in 2015-16, the Wake County Public School System implemented a school turnaround strategy known as the Elementary Support Model (ESM) in 12 selected schools. Schools in the ESM condition received a range of supports over three years, including governance reform, additional staffing, and instructional coaching (Paeplow et al., 2019). This program implementation is prototypical of the COS template; the intervention is non-randomly assigned to and delivered at the school level, but the investigators are focused on academic and behavioral outcomes measured at the student level.

1.2 Health Services Research

The COS design is also common in comparative effectiveness research (CER), which focuses on evaluating the causal effects of health care strategies on patient outcomes (Hernán, 2018). CER encompasses patient level treatments and the effects health care delivery, organization and financing, as well as public health interventions (Institute of Medicine, 2009). In CER, COS designs often mirror those in education where interventions are applied to entire hospitals. For example, the COS design has been applied at the hospital level to study the effect of the work environment and amount of autonomy given to nurses (Rao et al., 2017; Silber et al., 2016), the effect of medical residents' duty hours (Bilimoria et al., 2016; Patel et al., 2014; Silber et al., 2014), and the effect of Magnet certification for high quality nursing (Barnes et al., 2016; McHugh et al., 2013). The COS design in CER also arises at the level of the physician. In this setting, a treatment is applied to some physicians but not others, but outcomes are measured at the patient level. COS designs of this type include studies on the effects of training in a teaching hospital (Navathe et al., 2013a,b; Srinivas et al., 2013; Lorch et al., 2012) and the effect of a university-based surgical residency (Sellers et al., 2018).

1.3 Clustered Observational Studies

In both of these applied fields, a randomized trial would require randomly assigning entire clusters to treatment or control status. However, in educational and health settings, such randomization can often be impractical or even impossible, and so we require different analytic strategies to address questions of cause and effect. In such cases, critical research questions can only be investigated using clustered observational studies. The extant literature on the COS study design, however, critically assumes that the population of units within the clusters is not affected by treatment assignment (Keele and Zubizarreta, 2017; Pimentel et al., 2018; Hansen et al., 2014). Here, we focus on a feature of clustered treatment assignment that may result in a form of selection bias arising from differential selection of units – patients or students – across treated clusters. Specifically, we consider the possibility that the population of units within treated clusters changes in reaction to the treatment being assigned to the cluster. For example, once an intervention such as ESM is put into place, families with high achieving students may move to treated school catchment areas. Under this form of differential selection, differences in outcomes between treated and control students with similar pre-treatment characteristics may be a result of peer effects rather than the treatment effect, because schools with ESM may attract higher achieving students. That is, ESM may appear to have improved outcomes, when in fact it only attracted higher achieving students. Ogburn and VanderWeele (2014) refer to this as allocational interference.

In this manuscript, we present a series of new results for clustered observational studies. First, we develop a notational framework that allows for differential selection of units to clusters as a function of treatment assignment. We use the target trial framework to define three distinct trials that differ in terms of whether unit to cluster assignment is affected by the treatment. We highlight how extant COS research has been based on a hidden assumption that rules out the possibility of differential selection. We show that

identification for each target trial depends on conditioning on different types of covariates, to which we refer as the conditioning or adjustment set. We also outline the additional assumption that identifies causal effects under differential selection. We then conduct a simulation study investigating how bias can arise in the analysis of COS designs, when the wrong target trial is assumed. We present results from two different applications, one from education and one from health services research. Finally, we conclude with a discussion of key implications for applied research.

2 COS Framework

To understand issues of causal identification in this context, we use the target trial framework. Target trial emulation calls for applying design principles from randomized trials to the analysis of observational data (Hernán and Robins, 2016). More specifically, in the target trial framework, the investigator derives estimands and identification conditions from the hypothetical experimental trial that is being emulated. From a practical standpoint, the target trial of interest may be infeasible as an actual randomized trial. However, that is largely irrelevant for our technical, analytic purpose, which is to use the target trial to structure the design and analysis of an observational study. Next, we outline two hypothetical target trials that are relevant to COS designs.

2.1 Target Trials for the COS Design

Here, we provide an informal outline of two hypothetical target trials, where in both cases there are n units and m clusters. In both of these target trials, there are two stages of assignment, the order of which represent the key difference between these two target trials. In Figure 1, we provide a graphical illustration of these two target trials, with individual units represented by circles and groups or clusters represented by rectangles. In target trial

1, units are first assigned (randomly or not) to clusters and then clusters are randomly assigned to treatment or control. Target trial 1 also includes the setting where the unit-cluster pairing at the first stage is fixed, which is typical in the literature on CRTs. In target trial 2, on the other hand, we consider a scenario where at the first stage, m clusters are randomized to receive treatment or control, and at the second stage, n units are randomly assigned to the m clusters. This setting is also common in CRTs when units are recruited after clusters are randomized.

If the two stages of randomization are independent of each other—if the stage 2 randomization does not depend on the outcome of stage 1 assignments—we will show that the identification conditions for target trial 1 and 2 are equivalent. However, these two target trials differ if the treatment assignment depends on unit-cluster pairing in target trial 1, or if the unit-cluster pairing depends on clusters' treatment assignments in target trial 2. In particular, target trial 1 precludes the possibility of differential selection, but target trial 2 does not when unit selection is no longer randomized at the second stage. Next, we develop notation that allows us to formally describe these two target trials.

2.2 Notation

Let $\boldsymbol{A} = (A_1, \ldots, A_m)$ be the observed binary treatment indicators for all clusters, and let $\boldsymbol{J}(\boldsymbol{a}) = (J_1(\boldsymbol{a}), \ldots, J_n(\boldsymbol{a}))$ denote the clusters that the n subjects would be assigned to had the clusters received exposure: $\boldsymbol{a} = (a_1, \ldots, a_m)$. We use this potential-selections notation to capture the fact that the subject-cluster pairing may depend on treatment status of clusters, $\boldsymbol{A}$. We use $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{A})$ to denote the observed cluster values for the n subjects. Next, let $Y_i(\boldsymbol{a}, \boldsymbol{j})$ be the potential outcome for subject i had $\boldsymbol{A} = \boldsymbol{a}$ and $\boldsymbol{J} = \boldsymbol{j}$, where $\boldsymbol{j} = (j_1, \ldots, j_n)$. This notation allows for unit outcomes to depend on both treatment *and* cluster assignments. Finally, $Y_i = Y_i(\boldsymbol{A}, \boldsymbol{J})$ is the observed outcome for subject i.

For each unit i, we observe a vector of baseline covariates X_i that describe the units (age,

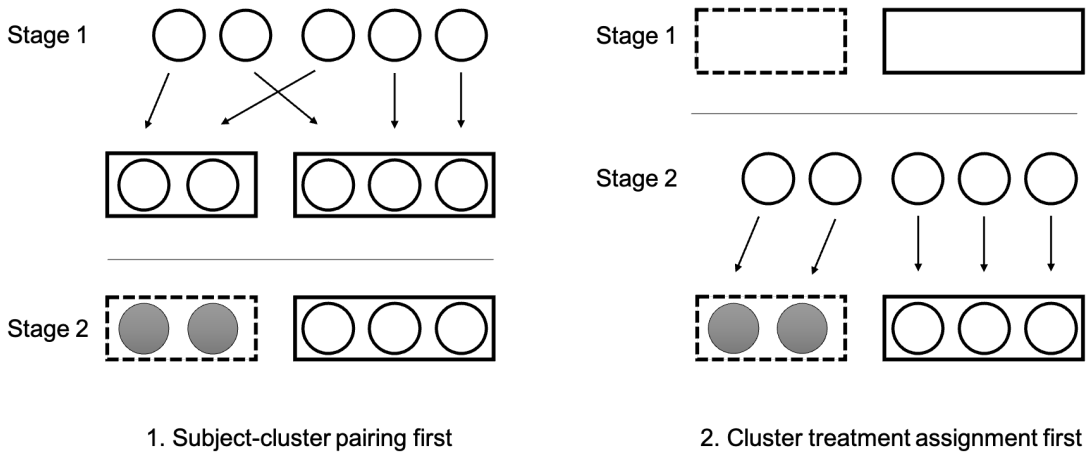

Figure 1: Schematic of two different types of target trials with different unit-cluster selection mechanisms. Dashed line represents cluster-level treatment assignment. The gray color represents treated units.

race, etc), and for each cluster j, we also observe baseline covariates W_j that describe cluster characteristics (school or hospital size). Denote $\boldsymbol{W} = (W_1, \ldots, W_m)$, $\boldsymbol{X} = (X_1, \ldots, X_n)$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Given the natural clustering in a COS, we also observe covariates that are aggregates of the individual-level covariates in the cluster denoted as $h(\boldsymbol{X}_{[j]})$, where $\boldsymbol{X}_{[j]} = \{X_1, \ldots, X_n : J_i = j\}$ is the collection of covariates for subjects belonging to the cluster j, which is a function of $(\boldsymbol{X}, \boldsymbol{J})$. Some commonly-used $h(\cdot)$ functions include the mean and quantile functions.

We assume that subjects' potential outcomes depend on $\boldsymbol{A}$ only through the exposure value of their own cluster, but not other clusters' exposure values, i.e., for any two exposure vectors $\boldsymbol{a}$ and $\boldsymbol{a}'$ such that $\boldsymbol{a}_{j_i} = \boldsymbol{a}'_{j_i}$, we have $Y_i(\boldsymbol{a}, \boldsymbol{j}) = Y_i(\boldsymbol{a}', \boldsymbol{j})$. Thus, the potential outcome $Y_i(\boldsymbol{a}, \boldsymbol{j})$ can be simplified and written as $Y_i(a_{j_i}, \boldsymbol{j})$, and the observed outcome satisfies $Y_i = Y_i(\boldsymbol{A}, \boldsymbol{J}) = Y_i(A_{J_i}, \boldsymbol{J})$. In the identification analysis that follows we focus on a common target estimand the average treatment effect formalized as:

$$E[Y_i(1, \boldsymbol{J}) - Y_i(0, \boldsymbol{J})]. \tag{1}$$

This estimand represents the average difference between the potential outcomes that would be observed if all clusters were treated versus if all clusters were untreated, while the cluster membership is kept as the observed (possibly random) value in the actual observational study, which for the second trial could depend on the observed treatment. In target trial 2, this estimand is similar to that of a natural direct effect in a mediation analysis (VanderWeele, 2016).

3 Identification

Next, we turn to issues of causal identification. Under the target trials articulated in Figure 1 with randomization of both units and clusters, the average treatment effect is identified. Here, we derive the relevant identification conditions for both of the target trials reconstituted as observational studies assuming treatments are not randomly assigned. We focus on the identification conditions through conditioning on baseline covariates. We outline that each target trial identification depends on different adjustment or conditioning sets. We demonstrate, in particular, that the presence of differential selection at the individual level requires conditioning on unit-level covariates.

3.1 Target Trial 1: Fixed Unit-Cluster Pairing

First, we focus on target trial 1 where unit-cluster pairing occurs before treatment assignment. For target trial 1, identification of the treatment effect is possible using a version of the standard conditional ignorability assumption altered to reflect cluster-level treatment assignment (VanderWeele, 2008; Hansen et al., 2014). That is, one assumes that treatment assignment is random within strata of the cluster-level covariates W_j and $h(\boldsymbol{X}_{[j]})$ which are aggregates of the individual-level covariates in the cluster. Formally, the key identification assumption is written as

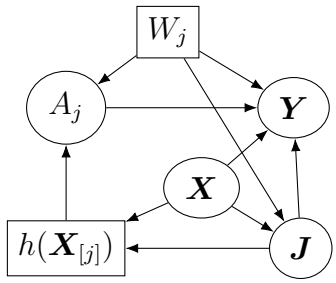

Figure 2: A DAG illustrating identification conditions for cluster j in target trial 1. Conditioning on W_j and $h(\boldsymbol{X}_{[j]})$ renders the treatment-outcome relationship unconfounded.

Assumption 1 (Target Trial 1)**.** *(i)* $\boldsymbol{J} = \boldsymbol{J}(1) = \boldsymbol{J}(0)$*; (ii) For every* i, j*,* a*, and* $\boldsymbol{j}$*,* $A_j \perp \{\boldsymbol{J}, \boldsymbol{X}, Y_i(a, \boldsymbol{j})\} \mid W_j, h(\boldsymbol{X}_{[j]})$.

Assumption 1(i) describes a unit-cluster pairing mechanism that is unaffected by cluster-level treatment assignments. Assumption 1(ii) says that each cluster's treatment assignment probability is a function of the cluster's characteristics and certain aggregate of the individual-level covariates in the cluster. We should note that outside of the COS setting, it is not common to require $\boldsymbol{X}$ conditionally independent of A_j. Figure 2 represents this assumption for each cluster as a directed acyclic graph (DAG). The key feature of this DAG is that conditioning on W_j and $h(\boldsymbol{X}_{[j]})$ blocks all backdoor paths from A_j to $\boldsymbol{Y}$ and renders the treatment-outcome relationship unconfounded (Pearl, 1995).

Next, we provide a formal statement of identification under this assumption. Define $\mu_{\text{wh}}(a, w, h) = E[Y_i \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h]$ and $\tau_{\text{wh}} = E[\mu_{\text{wh}}(1, W_{J_i}, h(\boldsymbol{X}_{[J_i]})] - E[\mu_{\text{wh}}(0, W_{J_i}, h(\boldsymbol{X}_{[J_i]})]$. Similarly, define $\mu_{\text{whx}}(a, w, h, x) = E[Y_i \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x]$ and $\tau_{\text{whx}} = E[\mu_{\text{whx}}(1, W_{J_i}, h(\boldsymbol{X}_{[J_i]}), X_i)] - E[\mu_{\text{whx}}(0, W_{J_i}, h(\boldsymbol{X}_{[J_i]}), X_i)]$.

Proposition 1 shows that there are two ways of identifying the treatment effect in target trial 1.

Proposition 1 (Target Trial 1)**.** *Under Assumption 1,* $\tau_{\text{wh}} = \tau_{\text{whx}} = E[Y_i(1, \boldsymbol{J})] - E[Y_i(0, \boldsymbol{J})]$.

The proof of Proposition 1 and all other proofs are in the supplementary material.

Proposition 1 shows that the treatment effect is identifiable even if we do not condition on unit-level covariates $\boldsymbol{X}$. This is due to the fact that in Assumption 1, $\boldsymbol{X}$ does not directly affect clusters' treatment assignment. However, adjusting for unit-level covariates may improve efficiency.

3.2 Target Trial 2: Cluster Treatment Assignment First

Next, we consider identification for target trial 2. Here, the identification conditions differ from target trial 1. Critically, we must now account for the possibility that when cluster treatment assignment precedes unit-cluster pairing, $\boldsymbol{J}$ is *post-treatment* and can directly affect subjects' outcomes. Notably, under target trial 2, we consider two different mechanisms for how subjects are selected into clusters. As such, we split target trial 2 into two target trials that we denote as 2(a) and 2(b). Specifically, we denote blinded unit-cluster pairing as target trial 2(a), and we denote unblinded unit-cluster pairing as target trial 2(b). Here, differential selection is possible under target trial 2(b), due to the unblinded unit-cluster pairing mechanism.

3.2.1 Target Trial 2(a): Blinded Unit-Cluster Pairing

For target trial 2(a), units are blinded to clusters' treatment status when selecting clusters, which we formalize through the following assumption:

Assumption 2 (Target Trial 2)**.** *(i)* $\boldsymbol{J} = \boldsymbol{J}(1) = \boldsymbol{J}(0)$*; (ii) for every* i, j, a*, and* $\boldsymbol{j}$*,* $A_j \perp \{\boldsymbol{J}, \boldsymbol{X}, Y_i(a, \boldsymbol{j})\} \mid W_j$.

Assumption 2 formalizes the scenario where at the first stage, clusters (e.g., schools, hospitals, physicians) adopt the treatment or not independently, and at the second stage, subjects select clusters with knowledge of clusters' characteristics $\boldsymbol{W}$ but with no knowledge of clusters' treatment assignment $\boldsymbol{A}$. The key difference under Assumption 2 we are no

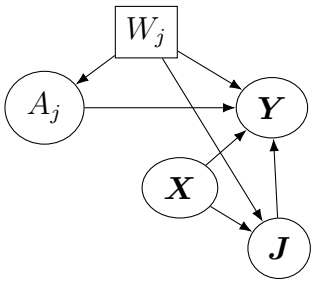

Figure 3: A DAG illustrating identification conditions for cluster j in target trial 2(a). Conditioning on W_j renders the treatment-outcome relationship unconfounded.

longer conditioning on $h(\boldsymbol{X}_{[j]})$. As such, for target trial 2(a) identification is possible by conditioning on reduced set of covariates as compared to target trial 1.

This scheme is illustrated by a DAG in Figure 3, which would arise, for example, when a new reading curriculum is assigned to some schools but not others. If students and their parents are unaware of the new reading curriculum when deciding which school to attend, this would preclude the possibility that the student population could shift to different schools in response to the treatment. On the other hand, it does allow for a scenario, where parents select schools based on school characteristics such as test scores history or lagged student demographics.

Next, we formalize identification under Assumption 2. Define $\mu_{\mathrm{w}}(a, w) = E[Y_i \mid A_{J_i} = a, W_{J_i} = w]$ and $\tau_{\mathrm{w}} = E[\mu_{\mathrm{w}}(1, W_{J_i})] - E[\mu_{\mathrm{w}}(0, W_{J_i})]$. Under Assumption 2, we can identify the average treatment effect in three different ways.

Proposition 2 (Target Trial 2(a))**.** *Under Assumption 2,* $\tau_{\mathrm{w}} = \tau_{\mathrm{wh}} = \tau_{\mathrm{whx}} = E[Y_i(1, \boldsymbol{J})] - E[Y_i(0, \boldsymbol{J})]$.

Proposition 2 shows that conditioning on $\boldsymbol{W}$ suffices to identify the treatment effect. This is because according to Assumption 2, $\boldsymbol{X}$ and $\boldsymbol{J}$ do not directly affect clusters' treatment assignment. However, adjusting for unit-level covariates and their aggregates may improve efficiency.

Note that target trial 2(a) is an instance of target trial 2 where the second stage randomization is independent of the first stage randomization. That is, the unit-cluster pairing does not depend on clusters' treatment assignments. In parallel, Assumption 2 and Proposition 2 can also be applied to target trial 1 when the treatment assignment depends on clusters' characteristics $\boldsymbol{W}$ but does not depend on unit-cluster pairing. Therefore, we see that if the two stages of randomization are independent of each other, the identification conditions and results for target trial 1 and 2 are equivalent.

In target trial 1 and 2(a) where the unit-cluster pairing does not depend on the treatment assignment, we can simplify the notation. That is, the index $\boldsymbol{j}$ in the definition of potential outcomes can be omitted, since $\boldsymbol{J}(1) = \boldsymbol{J}(0) = \boldsymbol{J}$. Specifically, we can define $Y_i(a_{J_i}) := Y_i(a_{J_i}, \boldsymbol{J})$ as the potential outcome for subject i. Under this notation, the possibility of units differentially selecting into clusters in response to treatment is precluded. Under this set of potential outcomes, the observed outcomes can be expressed as $Y_i = Y_i(A_{J_i}, \boldsymbol{J}) = Y_i(A_{J_i}) = A_{J_i}Y_i(1) + (1 - A_{J_i})Y_i(0)$, and the average treatment effect $E[Y_i(1, \boldsymbol{J}) - Y_i(0, \boldsymbol{J})]$ can be expressed in the following familiar form $E[Y_i(1) - Y_i(0)]$, and our identification result is consistent with the results in VanderWeele (2008). This discussion highlights that extant work on identification in COS designs has implicitly assumed that $\boldsymbol{J}$ is not affected by the treatment assignment and has ruled out the presence of differential selection (Hansen et al., 2014; Keele and Zubizarreta, 2017; Pimentel et al., 2018). This assumption holds in target trial 1 where unit-cluster pairing precedes clusters' treatment assignment. In target trial 2, this assumption is also innocuous when units are unaware that clusters have been assigned to the treatment. For example, checklists are often used to reduce medical errors. Hospitals may adopt such interventions with little to any awareness by the patients. In educational settings, many interventions may be alterations of the curriculum that students or parents are unaware of. Qualitative information on whether units are likely to be aware of the intervention will be critical to assessing the

plausibility of target trial 2(a).

3.2.2 Target Trial 2(b): Unblinded Unit-Cluster Pairing

Now we consider the scenario under target trial 2 with unblinded unit-cluster pairing. As we noted above, we refer to this as target trial 2(b). Here, the unit-cluster pairing mediates the effect of treatment on the outcome. Specifically, the unit-cluster pairing $\boldsymbol{J}$ plays the role of a mediator that is affected by the treatment and also can directly affect subjects' outcomes; see review of mediation analysis in VanderWeele (2016). In this case, the primary causal effect of interest is arguably the direct effect of the treatment, because the goal is to learn about the effect of treatment itself separated from any effect due to changing the unit composition of the clusters.

To make progress under this target trial, we introduce an additional assumption about the structure of interference to further simplify the definition of potential outcomes, which may be plausible for a variety of settings. For any two $\boldsymbol{j}$ and $\boldsymbol{j}'$ such that $a_{j_i} = a_{j'_i}, W_{j_i} = W_{j'_i}$, and $h(\boldsymbol{X}_{[j_i]}) = h(\boldsymbol{X}_{[j'_i]})$, we have $Y_i(a_{j_i}, \boldsymbol{j}) = Y_i(a_{j'_i}, \boldsymbol{j}')$. This assumption intuitively asserts that the potential outcomes of subject i remain the same as long as its associated cluster receives the same treatment, has the same cluster characteristics, and consists of subjects with the same individual covariate summaries. This type of assumptions is termed as stratified interference by Hudgens and Halloran (2008). Under the stratified interference assumption, the potential outcome can be further simplified and written as $Y_i(a, w, h) = Y_i(a_{j_i} = a, W_{j_i} = w, h(\boldsymbol{X}_{[j_i]}) = h)$. Under target trial 2(b), the causal estimand is expressed as

$$E[Y_i(1, W_{J_i}, h(\boldsymbol{X}_{[J_i]}))] - E[Y_i(0, W_{J_i}, h(\boldsymbol{X}_{[J_i]}))].$$

This estimand is the average effect of the treatment for each unit while fixing the cluster- and aggregated individual-level characteristics of the associated cluster to the natural values that occur. As such, this estimand quantifies the effect that is purely due to the treatment.

The stratified interference assumption renders this estimand equivalent to the estimand in (1).

Figure 4 contains the DAG for target trial 2(b). Unlike the DAGs for target trials 1 and 2(a), the DAG for target trial 3 is indexed by subject i to account for unit selection to clusters, and there are now arrows from $\boldsymbol{A}$ to $\boldsymbol{J}_{-i}$ and J_i because of unblinded unit-cluster pairing. Based on the DAG in Figure 4, we formalize the key assumption needed for identification using Assumption 3:

Assumption 3 (Target Trial 2(b)). $A_{J_i} \perp Y_i(a, w, h) \mid W_{J_i}, h(\boldsymbol{X}_{[J_i]}), X_i$ *for every* a, w, h, i.

Under this assumption, to disentangle treatment effect and peer effect, we propose comparing treated and control units that have similar observed characteristics and are in similar clusters with similar peers. The formal statement of identification under this assumption is contained in the following proposition:

Proposition 3 (Target Trial 2(b)). *Under Assumption 3,* $\tau_{\text{whx}} = E[Y_i(1, W_{J_i}, h(\boldsymbol{X}_{[J_i]}))] - E[Y_i(0, W_{J_i}, h(\boldsymbol{X}_{[J_i]}))]$.

Proposition 3 shows that under target trial 2(b), the treatment effect is identifiable conditioning on the cluster-level covariates W_{J_i}, individual-level covariates X_i, and aggregates of the individual-level covariates in the cluster $h(\boldsymbol{X}_{[J_i]})$. Here, conditioning on $W_{J_i}, h(\boldsymbol{X}_{[J_i]})$ is analogous to conditioning on mediators when the target parameter is the direct effect of the treatment (VanderWeele, 2016). Critically, we also need to adjust for X_i to account for possible differential unit-level distributions within clusters. Therefore, unlike target trials 1 and 2, identification now depends on conditioning on the full set of baseline covariates.

The key insight from the identification results is that different target trials imply different adjustment sets (e.g., covariate sets on which we must condition). For target trial 1, identification is dependent on a conditioning set that includes $\{W, h\}$. For target trial 2(a), identification requires a conditioning on $\{W\}$, while target trial 2(b) requires conditioning

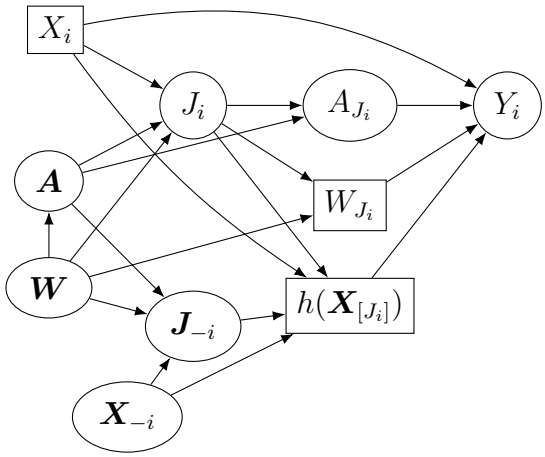

Figure 4: Directed acyclic graph (DAG) illustrating identification conditions for subject i in target trial 2(b). Conditioning on $W_{J_i}, h(\boldsymbol{X}_{[J_i]}), X_i$ renders the treatment-outcome relationship unconfounded.

on $\{W, h, X\}$. For applied analysts, knowing which target trial fits a given application will primarily depend on detailed knowledge of possible mechanisms for differential selection. That is, it is not possible to use the data to test between the different target trials. Critically, however, while it may not be possible to distinguish between the target trials in specific applications, we next use simulations to explore how using different conditioning sets can shed light on the appropriate target trial.

3.3 Estimation and Inference

If the appropriate adjustment set is selected, estimation is relatively straightforward. For example, to estimate τ_{w}, we can impose a parametric model of $\mu_{\mathrm{w}}(a, w)$, denoted by $\mu_{\mathrm{w}}(a, w; \eta_a)$. Let $\hat{\eta}_a$ denote the solution to the score equations corresponding to the likelihood of Y_i conditional on $A_{J_i} = a$ and W_{J_i}, the estimator of τ_{w} is given by

$$\hat{\tau}_{\mathrm{w}} = \frac{1}{n}\sum_{i=1}^{n} \mu_{\mathrm{w}}(1, W_{J_i}; \hat{\eta}_1) - \frac{1}{n}\sum_{i=1}^{n} \mu_{\mathrm{w}}(0, W_{J_i}; \hat{\eta}_0).$$

Estimators for $\hat{\tau}_{\text{wh}}$ and $\hat{\tau}_{\text{whx}}$ can be constructed in a similar fashion. This method of estimation is also called the parametric g-formula, see Hernan and Robins (2020, ch. 13) for a detailed review. Briefly, the estimation process consists of three steps. First, we fit two outcome models, one for treated and one for control. Second, we obtain the fitted values for all n subjects under the outcome models. Third, we standardize by separately averaging over the fitted values under treated and control, and calculate the difference. Variance estimators can be obtained using the block bootstrap which resamples both the clusters and all the units within the resampled clusters (Davison and Hinkley, 1997). We leave the development of more complex estimation methods to future work.

4 Simulations

We conduct a simulation study to further evaluate how the specification of adjustment set can affect the amount of bias when estimating treatment effects in the COS design, especially when the adjustment set does not match the correct target trial. We consider the following data-generating process. First, we generate the baseline covariates at the cluster- and unit- level in the population as:

$$W_j \sim N(0,1),\ j = 1, \ldots, m,$$
$$X_{i1} \sim N(0,1),\ X_{i2} \sim \text{Binom}(0.4),\ i = 1, \ldots, n.$$

For each target trial, we use these baseline covariates to govern how units are assigned to clusters. For target trial 1, we stipulate unit-cluster assignments with the following model:

$$P(J_i = j \mid X_{i1}, X_{i2}, \boldsymbol{W}) = \frac{\exp\{0.2W_j \cdot (1 + X_{i1} + X_{i2})\}}{\left[\sum_{j=1}^{m} \exp\{0.2W_j \cdot (1 + X_{i1} + X_{i2})\}\right]}. \tag{2}$$

Next, for a fixed unit-cluster pair, we calculate aggregate versions of X_{i1} and X_{i2}. Specifically, we aggregate to the 25%, 50%, 75% percentiles for X_{i1}'s and the mean for X_{i2}'s in cluster j, which are respectively denoted as $h_{j1}, h_{j2}, h_{j3}, h_{j4}$, for $j = 1, \ldots, m$. Finally, we assign

clusters to treatment via the following model: $\text{logit}\{P(A_j = 1 \mid W_j, h_{j1}, h_{j2}, h_{j3}, h_{j4})\} = 0.2W_j + 0.2(h_{j1} + h_{j2} + h_{j3}) + 0.2(h_{j4} - 0.4)$. Critically, consistent with target trial 1, treatment assignment depends on the cluster-level covariate and cluster aggregates of unit-level covariates.

For target trial 2, we assign the treatment to clusters according to $\text{logit}\{P(A_j = 1 \mid W_j)\} = 0.2W_j$, such that treatment assignment only depends on the cluster-level covariate, W_j. Then units are assigned to clusters using model (2) without information on clusters' treatment assignments. For target trial 3, we also assign the treatment according to $\text{logit}\{P(A_j = 1 \mid W_j)\} = 0.2W_j$, but we alter the model that assigns units to clusters using the following model:

$$P(J_i = j \mid X_{i1}, X_{i2}, \boldsymbol{W}) = \frac{\exp\{(0.2W_j + 0.2A_j) \cdot (1 + X_{i1} + X_{i2})\}}{\left[\sum_{j=1}^{m} \exp\{(0.2W_j + 0.2A_j) \cdot (1 + X_{i1} + X_{i2})\}\right]}.$$

The key difference is that unit to cluster assignment now depends on treatment assignment. For all three target trials, we generate outcomes using the same model:

$$Y_i = X_{i1} + X_{i2} + 0.4A_{J_i}(X_{i1} + X_{i2}) + 0.5(W_{J_i} + h_{J_i1} + h_{J_i2} + h_{J_i3} + h_{J_i4}) + 0.1e_{J_i} + \epsilon_i$$

where $e_j, \epsilon_i \sim N(0, 1)$ for $j = 1, \ldots, m$, $i = 1, \ldots, n$. The true average treatment effect is $0.4E[X_{i1} + X_{i2}] = 0.16$.

The primary element of the analysis we vary is the adjustment set. That is, for each target trial, we use three different adjustment sets: $\{W\}, \{W, h\}$, and $\{W, h, X\}$, through which we seek to understand how the adjustment set affects the treatment effect estimates. Specifically, we expect that under target trial 1, adjustment for $\{W, h\}$ should be sufficient for consistent estimation of the treatment effect. For target trial 2, adjustment for $\{W\}$ should be sufficient. For target trial 3, adjustment for $\{W, h, X\}$ is necessary. We also vary sample sizes and consider $(m, n) = (50, 4000), (100, 4000)$, and $(50, 8000)$. For each simulation scenario, we estimate treatment effects using separate linear model fits for the

treated and untreated units. The standard errors are obtained using the block bootstrap with 300 bootstrap samples. We used 1,000 simulation repetitions for each scenario.

Results from the simulation study are in Table 1. First, we review the results for target trial 1. For target trial 1, if we only adjust for $\{W\}$ this leads to a biased estimates. However, the bias is relatively modest. When we only adjust for $\{W\}$, the average treatment effect is 0.19 versus the true treatment effect of 0.16 – see row 1 of Table 1. When we adjust for either $\{W, h\}$ or $\{W, h, X\}$, the bias is negligible. For target trial 2, all three adjustment sets lead to unbiased treatment effect estimates, since all three adjustment sets contain $\{W\}$. For target trials 1 and 2, we find that doubling the number of clusters or doubling the total samples sizes substantially reduces the standard errors. However, in either case as long the adjustment set is appropriate, coverage rates perform as expected.

For target trial 3, specifying the correct adjustment set is critical. Under target trial 3, if we only adjusting for $\{W\}$, the estimate is substantially biased. When we only adjust for $\{W\}$, the estimated treatment effect is too large by over a factor of 4. That is, the estimated treatment effects in this scenario are approximately 0.73 relative to the true treatment effect of 0.16. When we adjust for $\{W, h\}$, the bias is still present but much more modest with the estimate effect being approximately 0.19. Finally, if we adjust for the full set of covariates, $\{W, h, X\}$, treatment effects are unbiased. For target trial 3, doubling the number of units does not lead to a reduction in the standard error estimates. That is, additional units do not increase the information in the data given the clustering of units. Here, doubling the number of clusters reduces the size of the standard errors and produces coverage probabilities that are close to the nominal level.

The results from the simulation study agree with the identification results in Section 3. Unbiased estimates depend critically on matching the correct adjustment set to the appropriate target trial. Critically, the key threat is from under-specification. Only under target trial 2 will adjustment for $\{W\}$ alone result in unbiased estimates. In general, we find that

Table 1: Mean, standard deviation (SD), average standard error (SE), and coverage probability (CP) of 95% asymptotic confidence interval for the true average treatment effect $\tau = 0.16$ based on 1,000 simulations.

Target trial	Sizes	Adjustment Set	Mean	SD	SE	CP
1	$m = 50, n = 4000$	W	0.185	0.106	0.104	0.929
		W, h	0.160	0.051	0.080	0.974
		W, h, X	0.160	0.049	0.078	0.966
	$m = 100, n = 4000$	W	0.199	0.102	0.099	0.916
		W, h	0.158	0.043	0.045	0.954
		W, h, X	0.158	0.040	0.042	0.946
	$m = 50, n = 8000$	W	0.170	0.077	0.078	0.946
		W, h	0.161	0.040	0.055	0.979
		W, h, X	0.162	0.038	0.053	0.980
2	$m = 50, n = 4000$	W	0.159	0.098	0.104	0.961
		W, h	0.161	0.050	0.066	0.975
		W, h, X	0.161	0.048	0.063	0.980
	$m = 100, n = 4000$	W	0.160	0.099	0.099	0.943
		W, h	0.160	0.043	0.044	0.952
		W, h, X	0.160	0.041	0.042	0.947
	$m = 50, n = 8000$	W	0.161	0.076	0.078	0.955
		W, h	0.159	0.042	0.053	0.978
		W, h, X	0.159	0.041	0.051	0.972
3	$m = 50, n = 4000$	W	0.732	0.104	0.104	0.001
		W, h	0.191	0.070	0.109	0.957
		W, h, X	0.161	0.067	0.101	0.979
	$m = 100, n = 4000$	W	0.739	0.097	0.099	0.000
		W, h	0.188	0.052	0.054	0.926
		W, h, X	0.160	0.050	0.052	0.962
	$m = 50, n = 8000$	W	0.737	0.075	0.078	0.000
		W, h	0.187	0.072	0.118	0.972
		W, h, X	0.156	0.071	0.112	0.980

adjustment for the full set of covariates, $\{W, h, X\}$, never leads to biased estimates. Given that in many applied applications there may be some uncertainty in terms of which target trial is appropriate, it would be wise to adjust for the full set of covariates.

5 Applications

We present two empirical applications: one from education and one from health services research, which are selected to contrast how key aspects of COS designs can vary depending on the applied context. For each application, we focus on the likelihood of differential selection of units to clusters. In both cases, differential selection is possible, but more likely in one case than in the other. In either case, we are unable to rule out the presence of differential selection.

5.1 Summer School Reading Intervention

In the first application, the empirical question of interest is whether a summer school reading intervention in Wake County, NC improved students' reading scores (Pimentel et al., 2018; Page et al., 2020). Specifically, in the summer 2013, the Wake County Public School System (WCPSS) selected myON, a computer-aided reading program, for use in the summer school program for elementary school students. myON is a web-based software product designed to increase summer school attendees' reading comprehension. Due to technical constraints, only some summer-school sites used myON. WCPSS officials selected the schools that used myON, and principals and schools themselves had no input on program participation. Students at selected schools used the program for up to thirty minutes during the daily summer-school literacy block and could continue using it at home with a device and internet connection. Overall, 3,434 students from 49 different WCPSS elementary schools attended summer school. Of these, 1,371 summer-school students from 20

schools used myON. The primary outcome is student-level reading performance measured via standardized test scores.

To begin, we consider which target trial is appropriate for analysis of the myON intervention. As this application illustrates, selecting the appropriate target trial analogue requires a detailed understanding of the substantive context. For myON, treatment assignment clearly is clustered in that the treatment was applied to all students in selected summer school sites. In general, we judge that the possibility of differential selection at the individual level is quite small. Specifically, students are selected for summer school based on their school-year performance and are residentially zoned into a particular summer school site. Further, summer school selection for myON occurred at the district offices. Students and parents were likely unaware of which summer school sites would be using myON. This is because myON was one relatively small part of the summer school curriculum, and the district did not advertise its use to students or parents. As such, we have no reason to believe that parents would have reacted to the selection of certain sites for myON by shifting student enrollment patterns in a way that would cause differential selection. Nevertheless, we judge that either target trial 2(a) or 2(b) is more appropriate than target trial 1 because the selection of the summer school sites precedes students' summer school selection.

It is important to note however, that we cannot rule out the possible presence of differential selection that would occur under target trial 2(b). Our evidence against differential selection is based on qualitative reasoning and not a statistical test. However, we can use balance tests for additional indirect evidence that target trial 2(a) holds. Table 2 contains balance statistics for student- and school-level covariates. If differential selection did occur, we would expect student level covariates to be correlated with school level treatment assignment. For the myON application, we find that there are clear differences between treated and control schools in terms of school-level covariates such as proficiency in math and the share of teachers who are novices, but differences across student-level covariates

are small.

Critically, our identification results do have observable implications for the role of confounders. Under target trial 2(a) to identify the effect of the myON intervention, we need to condition only on school-level covariates. Student-level covariates may improve efficiency of our estimates but are unnecessary for identification of the effect of interest. That is, if the assumptions of target trial 2(a) hold, we should not observe large differences in the magnitude of the point estimate across specifications that do and do not control for student-level covariates. In the analysis that follows, we consider specifications that include and omit student-level covariates.

For this analysis, we estimate the myON treatment effect using the parametric g-formula based on a linear regression. We included quadratic terms for all continuous covariates. More flexible methods of estimation could be used to further expand the specification. We used 1,000 resamples from the block bootstrap to obtain Efron's percentile confidence intervals. For the specification that omits student-level covariates, the estimated treatment effect is 0.011, which implies that myON increased test scores 0.011 standard deviations. However, the confidence interval includes zero (95% CI: -0.012, 0.034). When we include student-level covariates, the estimated effect is 0.017 but the confidence interval is shorter (95% CI: 0.003, 0.031). In sum, including student-level covariates increases the precision of our estimate slightly but does not substantively change the magnitude, which provides further evidence that differential selection did not occur for the myON intervention.

5.2 Surgical Training

One strand of health services research focuses on whether certain aspects of surgical training have an effect on patient outcomes (Asch et al., 2009; Bansal et al., 2016; Zaheer et al., 2017; Sullivan et al., 2012). Here, we re-analyze one study from this literature. Sellers et al. (2018) studied whether surgeons from university-based residency programs produce

superior patient outcomes compared to surgeons trained in community-based residency programs. In their study, they used a data set that merges the American Medical Association (AMA) Physician Masterfile with all-payer hospital discharge claims from New York, Florida and Pennsylvania from 2012–2013. They collected data on residency type, and surgeons were classified as having attended either a university-based residency (UBR) or a non-university based residency (NUBR) based on the program listed in the AMA Masterfile. Data on surgeon age, sex and year of training completion were also collected. Surgeon experience was defined as year of training completion subtracted from year of operation. They compared surgeon performance between UBR and NUBR surgeons for patients that underwent one of 44 common operations performed by general surgeons in an inpatient setting. Operations were selected to capture a standard set of procedures routinely performed by general surgeons (Sellers et al., 2018). The data also contain patient sociodemographic and clinical characteristics including 31 comorbidities based on Elixhauser indices (Elixhauser et al., 1998). The primary outcome is a binary indicator of any postoperative complications that arise during the hospitalization. Complications were identified using ICD-9 diagnosis codes and collapsed into a binary variable indicating the development of 1 or more complications. For patients treated by UBR surgeons, 12.7% had a post-operative complication. For patients treated by NUBR surgeons, 14% had a post-operative complication. If we estimate the unadjusted treatment effect via a regression model with clustering at the surgeon level, the difference is statistically significant ($p = 0.001$).

The UBR study fits the COS template: all patients treated by a UBR surgeon are exposed to the treatment and vice-versa. However, the structure of the data for this application is quite different compared to the myON application. That is, the number of clusters and units is much larger. There are 498 treated surgeons and 1201 control surgeons. Overall, there are 86,305 patients operated on by UBR surgeons, and 193,307 patients operated on by NUBR surgeons. The number of patients treated by each surgeon

varied from five to 1,074 over the two year period. Thus there are many more clusters, and there is considerable variation in the number of patients per surgeon. Moreover, we have 88 patient-level covariates but only five surgeon-level covariates.

Next, we consider the possibility of differential selection. Unlike in the myON application, we have little qualitative evidence to rule out the possibility of differential selection. That is, it may be the case that if UBR surgeons are viewed as more skilled, they will be assigned patients that have more complex pre-operative conditions or with a generally worse prognosis. As such, differential selection is an open possibility in this application. Again, we can use balance statistics to shed light on this possibility. Specifically, we consider whether UBR patients are observably different from NUBR patients by examining standardized mean differences between patients of NUMBER and UBR surgeons. Surprisingly, we found that none of the patient-level covariates had standardized differences larger than 0.10. This suggests that differential selection may not be in operation. Still, we cannot rule out that UBR patients have higher levels of unobserved frailty.

Next, we estimate the UBR effect using the parametric g-formula via linear regression. In our analysis, we used three different specifications. The first specification only controls for surgeon-level variables. This specification is consistent with target trial 1. In the second specification, in addition to the surgeon-level covariates we included the patient-level variables aggregated to the surgeon level. For the aggregation, we used the average. This specification would identify the average treatment effect under the assumptions of target trial 2(a). In the third specification, we include patient-level variables as well. This specification would identify the average treatment effect under the assumptions of target trial 2(b). We used 1,000 resamples from a block bootstrap to obtain Efron's percentile confidence intervals.

For the surgeon only specification, UBR surgeons had lower complications by 1.5 percentage points (95% CI: -0.022, -0.008). Thus controlling for surgeon level covariates leaves

the estimate nearly unchanged compared to the unadjusted estimate. For the second specification, the difference between UBR and NUBR surgeons is -0.008 (95% CI: -0.014, -0.003). Now the estimated difference in complication rates is quite small. As such, adding the aggregated patient covariates reduces the magnitude of the treatment effect substantially. For the third specification, the difference between UBR and NUBR surgeons is -0.006 (95% CI: -0.010, -0.002). As such, adding patient level covariates does not change the estimated effect. This suggests that in this application, patient assignment to surgeons is not a function of the training, but instead follows some unrelated cluster level mechanism.

6 Discussion

The literature on the COS design has primarily focused on estimation methods and has operated under identification assumptions that are essentially borrowed directly from study designs with unclustered treatment assignment. However, those identification assumptions imply that differential selection does not occur. That is, prior work has assumed that when treatments are assigned to clusters, this assignment has no effect on the population of units within the clusters. While this assumption may be innocuous in some settings, it is implausible in others. For example, in settings where the treatment is a school-wide reform effort, such as Success for All (Borman et al., 2007), parents may react and move to or away from the schools that are exposed to the treatment. In this paper, we consider how differential selection changes the identification assumptions in the context of COSs.

We used the target trial framework to formalize identification conditions for the COS design with and without differential selection. We outlined three possible target trials to describe different scenarios for both the assignment of treatments and units to clusters. In this framework, we show that for each target trial, the set of covariates that render treatment assignment ignorable differ. Under target trial 1, analysts need to condition on cluster-level covariates and cluster-level aggregates. Under target trial 2, analysts need to

condition on cluster-level covariates. Only under target trial 3, do investigators need to condition on cluster covariates, cluster aggregates, and unit-level covariates. For target trial 3, it is critical to condition on unit-level covariates to control for the possible differential mix in cluster populations.

Our work has key implications for applied research. In many cases, researchers may not have definitive information on whether differential selection is present in a COS design. As we demonstrated in our empirical examples, we may be able to reason about the likelihood of differential selection but still be unable to rule it out. Critically, we show that conditioning on the full set of covariates will reduce bias if differential selection is present. However, conditioning on the full set of covariates does no harm if differential selection did not occur. As such, researchers should consider more expansive specifications to reduce possible bias from differential selection. As an alternative, investigators can compare a specification that omits unit-level covariates and a specification that includes unit-level covariates. If the magnitude of the treatment effect estimate differs across these two specifications, it provides some evidence for differential selection. As we highlighted, COS designs are common in many areas of applied research. Our work provides researchers in these areas with guidance on how to consider the possibility of differential selection and how to select specifications that reduce bias.

Table 2: Balance on student- and school-level covariates for myON application.

Student Covariates	Treated Mean Before	Control Mean Before	Std. Difference
Reading pretest score	437.00	437.90	-0.02
Math pretest score	60.25	60.56	-0.02
Male (0/1)	0.36	0.40	-0.09
Special education (0/1)	0.47	0.43	0.09
Hispanic (0/1)	0.53	0.52	0.02
African-American (0/1)	0.22	0.22	0.00
School Covariates			
Composite proficiency	60.74	58.56	0.21
Proficient in reading	58.48	57.27	0.11
Proficient in math	60.68	58.41	0.20
Free/reduced lunch eligible	0.50	0.51	-0.10
English language learners	0.13	0.15	-0.29
Novice teachers	0.19	0.17	0.28
Staff turnover	0.11	0.12	-0.28
Nonwhite teachers	0.14	0.18	-0.26
Title I school	0.90	0.93	-0.11
Schools	20	29	
Summer school students	1,371	2,063	

Note: Standardized difference for a given variable is computed as the mean difference between treatment and comparison schools or students divided by the pooled standard deviation.

References

Adelson, J. L., McCoach, D. B., and Gavin, M. K. (2012). Examining the effects of gifted programming in mathematics and reading using the ecls-k. *Gifted Child Quarterly*, 56(1):25–39.

Asch, D. A., Nicholson, S., Srinivas, S., Herrin, J., and Epstein, A. J. (2009). Evaluating obstetrical residency programs using patient outcomes. *Jama*, 302(12):1277–1283.

Bansal, N., Simmons, K. D., Epstein, A. J., Morris, J. B., and Kelz, R. R. (2016). Using patient outcomes to evaluate general surgery residency program performance. *JAMA surgery*, 151(2):111–119.

Barnes, H., Rearden, J., and McHugh, M. D. (2016). Magnet® hospital recognition linked to lower central line-associated bloodstream infection rates. *Research in nursing & health*, 39(2):96–104.

Bilimoria, K. Y., Chung, J. W., Hedges, L. V., Dahlke, A. R., Love, R., Cohen, M. E., Hoyt, D. B., Yang, A. D., Tarpley, J. L., Mellinger, J. D., et al. (2016). National cluster-randomized trial of duty-hour flexibility in surgical training. *New England Journal of Medicine*, 374(8):713–727.

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal*, 44(3):701–731.

Bryk, A. S., Gomez, L. M., Grunow, A., and LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.

Coleman, J. S. and Hoffer, T. (1987). *Public and Private Schools: The Impact of Communities*. Basic Books, New York, NY.

Coleman, J. S., Hoffer, T., and Kilgore, S. (1982). *High School Achievement: Public, Catholic, and Private Schools Compared*. Basic Books, New York, NY.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27.

Hansen, B. B., Rosenbaum, P. R., and Small, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, 109(505):133–144.

Hedges, L. V. and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1):60–87.

Hernán, M. A. (2018). The c-word: Scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, 108(5):616–619.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hoffer, T., Greeley, A. M., and Coleman, J. S. (1985). Achievement growth in public and catholic schools. *Sociology of Education*, 58(1):74–97.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Institute of Medicine (2009). *Initial national priorities for comparative effectiveness research*. National Academies Press, Washington, D.C.

Keele, L. J. and Zubizarreta, J. (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *Journal of the American Statistical Association*, 112(518):547–560.

Lorch, S. A., Baiocchi, M., Ahlberg, C. E., and Small, D. S. (2012). The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics*, pages peds–2011.

McGuinn, P. J. (2006). *No Child Left Behind and the transformation of federal education policy, 1965-2005*. Univ Pr of Kansas.

McHugh, M. D., Kelly, L. A., Smith, H. L., Wu, E. S., Vanak, J. M., and Aiken, L. H. (2013). Lower mortality in magnet hospitals. *Medical care*, 51(5):382.

Mehta, J., Schwartz, R. B., and Hess, F. M. (2012). *The futures of school reform*. Harvard Education Press.

Navathe, A. S., Silber, J. H., Small, D. S., Rosen, A. K., Romano, P. S., Even-Shoshan, O., Wang, Y., Zhu, J., Halenar, M. J., and Volpp, K. G. (2013a). Teaching hospital financial status and patient outcomes following acgme duty hour reform. *Health services research*, 48(2pt1):476–498.

Navathe, A. S., Silber, J. H., Zhu, J., and Volpp, K. G. (2013b). Does admission to a teaching hospital affect acute myocardial infarction survival? *Academic Medicine*, 88(4):475–482.

Ogburn, E. L. and VanderWeele, T. J. (2014). Causal diagrams for interference. *Statistical science*, 29(4):559–578.

Paeplow, C., Singh, M., and Scrimgeour, M. (2019). Elementary Support Model Implementation and Outcomes: 2014-15 to 2017-18. *Wake County Public School System*.

Page, L. C., Lenard, M., and Keele, L. (2020). The design of clustered observational studies in education. *AERA Open*, 6(3):1–14.

Patel, M. S., Volpp, K. G., Small, D. S., Hill, A. S., Even-Shoshan, O., Rosenbaum, L., Ross, R. N., Bellini, L., Zhu, J., and Silber, J. H. (2014). Association of the 2011 acgme resident duty hour reforms with mortality and readmissions among hospitalized medicare patients. *Jama*, 312(22):2364–2373.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pimentel, S. D., Page, L. C., Lenard, M., and Keele, L. J. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. *Annals of Applied Statistics*, 12(3):1479–1505.

Rao, A. D., Kumar, A., and McHugh, M. (2017). Better nurse autonomy decreases the odds of 30-day mortality and failure to rescue. *Journal of Nursing Scholarship*, 49(1):73–79.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2):173.

Sellers, M. M., Keele, L. J., Sharoky, C. E., Wirtalla, C., Bailey, E. A., and Kelz, R. R. (2018). Association of surgical practice patterns and clinical outcomes with surgeon training in university-or nonuniversity-based residency program. *JAMA surgery*, 153(5):418–425.

Silber, J. H., Romano, P. S., Itani, K. M., Rosen, A. K., Small, D., Lipner, R. S., Bosk, C. L., Wang, Y., Halenar, M. J., Korovaichuk, S., et al. (2014). Assessing the effects of the 2003 resident duty hours reform on internal medicine board scores. *Academic Medicine*, 89(4):644.

Silber, J. H., Rosenbaum, P. R., McHugh, M. D., Ludwig, J. M., Smith, H. L., Niknam, B. A., Even-Shoshan, O., Fleisher, L. A., Kelz, R. R., and Aiken, L. H. (2016). Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA surgery*, 151(6):527–536.

Srinivas, S. K., Fager, C., and Lorch, S. A. (2013). Variations in postdelivery infection and thrombosis by hospital teaching status. *American journal of obstetrics and gynecology*, 209(6):567–e1.

Sullivan, M. C., Sue, G., Bucholz, E., Yeo, H., Bell Jr, R. H., Roman, S. A., and Sosa, J. A. (2012). Effect of program type on the training experiences of 248 university, community, and us military-based general surgery residencies. *Journal of the American College of Surgeons*, 214(1):53–60.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, 27(11):1934–1943.

VanderWeele, T. J. (2016). Mediation analysis: a practitioner's guide. *Annual review of public health*, 37:17–32.

Zaheer, S., Pimentel, S. D., Simmons, K. D., Kuo, L. E., Datta, J., Williams, N., Fraker, D. L., and Kelz, R. R. (2017). Comparing international and united states undergraduate medical education and surgical outcomes using a refined balance matching methodology. *Annals of surgery*, 265(5):916–922.

Supplementary Material

S1 Technical Proofs

S1.1 Proof of Proposition 1

First note

$$
\begin{aligned}
&E[Y_i(a, \boldsymbol{J}) \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x] \\
&= \sum_{j=1}^{m} E[Y_i(a, \boldsymbol{J}) \mid A_j = a, W_j = w, h(\boldsymbol{X}_{[j]}) = h, X_i = x, J_i = j] \\
&\qquad P(J_i = j \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x) \\
&= \sum_{j=1}^{m} E[Y_i(a, \boldsymbol{J}) \mid W_j = w, h(\boldsymbol{X}_{[j]}) = h, X_i = x, J_i = j] P(J_i = j \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x) \\
&= \sum_{j=1}^{m} E[Y_i(a, \boldsymbol{J}) \mid W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x, J_i = j] P(J_i = j \mid W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x) \\
&= E[Y_i(a, \boldsymbol{J}) \mid W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x]
\end{aligned}
$$

where the second equality is because Assumption 1 implies that $A_j \perp \{X_i, Y_i(a, \boldsymbol{J}), J_i\} \mid W_j, h(\boldsymbol{X}_{[j]})$ for every i, j, and thus $A_j \perp Y_i(a, \boldsymbol{J}) \mid W_j, h(\boldsymbol{X}_{[j]}), X_i, J_i$ for every i, j, and the third equality is because $A_{J_i} \perp J_i \mid W_{J_i}, h(\boldsymbol{X}_{[J_i]}), X_i$ from $P(A_{J_i} = 1 \mid J_i = j, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x) = P(A_j = 1 \mid W_j = w, h(\boldsymbol{X}_{[j]}) = h, X_i = x) = \pi_A(w, h, x) = P(A_{J_i} = 1 \mid W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x)$ also from Assumption 1, where $\pi_A(w, h, x) := P(A_j = 1 \mid W_j = w, h(\boldsymbol{X}_{[j]}) = h, X_i = x)$. Then,

$$
\begin{aligned}
&E[Y_i \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x] \\
&= E[Y_i(a, \boldsymbol{J}) \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x] \\
&= E[Y_i(a, \boldsymbol{J}) \mid W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x].
\end{aligned}
$$

Hence, $E[\mu_{\text{whx}}(a, W_{J_i}, h(\boldsymbol{X}_{[J_i]}), X_i)] = E[Y_i(a, \boldsymbol{J})]$. The above results also hold without conditioning on X_i and thus $E[\mu_{\text{wh}}(a, W_{J_i}, h(\boldsymbol{X}_{[J_i]}))] = E[Y_i(a, \boldsymbol{J})]$.

S1.2 Proof of Proposition 2

First note that Assumption 2 implies Assumption 1. Hence the results proved in Proposition 1 still holds under Assumption 2. The results when conditioning on W_{J_i} can be proved in the same way.

S1.3 Proof of Proposition 3

From Assumption 3,

$$
\begin{aligned}
&E[Y_i \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x] \\
&= E[Y_i(a, w, h) \mid A_{J_i} = a, W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x] \\
&= E[Y_i(a, w, h) \mid W_{J_i} = w, h(\boldsymbol{X}_{[J_i]}) = h, X_i = x],
\end{aligned}
$$

Hence, $E[\mu_{\text{whx}}(a, W_{J_i}, h(\boldsymbol{X}_{[J_i]}), X_i)] = E[Y_i(a, W_{J_i}, h(\boldsymbol{X}_{[J_i]}))]$.