Received: 1 October 2021 | Revised: 20 April 2022 | Accepted: 1 May 2022

DOI: 10.1002/sim.9437

RESEARCH ARTICLE

Statistics in Medicine WILEY

Propensity score methods for observational studies with clustered data: A review

Ting-Hsuan Chang[1] | **Elizabeth A. Stuart**[2,3,4]

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[2]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[4]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

Correspondence
Elizabeth A. Stuart, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, Room 804, Baltimore, MD 21205, USA.
Email: estuart@jhu.edu

Propensity score methods are a popular approach to mitigating confounding bias when estimating causal effects in observational studies. When study units are clustered (eg, patients nested within health systems), additional challenges arise such as accounting for unmeasured confounding at multiple levels and dependence between units within the same cluster. While clustered observational data are widely used to draw causal inferences in many fields, including medicine and healthcare, extensions of propensity score methods to clustered settings are still a relatively new area of research. This article presents a framework for estimating causal effects using propensity scores when study units are nested within clusters and are nonrandomly assigned to treatment conditions. We emphasize the need for investigators to examine the nature of the clustering, among other properties, of the observational data at hand in order to guide their choice of causal estimands and the corresponding propensity score approach.

KEYWORDS

causal inference, clustered data, multilevel, observational studies, propensity score

1 | INTRODUCTION

Observational studies are commonly used to evaluate the effect of an intervention or exposure (hereafter collectively referred to as "treatment") in the medical and healthcare literature, because random assignment to treatment conditions may be unethical or impossible. Unlike in randomized control trials (RCTs), where the treatment effect can be estimated unbiasedly by simply comparing the outcome between treatment groups, in observational studies, the contrast in outcome may not accurately reflect the treatment effect due to a potential lack of comparability between groups on pretreatment characteristics associated with the outcome (eg, prognostic factors). This distortion of the treatment effect is often referred to as "confounding bias."[1] While regression adjustment is frequently used to adjust for confounders, over the past two decades, propensity score methods have seen increasing application in observational studies as a less parametric approach to reducing confounding bias compared to regression adjustment.[2,3] Propensity score methods are less parametric in the sense that they are "design-based," in which one attempts to create a sample with treatment groups balanced on baseline characteristics, thus mimicking an RCT at least with respect to the observed characteristics, and confirming that balance is obtained, before analyzing the outcome data.[4,5]

Many of the settings in which treatment effects are being estimated are multilevel (eg, students nested within classrooms or schools, patients nested within hospitals or health systems), and this structure is common in many disciplines, such as education, psychology, epidemiology, and medicine. However, extensions of propensity score methods to multilevel settings are still a relatively new area of research, and even defining the causal estimands in this context is not always straightforward. Several articles have provided a review of using propensity score methods with clustered data,[6-8] but most of these are geared toward an education or social sciences audience. This article aims to bring attention to this topic among clinical and public health researchers by providing an up-to-date guide for estimating causal effects with clustered observational data using existing propensity score methods. For simplicity, we will focus on two-level hierarchical structures and binary treatment assignment at the individual level (ie, individuals are nested within clusters and are nonrandomly assigned to treatment conditions), but the ideas can be easily extended to higher level structures. The different implications of cluster-level treatment assignment (ie, when the treatment is administered at the cluster level) will be briefly discussed.

The article proceeds as follows. The remainder of this section discusses two different contexts that involve clustered observations and their implications for the basic ideas and assumptions underlying the use of propensity scores for causal inference. Section 2 discusses models for the estimation of propensity scores with clustered data. Section 3 reviews three common uses of the propensity score—matching, stratification, and weighting—and their extension to clustered settings, each followed by estimation of the treatment effect. Lastly, Section 4 emphasizes key points, provides a brief discussion on cluster-level treatment assignment, and suggests directions for future research.

1.1 | Two perspectives on clustering

Before conducting propensity score analysis in the context of clustered data, it is important to recognize the nature of the clustering. Here we provide a brief review of two perspectives on clustering detailed in Thoemmes and West[8]: clustering as a central vs incidental feature of the design. In cases where clustering is a central feature, treatment implementation and the assignment mechanism (ie, the selection of individuals into treatment conditions) may vary across clusters.[8] For example, patients' insurance status may strongly predict clinical decisions in some but not other hospitals. Thus, the propensity score analysis in these cases seeks to approximate a multisite randomized trial where random assignment is implemented within clusters.[8] Achieving balance on pretreatment characteristics within each cluster is particularly important if cluster-specific treatment effects are of interest to the investigator.[8] The overall treatment effect can then be estimated by pooling the cluster-specific treatment effects, for example, with weights proportional to cluster sizes. On the other hand, in cases where clustering is an incidental feature, treatment implementation and the assignment mechanism are presumably identical across clusters; the propensity score analysis thus seeks to approximate a single-level randomized experiment where individuals are grouped simply by coincidence (eg, by being in the same room at the same time when treatment is assigned).[8] In these cases the estimand of interest is the treatment effect in the entire study population, and thus achieving covariate balance over the entire sample is the primary goal.[8] With incidental clustering, cluster-level covariates are generally not as meaningful and important as individual-level covariates in affecting causal conclusions. Hence, most of the issues related to clustering discussed in this article will generally be less of a concern for studies with incidental clustering and more prominent for the first case.

Being explicit about the clustering nature of the data can help guide the choice of treatment effect estimands (overall and/or cluster-specific treatment effects), important covariates to adjust for (individual-level and/or cluster-level confounders), and appropriate propensity score strategies. In this article we thus adopt Thoemmes and West's[8] perspectives on clustering and will discuss in further detail the additional nuances and suitable approaches for each type of clustering.

1.2 | Notations, assumptions, and propensity score analysis with clustered data

We now provide an introduction to defining causal effects and using propensity scores for causal inference in observational studies while raising potential issues and complexities specific to clustered settings. For a two-level structure, we use h to index clusters and k to index individuals within a cluster ($k = 1, 2, \dots , n_h$, where n_h is the number of individuals in cluster h), with the assumption that cluster membership is fixed for each individual. Let $N = \sum_h n_h$ be the total number of individuals in the study population.

The treatment effect is defined based on Rubin's causal model,[9,10] including the stable unit treatment value assumption, or SUTVA for short.[11] SUTVA has two components: (i) an individual's outcome is unaffected by the treatment condition received by other individuals; (ii) there is only one version of each treatment condition. Under SUTVA, each individual has two potential outcomes associated with a binary treatment: $Y_{hk}(1)$ (potential outcome under treatment) and $Y_{hk}(0)$ (potential outcome under the control condition). SUTVA is more likely to be violated in multilevel settings compared to single-level settings, especially when clustering is a central feature. An example of violation to component (ii) is when there are variations in a particular surgical procedure (treatment) across hospitals (clusters). Component (ii) can be relaxed to allow these variations in treatment but for each individual they should all have the same effect on the outcome; that is, each individual should have a well defined outcome under "treatment."[12] Component (i) is questionable in settings where individuals in the same cluster are highly connected, for example, students in the same classroom, such that their outcomes may be affected by the treatment received by their peers. In their study on the effect of kindergarten retention, Hong and Raudenbush[13] developed a causal model that relaxes this component of SUTVA by allowing school assignment and the retention rate of a child's school to affect the child's potential outcomes. Specifically, they modeled peer effects on a child's potential outcomes as operating through a scalar function that characterized the school retention rate as low or high. Hence, each child had four instead of two potential outcomes, depending on whether the child was retained and whether the child attended a low or high-retention school.[13] Most of the existing propensity score methods in the multilevel context, however, are still carried out assuming SUTVA. Therefore, the concepts and methods discussed in this article will be based on Rubin's potential outcomes framework and with SUTVA assumed unless otherwise noted.

The individual treatment effect is defined as the difference between the two potential outcomes, $Y_{hk}(1) - Y_{hk}(0)$, which can never be directly observed because only the potential outcome corresponding to the treatment condition received is observed for each individual. Inference thus is often interested in treatment effects averaged across individuals. The average treatment effect (ATE) in the population is defined as the expected value of the individual treatment effects; here we average across individuals such that each individual receives equal weight:

$$\text{ATE} = E\left[Y_{hk}(1) - Y_{hk}(0)\right].$$

Another treatment effect estimand that is commonly of interest is the average treatment effect on the treated (ATT):

$$\text{ATT} = E\left[Y_{hk}(1) - Y_{hk}(0) \,|\, Z_{hk} = 1\right],$$

where Z_{hk} denotes the assigned treatment condition (1 if treated and 0 otherwise). ATT is estimated instead of ATE when interest lies in the treatment effect among those who would receive the treatment in reality or for whom the treatment is intended.[14,15]

As mentioned earlier, cluster-specific ATEs or ATTs may be of interest when clustering is a central feature; these are defined as:

$$\text{ATE}_h = E\left[Y_{hk}(1) - Y_{hk}(0) | H = h\right] \text{ or,}$$

$$\text{ATT}_h = E\left[Y_{hk}(1) - Y_{hk}(0) \,|\, Z_{hk} = 1, H = h\right],$$

where H denotes cluster membership. In some cases it may also be useful to estimate the between-cluster variance of the ATEs or ATTs to examine treatment effect heterogeneity.

Confounding can occur at different levels in multilevel settings. Let $\boldsymbol{X}$ and $\boldsymbol{V}$ denote a set of observed, and supposedly correctly measured, individual-level and cluster-level covariates, respectively. We note that cluster-level covariates may be aggregate versions of the individual-level covariates. For unbiased estimation of the treatment effect using propensity scores, the following identifying assumptions, which jointly define the strong ignorability condition, are assumed[16]:

(i) *Unconfoundedness*:

$$(Y_{hk}(1), Y_{hk}(0)) \perp Z_{hk} \mid (\boldsymbol{X}_{hk}, \boldsymbol{V}_h).$$

(ii) *Positivity* or *overlap*:

$$0 < \Pr[Z = 1|\boldsymbol{X}, \boldsymbol{V}] < 1.$$

Unconfoundedness states that treatment assignment is independent of the potential outcomes conditional on a set of observed covariates. For estimation of the ATT, only $Y_{hk}(0) \perp Z_{hk} \mid (\boldsymbol{X}_{hk}, \boldsymbol{V}_h)$ is needed. Violation of this assumption—specifically, the omission of cluster-level confounders in the analysis (ie, if $\boldsymbol{V}$ does not include all cluster-level confounders)—is referred to as the "unmeasured context" problem by Arpino and Mealli,[17] and is a focus of much of the existing literature on propensity score methods with clustered data.[3,17-19]

A variation on the unconfoundedness assumption in the multilevel causal inference literature is "latent unconfoundedness"[18,20]:

$$(Y_{hk}(1), Y_{hk}(0)) \perp Z_{hk} \mid (\boldsymbol{X}_{hk}, \boldsymbol{V}_h, \alpha_h),$$

where α_h is cluster-specific and captures the effect of all unobserved cluster-level confounders, meaning that this modified version of (i) does not require all cluster-level confounders to be observed and included in $\boldsymbol{V}$. As discussed further below, α_h may be considered as a fixed parameter for each cluster h, or a random variable drawn from a distribution.[18]

Positivity implies that there exists both treated and control individuals for all combinations of the observed covariates. For estimation of the ATT, positivity is only needed for combinations of the observed covariates that are present among the treated individuals. If these assumptions hold, we say that treatment assignment is strongly ignorable given $\boldsymbol{X}$ and $\boldsymbol{V}$. When estimating cluster-specific treatment effects, only the individual-level covariates $\boldsymbol{X}$ are required since $\boldsymbol{V}$ is constant within clusters. That is, to estimate ATT$_h$ (effects within cluster h) we only need to assume that treatment assignment is strongly ignorable given $\boldsymbol{X}$. However, as will be further discussed in subsequent sections, it is often difficult or impossible to estimate cluster-specific treatment effects with small clusters; one extreme example being when all individuals in a cluster receive the same treatment condition, which is more likely to happen when the clusters are small.

The propensity score will be a key tool for estimating treatment effects in the settings we consider. For individual k in cluster h, their propensity score is their probability of receiving treatment conditional on the values of their observed covariates:

$$e_{hk} = \Pr\left[Z_{hk} = 1|\boldsymbol{X}_{hk}, \boldsymbol{V}_h\right].$$

Rosenbaum and Rubin[16] showed that the propensity score is a balancing score, meaning that conditional on the propensity score, the distribution of the covariates entered in the propensity score model would be balanced across treatment groups, and that the treatment assignment is strongly ignorable given e_{hk} if it is strongly ignorable given $\boldsymbol{X}$ and $\boldsymbol{V}$. Given these properties, the propensity score alone can be used to equate the treatment groups on observed covariates ($\boldsymbol{X}$ and $\boldsymbol{V}$). As will be seen in Section 2, to alleviate the unmeasured context problem (ie, if $\boldsymbol{V}$ fails to include all cluster-level confounders), the propensity score model can also include a cluster-specific term (α_h), based on the latent unconfoundedness assumption described earlier:

$$e_{hk} = \Pr\left[Z_{hk} = 1|\boldsymbol{X}_{hk}, \alpha_h\right] \text{ if } \alpha_h \text{ is considered as a fixed parameter, or}$$

$$e_{hk} = \Pr\left[Z_{hk} = 1|\boldsymbol{X}_{hk}, \boldsymbol{V}_h, \alpha_h\right] \text{ if } \alpha_h \text{ is considered as a random variable.}$$

Propensity score analysis for causal inference generally involves the following key steps: (1) estimate the propensity score; (2) induce covariate balance between treatment groups using an appropriate strategy, including matching units on propensity scores, grouping units into strata with similar propensity scores, or applying propensity score weights; (3) check covariate balance with metrics such as the standardized mean difference. This may be done both within and across clusters when clustering is a central feature; the thinking behind this is articulated further below. Researchers may repeat steps (1)-(3) to re-estimate the propensity scores using a different model until adequate balance is achieved; (4) estimate the ATE(s) or the ATT(s); (5) perform sensitivity analysis to assess the impact of unobserved confounders on treatment effect estimates. We will discuss step (1) in the next section and steps (2)-(4) in Section 3. Many sensitivity analysis methods (step 5) have been developed in the single-level context,[21-23] but sensitivity analysis for clustered data is an area in need of more research. For a review of propensity score methods in a single-level context, see, for example, Austin[24] and Stuart.[25] Those interested in carrying out the above in a multilevel context may consider the R package *multilevelPSA*, which is specifically developed to implement propensity score analysis for multilevel data, including estimating a separate propensity score model for each cluster, assessing covariate balance and summarizing or visualizing results at different levels.[26]

2 | PROPENSITY SCORE ESTIMATION WITH CLUSTERED DATA

For binary treatments, the following logistic regression models are often used to estimate propensity scores in a multilevel context, depending on assumptions of the assignment mechanism and specifically, whether unconfoundedness holds conditional on the observed covariates (which may include covariates at any level) only or additionally on cluster membership (α_h in the latent unconfoundedness assumption).

2.1 | Single-level models

Single-level models include the observed individual-level covariates $\boldsymbol{X}_{hk}$ and cluster-level covariates $\boldsymbol{V}_h$ (the index k in the subscript is omitted since $\boldsymbol{V}$ is constant within clusters) as predictors:

$$\text{logit}\,(e_{hk}) = \alpha + \boldsymbol{X}_{hk}\boldsymbol{\beta} + \boldsymbol{V}_h\boldsymbol{\gamma}.$$

α is an intercept, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of coefficients for the individual- and cluster-level covariates, respectively. Interaction (either same-level or cross-level) and polynomial terms may be included for more flexible modeling of the propensity score. Single-level models assume that no confounding remains after conditioning on the observed covariates ($\boldsymbol{X}_{hk}$ and $\boldsymbol{V}_h$), which is unlikely in reality.[3] These models can be a reasonable choice if the assignment mechanism is independent of cluster membership, which is more plausible when clustering is incidental in the observational study.[8] The propensity scores estimated from single-level models are used to achieve balance on the observed covariates across clusters but not necessarily within clusters.[27]

2.2 | Fixed effects models

Fixed effects models assume unconfoundedness conditional on the observed individual-level covariates *and* on cluster membership, thus the inclusion of cluster indicator variables or equivalently, cluster-specific intercepts:

$$\text{logit}\,(e_{hk}) = \alpha_h + \boldsymbol{X}_{hk}\boldsymbol{\beta},$$

where α_h is a fixed intercept term for cluster h. Fixed effects models do not include cluster-level covariates $\boldsymbol{V}_h$ because the cluster-specific intercept captures the effects of both observed and unobserved cluster-level covariates, thus protecting against the unmeasured context problem.[3,17,28] However, the number of parameters to be estimated increases considerably when there are many clusters, especially when interactions between cluster membership and the individual-level covariates are considered, which is equivalent to allowing the slopes of the individual-level covariates to vary across clusters[8]:

$$\text{logit}\,(e_{hk}) = \alpha_h + \boldsymbol{X}_{hk}\boldsymbol{\beta}_h.$$

In these cases the clusters need to be large enough and have a reasonable number of individuals in each treatment condition in order to support estimation of the model.[3,8]

2.3 | Random effects models

An alternative to fixed effects models that reduces the number of parameters are random effects models:

$$\text{logit}\,(e_{hk}) = \alpha_h + \boldsymbol{X}_{hk}\boldsymbol{\beta} + \boldsymbol{V}_h\boldsymbol{\gamma}, \alpha_h \sim N\left(0, \sigma^2\right).$$

The main difference between fixed and random effects models is that the intercept term α_h is considered fixed for each cluster in fixed effects models (by including cluster indicator variables), whereas α_h usually follows a normal distribution in random effects models. In addition to a random intercept, random slopes can also be specified:

$$\text{logit}\,(e_{hk}) = \alpha_h + \boldsymbol{X}_{hk}\boldsymbol{\beta}_h + \boldsymbol{V}_h\boldsymbol{\gamma}_h,$$

$$\alpha_h \sim N\left(0, \sigma^2\right), \boldsymbol{\beta}_h \sim N\left(0, \Sigma_\beta\right), \boldsymbol{\gamma}_h \sim N\left(0, \Sigma_\gamma\right).$$

Kim and Seltzer[29] found that random intercept and slopes (RIS) models yielded improved covariate balance within clusters compared to random intercept only (RI) models when the effects on treatment assignment of certain individual-level covariates vary across clusters. The choice of whether to include random slopes, though, should hinge upon feasibility as well as the (presumed) degree of robustness of the model to omission of random slopes, which may vary by study.[7] One caveat of random effects models is that the observed covariates and the unobserved cluster-level covariates are assumed to be independent,[19] and because the random effects shrink toward zero, there may be more residual imbalance as compared to the fixed effects approach.[3]

Both fixed effects and random effects models acknowledge the potential existence of unobserved cluster-level confounders and follow the latent unconfoundedness assumption with the inclusion of α_h.[18] Several studies[3,8,17] have compared the two models with single-level models for propensity score estimation. Although these studies have very different simulation setups and use the propensity scores in different ways (matching, stratification, or weighting), they all showed that fixed and random effects models outperformed single-level models (which include *observed* individual- and cluster-level covariates) in reducing bias, due to their ability to capture remaining cluster heterogeneity.

Fixed effects models with cluster-specific slopes and RIS models mimic the assignment mechanism in a multisite randomized trial, and the estimated propensity scores are not comparable across clusters because of the different regression slopes.[8] As a result, these models aim to achieve balance within clusters rather than across clusters and thus are preferable when ATE$_h$ or ATT$_h$ may be of interest.

Fixed effects models with cluster-specific slopes are generally the ideal approach when clustering is a central feature, both because they acknowledge the potential heterogeneity in the assignment mechanism across clusters, and because they have the advantage of automatically controlling for all cluster-level covariates so that only the individual-level covariates are needed. However, in reality cluster sizes are often too small to support propensity score modeling within clusters.[27] A few strategies have been proposed to overcome the issue with small clusters by combining similar clusters into larger "groups" or "classes."[30-33] For example, Kim and Steiner[30] identified homogeneous classes of clusters with respect to the treatment assignment model using latent class analysis; Lee et al[33] formed cluster groups with similar treatment prevalence. A distinct propensity score model is then constructed for each group, or class, of clusters. These strategies were shown to be effective in reducing bias when cluster sizes are too small to estimate fixed cluster effects.

3 | THE USE OF PROPENSITY SCORES AND TREATMENT EFFECT ESTIMATION WITH CLUSTERED DATA

The previous section discussed how propensity scores can and should be estimated in clustered settings; we now turn to how those propensity scores can be used in this context. In particular, this section reviews the extension to clustered data of three common methods that use propensity scores—matching, stratification, and weighting—and the estimation of treatment effects when using these methods. We remind the reader that the use of propensity scores can be viewed as the preprocessing stage to induce comparability between treatment groups on observed covariates in the data, which sets the basis for the outcome analysis stage where treatment effects are estimated with the preprocessed data.[25,34]

3.1 | Matching

Propensity score matching is the process of creating matched sets of treated and control individuals with similar values of the propensity score. It is generally used to estimate the ATT[35]; hence we focus on the estimation of ATT in this subsection. Two common algorithms for forming matched sets are greedy matching and optimal matching.[36] For more details on these matching algorithms, we refer the reader to Austin,[24] Stuart,[25] and Gu and Rosenbaum.[37] Here we focus on greedy *k*:1 nearest neighbor matching, which is one of the most common matching methods.[25] In *k*:1 nearest neighbor matching, each treated individual is matched with one or multiple controls whose propensity score values are closest to theirs; unmatched controls are then discarded once all treated individuals have found their matches. To ensure quality matches, researchers may impose a prespecified caliper distance (eg, 0.2 of the SD of the logit of the propensity score) on the absolute difference in the propensity scores of matched individuals.[38] Treated individuals with no controls falling within the caliper are discarded, which will then estimate a restricted ATT.

The considerations for matching in general apply to both single-level and multilevel settings. With clustered data, researchers need to also choose between matching within or across clusters, or a combination of both ("hybrid" matching strategies).

3.1.1 | Within-cluster matching

Within-cluster matching requires that matched individuals belong to the same cluster, which controls for all cluster-level covariates automatically. Treatment effect estimates will then be free from the bias caused by unobserved cluster-level confounders, which is desirable when strong cluster-level confounding is likely. Hence, cluster-specific propensity score estimation (ie, using a fixed effects model with cluster-specific slopes) followed by within-cluster matching is the ideal approach when clustering is a central feature and the aim is to approximate a multisite randomized trial. This achieves balance within clusters and could yield cluster-specific effect estimates if the clusters are sufficiently large.[8] However, this approach may not always be feasible: with small cluster sizes, there may be insufficient propensity score overlap between treatment groups within clusters and many unmatched individuals may be discarded, leading to a more limited estimand, or the quality of the matches may be relatively low in terms of their similarity on the individual-level covariates.

3.1.2 | Across-cluster matching

The other extreme is across-cluster matching, which selects matches using the entire study population, irrespective of cluster membership. Therefore, across-cluster matching requires a joint propensity score estimation model so that the propensity scores across clusters are comparable.[8] This achieves balance on the observed covariates in the entire matched dataset but may not produce balance on the individual-level covariates within each cluster.[8] Hence, across-cluster matching is most suitable under incidental clustering and when the estimand is the overall ATT.[8] Across-cluster matching may retain a larger sample size compared to within-cluster matching. However, it relies on stronger identifying assumptions than within-cluster matching, as it requires the correct measurement and modeling of cluster-level covariates.[30,39] The choice between within- and across-cluster matching thus exemplifies a bias-variance tradeoff and depends on how important it is to achieve balance on cluster-level covariates.

3.1.3 | Hybrid matching strategies

In many ways the choice of matching approaches comes down to how important it is to get balance on individual-level vs cluster-level covariates, often based on how important each set are in terms of predicting the outcome. Thus, alternative matching strategies have been developed to combine the benefits of both within- and across-cluster matching. The general idea is to prioritize within-cluster matches, but not at the expense of poor balance on the individual-level covariates. Rickles and Seltzer[40] proposed a two-stage propensity score modeling (2SM) approach, which is an extension of Stuart and Rubin's[41] matching method with multiple control groups to the multilevel setting. The 2SM approach first attempts matching within clusters. For treated individuals without an acceptable match (ie, too dissimilar in terms of the individual-level covariates), matches ("nonlocal" controls) are then searched across clusters but within a preidentified group of clusters that share similar characteristics as the cluster to which the treated individual belongs.[40] Because across-cluster matching may lead to poor covariate balance if it is based on propensity score models that vary across clusters, for each treated individual looking for nonlocal matches Rickles and Seltzer[40] recalculated the propensity scores of nonlocal controls using the same model as the treated individual to ensure comparability of the propensity scores. In addition, Rickles and Seltzer[40] addressed any potential imbalance in cluster-level covariates by adjusting for between-cluster differences in each nonlocal control's observed outcome before estimating treatment effects. Another popular hybrid matching strategy is "preferential" within-cluster matching, proposed by Arpino and Cannas.[42] Like the 2SM approach, preferential within-cluster matching prioritizes matching within clusters and searchers for matches across clusters only if no acceptable within-cluster match can be found, but does not restrict across-cluster matching to a group of similar clusters.[42] The R package *CMatching* provides functions that implement within-cluster matching and preferential within-cluster matching.[43]

3.1.4 | Treatment effect estimation after matching

After a matched dataset with adequate balance is formed, researchers can proceed to the outcome analysis stage by comparing outcomes among the matched sample. In practice, a "doubly robust" approach, in which the ATT is estimated using a (multilevel) outcome model that includes both the treatment condition and observed covariates as predictors, is recommended.[44] Note that weights may be applied if necessary, for example, when matching with replacement. In essence "doubly robust" means that when the propensity score and outcome models are used in combination, only one of the two models is needed to be correctly specified to provide a consistent estimate of the treatment effect.[45] Treatment effect heterogeneity across clusters can be evaluated by allowing the slope associated with the treatment variable to vary by cluster.

A less ideal (because of the potential for residual confounding after matching) but simpler alternative is to estimate the ATT nonparametrically as follows:

$$\hat{\tau}^{\mathrm{ATT}} = \frac{1}{N_1} \sum_{h,k} \left(Y_{hk} - \hat{Y}_{hk}(0) \right) Z_{hk}.$$

$\hat{\tau}^{\mathrm{ATT}}$ estimates the average of the individual treatment effects among the treated, since $Z_{hk} = 0$ for controls and here N_1 is the number of treated individuals in the matched dataset. $\hat{Y}_{hk}(0)$ is the estimated potential outcome under the control condition, which can be estimated by averaging the outcomes of all controls matched to individual k in cluster h. If the ATT is estimated within each cluster, one can obtain an overall ATT estimate by pooling the cluster-specific ATTs with weights proportional to the number of treated individuals in the cluster. The SE estimate of the nonparametric ATT estimator can be obtained via cluster bootstrapping (resampling the clusters rather than individuals). However, Abadie and Imbens[46] found that bootstrap SEs were not valid for nearest-neighbor matching with replacement and a fixed number of matches. An alternative is to simply regress the outcome on the treatment variable in the matched dataset, and obtain SEs that account for the within-cluster dependence in the outcome due to shared cluster-level characteristics.[42]

3.2 | Stratification

Propensity score stratification involves dividing the sample into mutually exclusive strata based on the quantiles of the estimated propensity scores. Therefore, the distribution of the observed covariates should be roughly similar between the treatment groups within each stratum. Recent studies on the choice of the number of strata can be found in Neuhäuser et al[47] and Orihara and Hamada.[48] Another method that is closely related to propensity score stratification is optimal full matching (OFM).[49] Unlike the matching methods discussed in Section 3.1, which may discard unmatched individuals, full matching forms matched sets that contain at least one treated and at least one control individual using the entire sample.[49,50] OFM is optimal in the sense that the mean of the within-stratum distances in the propensity score between treatment conditions is minimized. OFM can thus be viewed as a form of propensity score stratification where the number of stratum and the grouping of individuals are optimized to minimize the propensity score differences within strata.

Like matching, stratification (or OFM) can be done either within or across clusters in multilevel settings, with the same considerations as with within- vs across-cluster matching. However, stratification across clusters is more commonly done in applied studies[27,51] as stratification within clusters is often hindered by small cluster sizes.

3.2.1 | Treatment effect estimation after stratification

After a stratified sample is formed, treatment effects (or cluster-specific treatment effects if stratification is done within clusters) can be estimated by taking the weighted average of the stratum-specific treatment effects. Similar to matching, stratum-specific treatment effects can be estimated nonparametrically (eg, by simply comparing the difference in the mean outcome between treated and control individuals irrespective of cluster membership) or more ideally, using a doubly robust approach (eg, by regressing the outcome on the treatment variable and all observed covariates, and incorporating fixed or random cluster effects). Stratum-specific treatment effects are then weighted by the size of each stratum for estimating the ATE, or by the number of treated individuals in each stratum for estimating the ATT.

Another way to estimate the overall treatment effect, as demonstrated in Leite et al,[7] is by applying the "marginal mean weights" provided by Hong,[52] which are weights based on stratum membership and treatment assignment:

$$w^{\text{ATE}} = \frac{N_s}{N_{z,s}} \Pr[Z = z],$$
$$w^{\text{ATT}} = z + (1 - z)\frac{N_{1,s}}{N_{0,s}} \times \frac{1 - \Pr[Z = 1]}{\Pr[Z = 1]},$$

where N_s is the total number of individuals in stratum s and $N_{z,s}$ is the number of individuals with treatment level z within stratum s. Investigators can then construct a weighted outcome model (preferably a model the accounts for the cluster structure) to obtain the ATE or ATT.[7]

3.3 | Weighting

Propensity score weighting is a technique that directly uses the estimated propensity scores to adjust for confounding bias. In multilevel settings, weights are constructed in the same manner as in single-level settings. For estimating the ATE, propensity score weights, or "inverse probability of treatment weights" (IPTW), are: $w_{hk} = \frac{1}{\hat{e}_{hk}}$ for treated individuals and $w_{hk} = \frac{1}{(1-\hat{e}_{hk})}$ for control individuals, where $\hat{e}_{hk}$ is the estimated propensity score. For estimating the ATT, treated individuals receive a weight of 1, while the weight for control individuals is $\frac{\hat{e}_{hk}}{(1-\hat{e}_{hk})}$. Similar to single-level settings, the weights are applied to create a pseudo-sample balanced on the observed covariates. In the case of estimating the ATE, both treatment groups are weighted to resemble the combined group in terms of the observed covariates, whereas in the case of estimating the ATT, control individuals are weighted to resemble the treated group.

Extreme weights can often lead to inflated variance estimates. To prevent extreme weights, one may consider the use of stabilized weights[53]: $w_{hk} = \frac{P(Z=1)}{\hat{e}_{hk}}$ for treated individuals and $w_{hk} = \frac{P(Z=0)}{1-\hat{e}_{hk}}$ for control individuals when estimating the ATE. In multilevel settings, one may replace the across-cluster average probability of treatment (or no treatment) in the numerator of these "marginal stabilized weights" by the cluster-specific probability $P(Z = z|H = h)$ to create "cluster-mean stabilized weights."[54] Schuler et al[54] found that cluster-mean stabilized weights outperformed marginal stabilized weights in the context of a continuous treatment.

3.3.1 | Treatment effect estimation after weighting

The ATE or the ATT can simply be estimated by the difference of the weighted means of the outcomes between the two treatment groups ("marginal" estimator),[3] ignoring the cluster structure:

$$\hat{\tau}_{ma} = \frac{\sum_{h,k} Z_{hk} Y_{hk} w_{hk}}{\sum_{h,k} Z_{hk} w_{hk}} - \frac{\sum_{h,k} (1 - Z_{hk})\, Y_{hk} w_{hk}}{\sum_{h,k} (1 - Z_{hk})\, w_{hk}}.$$

On the other hand, a cluster-weighted estimator first estimates the cluster-specific treatment effects ($\hat{\tau}_h$), then takes a weighted average of these to estimate the overall treatment effect ($\hat{\tau}_{cl}$)[3]:

$$\hat{\tau}_{cl} = \frac{\sum_h w_h \hat{\tau}_h}{\sum_h w_h},$$

where

$$\hat{\tau}_h = \frac{\sum_{k=1}^{n_h} Z_{hk} Y_{hk} w_{hk}}{\sum_{k=1}^{n_h} Z_{hk} w_{hk}} - \frac{\sum_{k=1}^{n_h} (1 - Z_{hk}) Y_{hk} w_{hk}}{\sum_{k=1}^{n_h} (1 - Z_{hk}) w_{hk}} \text{ and } w_h = \sum_{k=1}^{n_h} w_{hk}.$$

Li et al[3] recommended using the bootstrap (resampling the clusters) to estimate the SEs of these nonparametric estimators.

More recently, Yang[18] and He[19] proposed alternative propensity score estimation methods to improve the marginal estimator ($\hat{\tau}_{ma}$) such that it is robust to unobserved cluster-level confounding: Yang[18] proposed a calibration method that involves defining the weights in a way that satisfies two constraints: (i) for each observed covariate, the sum in the full sample equals the weighted sum in each treatment group; (ii) in each cluster, the sum of the propensity score weights in each treatment group equals the size of the cluster. He[19] proposed conditional propensity score estimation, which utilizes a sufficient statistic for the unobserved cluster-level covariates, for example, a function of the treatment allocations in the cluster. Both methods were shown to reduce bias in $\hat{\tau}_{ma}$ due to omitted cluster-level confounders under certain conditions, though Yang[18] cautioned that the robustness of the calibrated estimator may not hold if some individual-level confounders were omitted.

In terms of parametric estimation, doubly robust estimators for the ATE and the ATT have been formulated as follows[3,55,56]:

$$\hat{\tau}_{dr}^{\text{ATE}} = \sum_{h,k} \frac{1}{N} \left\{ \left[\frac{Z_{hk} Y_{hk}}{\hat{e}_{hk}} - \frac{\left(Z_{hk} - \hat{e}_{hk}\right) \hat{Y}_{hk}^1}{\hat{e}_{hk}} \right] - \left[\frac{(1 - Z_{hk}) Y_{hk}}{1 - \hat{e}_{hk}} - \frac{\left(Z_{hk} - \hat{e}_{ij}\right) \hat{Y}_{hk}^0}{1 - \hat{e}_{hk}} \right] \right\};$$

$$\hat{\tau}_{dr}^{\text{ATT}} = \sum_{h,k} \frac{1}{N_1} \left[Z_{hk} Y_{hk} - \frac{(1 - Z_{hk}) Y_{hk} \hat{e}_{hk} + \left(Z_{hk} - \hat{e}_{hk}\right) \hat{Y}_{hk}^0}{1 - \hat{e}_{hk}} \right].$$

$\hat{Y}_{hk}^1$ and $\hat{Y}_{hk}^0$ are the fitted outcomes based on the outcome model in the treated and control groups, respectively.[3] As with the nonparametric estimators, the SEs of the doubly robust estimators can be estimated via cluster bootstrapping.[3] Alternatively, researchers can fit an outcome model on the propensity score weighted sample to obtain the treatment effect and cluster robust SEs.

Due to the direct use of propensity scores, weighting-based estimators are particularly sensitive to incorrect specifications of the propensity score model. Potential solutions include using a doubly robust approach (combining the propensity score and outcome models), or the calibrated propensity score weighting estimator proposed by Yang,[18] which was shown to be robust to propensity score model misspecification.

4 | DISCUSSION

For clustered observational studies in which individuals are nonrandomly assigned to treatment, propensity score methods should be modified to incorporate the cluster structure for two main reasons: (1) to account for within-cluster correlation in both the propensity score and the outcome due to shared cluster-level characteristics; (2) to account for potential unobserved cluster-level confounding. While these may be less of an issue when clustering is incidental (and thus individuals in the same cluster may not share much in common), taking account of the cluster structure when clustering is a central feature not only reduces bias, but is also necessary for valid SE estimation.

Regardless of the propensity score method used, it is unanimously agreed in the literature that taking account of the cluster structure in either the propensity score estimation (eg, using a multilevel model to estimate propensity scores) or the outcome analysis stage (eg, using the cluster-weighted IPTW estimator or a multilevel outcome model) can significantly reduce bias. Incorporating cluster information in both stages yields the least bias.[3,57] In addition, Li et al[3] found that the choice of the outcome model may be more important for reducing bias than the propensity score model. Specifically, their simulation results suggested that failing to include the cluster structure in the outcome model may lead to larger bias compared to failing to include it in the propensity score model when unmeasured cluster-level confounding is present.[3] More recently, Suk et al[58] extended the causal forest algorithm in Athey and Wager[59] to multilevel observational settings, showing promising performance when propensity scores estimated from multilevel models were used.

This review focuses exclusively on clustered observational studies where treatment is assigned to individuals. However, some research questions involve treatment assignment at the cluster-level, such as an intervention at a whole school or community level. In this case, it is thus more important to control for cluster-level covariates than to control for individual-level covariates[60]; note too that the cluster-level covariates may be aggregate versions of the individual-level covariates. The outcome of interest can be measured either at the individual or the cluster level. For studies that measure the outcome at the cluster level, the ideas discussed under the usual unstructured data context can be directly applied, as the clusters themselves will be the units for analysis.[61] On the other hand, propensity score methods require some modifications for studies where interest is in the cluster-level treatment effect on individual-level outcomes, which is common for many health care interventions, for example, interventions on the health systems. These studies may approximate a clustered randomized control trial (clustered RCT), where randomization of treatment occurs at the cluster level.[60] In terms of matching strategies for studies that emulate a clustered RCT, some have suggested to match the clusters only or to match clusters first and then match individuals within matched clusters.[30,62] Recent work by Zubizarreta and Keele,[63] however, proposed an optimal matching strategy in which individuals are first matched across all possible combinations of treated and control clusters, and then clusters are matched considering the individual-level matches. Other methodological work that examines the effect of cluster-level treatment on individual outcome includes Leyrat et al[64] and Balzer et al.[65]

In the remaining of this section, we provide suggestions on areas for future research in the realm of multilevel causal inference.

4.1 | Violation of SUTVA

As mentioned in Section 1, the SUTVA assumption is likely untenable in multilevel settings where cluster members are interconnected and influential to one another. However, most of the existing literature reviewed in this article have proceeded under SUTVA, both within and across clusters, without an explicit discussion on the violations to the assumption. A few studies have adopted a modified version of SUTVA that allows "partial interference"—that is, that there is interference between individuals in the same cluster, but no interference between individuals from different clusters.[66] Examples include a kindergarten retention study in Hong and Raudenbush[13] and a study on the effect of students' eating habits and supportive school environments on obesity in Eckardt.[67] Both studies capture the interference within schools through a binary variable indicating the school's treatment pattern which, together with the student's own treatment condition, defines the potential outcomes. This approximates a two-stage randomization where schools are first randomly assigned to treatment allocation plans, and students are then assigned to treatment conditions within schools. Other examples can be found more broadly in the social networks and infectious disease literatures, in which the spillover effect (the effect of treatment received by others) is commonly of interest. For more propensity score-based causal inference methods in these contexts, see, for example, Tchetgen Tchetgen and VanderWeele[68] and Forastiere et al.[69] Causal inference with interference is a highly active area of research, and propensity score methods allowing different interference patterns between individuals in multilevel settings is an important direction for future work.

4.2 | Nonparametric propensity score estimation

In Section 2, we presented logistic regression models that are frequently used for the estimation of propensity scores with clustered data, but researchers can also use nonparametric techniques to estimate propensity scores; such an approach can be especially beneficial when the number of covariates is large. Nonparametric techniques, such as generalized boosted modeling, random forest, and neural networks, were shown to be promising alternatives to logistic regression for estimating propensity scores in single-level settings.[70,71] A recent study by Han and Kwak[72] showed that propensity scores estimated by a conditional inference tree produced better covariate balance than did propensity scores estimated by multilevel logistic regression models with two-level clustered data, but the comparison was evaluated in an example with fairly large clusters (24 countries with 144 to 1532 schools per country). The performance of nonparametric propensity score estimation has yet to be investigated with different machine learning techniques and in different multilevel settings.

4.3 | Insufficient overlap

Insufficient overlap (ie, violation of the positivity condition), either within or across clusters, between the treated and control groups is another issue that could be further addressed in future works. For example, the preferential within-cluster matching approach introduced by Arpino and Cannas[42] was shown to be effective when the propensity score distribution showed sufficient overlap in the full sample, but its performance in situations where overlap is poor remains unclear.[42] Propensity score weighting generally does not discard individuals, but insufficient covariate overlap between treatment groups can lead to extreme weights, which may in turn result in bias and large variability of the treatment effect.[73] Excluding individuals with propensity scores falling outside a pre-specified range, or reducing large weights to a pre-specified maximum value are common solutions to the insufficient overlap problem.[74,75] However, the methods for resolving insufficient overlap were developed under single-level settings and it is unclear how they would work or how they should be modified in the multilevel context.

4.4 | Variance calculation

Variance estimation remains a somewhat open issue in this literature. Two common ways to estimate the variance of the ATE/ATT with clustered data are to calculate a cluster robust variance, or to perform cluster bootstrapping (resampling clusters from the original sample)[6]; the key is that the within-cluster correlation induced by the clustered structure should be taken care of in some way. Because robust (or "sandwich") variance estimators do not account for the uncertainty in estimating propensity scores, some have recommended the bootstrapping approach in the single-level context (resampling individuals from the original sample, and reestimate propensity scores in each bootstrap sample).[76] However, even in the single-level context, more research is needed to examine the performance of bootstrap variances—and what the resampling units should be (eg, individuals or matched pairs)—under different settings and with different propensity score methods. For example, as mentioned in the previous section, bootstrap SEs are invalid if propensity score matching *with* replacement was done,[46] but Austin and Small[77] found that bootstrap SEs performed well if matching was done *without* replacement. A multilevel structure further complicates the operation of bootstrapping; Li et al[3] suggested cluster bootstrapping to calculate the variance of IPTW-based estimators, but the validity of bootstrapping in the context of matching or stratification with clustered data remains unclear and understudied. Moreover, it is unclear how variance should be estimated if across-cluster matching or a hybrid matching approach was done, given that the variance estimation would potentially need to account for the clusters as well as for the matched pairs/sets, but the matched pairs/sets would not be nested within clusters. Further work is needed to understand the best approaches for variance estimation for the various contexts and methods.

4.5 | Sensitivity analysis

While there are plenty of sensitivity analysis methods in the causal inference literature, especially those that assess the sensitivity of results to an unobserved confounder, these methods have not yet been well established in multilevel settings. Recently, Scott et al[78] extended the sensitivity analysis approaches in Imbens,[22] Carnegie et al,[79] and Middleton et al[80] to two-level settings, allowing the evaluation of robustness to both individual- and cluster-level unmeasured confounding. Further development and application of sensitivity analysis methods for unmeasured confounding at multiple levels could be an important area for future research.

This review presents a framework for examining treatment effects when study units are clustered in some way. Choice of the strategies and modeling approaches presented in this review should be based upon the study objective, study design, and assumptions of how cluster-level characteristics may impact causal inferences. With the increasing use of large-scale, multilevel data for clinical and public health research, it is important for investigators to be aware of such structure and explicate its implications for drawing causal conclusions.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study

ORCID

Ting-Hsuan Chang https://orcid.org/0000-0001-6002-2947

REFERENCES

1. VanderWeele TJ. Confounding and effect modification: distribution and measure. *Epidemiol Method*. 2012;1(1):55-82. doi:10.1515/2161-962X.1004
2. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59(5):437.e1-437.e24. doi:10.1016/j.jclinepi.2005.07.004
3. Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. *Stat Med*. 2013;32(19):3373-3387. doi:10.1002/sim.5786
4. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20-36. doi:10.1002/sim.2739
5. Greifer N, Stuart EA. Matching methods for confounder adjustment: an addition to the epidemiologist's toolbox. *Epidemiol Rev*. 2022;43(1):118-129. doi:10.1093/epirev/mxab003
6. Cafri G, Wang W, Chan PH, Austin PC. A review and empirical comparison of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. *Stat Methods Med Res*. 2018;28(10–11):3142-3162. doi:10.1177/0962280218799540
7. Leite WL, Jimenez F, Kaya Y, Stapleton LM, MacInnes JW, Sandbach R. An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivar Behav Res*. 2015;50(3):265-284. doi:10.1080/00273171.2014.991018
8. Thoemmes FJ, West SG. The use of propensity scores for nonrandomized designs with clustered data. *Multivar Behav Res*. 2011;46(3):514-543. doi:10.1080/00273171.2011.569395
9. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701. doi:10.1037/h0037350
10. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945-960. doi:10.2307/2289064
11. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591-593. doi:10.2307/2287653
12. Robins JM, Greenland S. Causal inference without counterfactuals: comment. *J Am Stat Assoc*. 2000;95(450):431-435. doi:10.2307/2669381
13. Hong G, Raudenbush SW. Evaluating kindergarten retention policy. *J Am Stat Assoc*. 2006;101(475):901-910. doi:10.1198/016214506000000447
14. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1):4-29.
15. Greifer N, Stuart EA. Choosing the estimand when matching or weighting in observational studies [preprint]. arXiv 2021. Accessed April 19, 2022. 10.48550/arXiv:2106.10577
16. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41
17. Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. *Comput Stat Amp Data Anal*. 2011;55(4):1770-1780. doi:10.1016/j.csda.2010.11.008
18. Yang S. Propensity score weighting for causal inference with clustered data. *J Causal Infer*. 2018;6(2):20170027. doi:10.1515/jci-2017-0027
19. He Z. Inverse conditional probability weighting with clustered data in causal inference [preprint]. arXiv; 2018. 10.48550/arXiv.1808.01647. Accessed July 7, 2021.
20. Kim G-S, Paik MC, Kim H. Causal inference with observational data under cluster-specific non-ignorable assignment mechanism. *Comput Stat Data Anal*. 2017;113:88-99. doi:10.1016/j.csda.2016.10.002
21. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J R Stat Soc Ser B*. 1983;45(2):212-218. doi:10.1111/j.2517-6161.1983.tb01242.x
22. Imbens GW. Sensitivity to exogeneity assumptions in program evaluation. *Am Econ Rev*. 2003;93(2):126-132. doi:10.1257/000282803321946921
23. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017;167(4):268-274. doi:10.7326/M16-2607
24. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786
25. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21. doi:10.1214/09-STS313
26. Bryer J. multilevelPSA: multilevel propensity score analysis. R package version 1.2.5; 2018.
27. Griswold ME, Localio AR, Mulrow C. Propensity score adjustment with multilevel data: setting your sites on decreasing selection bias. *Ann Intern Med*. 2010;152(6):393-395. doi:10.7326/0003-4819-152-6-201003160-00010
28. Allison P. *Fixed Effects Regression Models*. Thousand Oaks, CA: SAGE Publications, Inc; 2009. doi:10.4135/9781412993869
29. Kim J, Seltzer M. Causal inference in multilevel settings in which selection processes vary across schools. CSE technical report 708. National Center for Research on Evaluation, Standards, and Student Testing (CRESST); 2007. doi:10.1037/e644002011-001

30. Kim J-S, Steiner PM. Multilevel propensity score methods for estimating causal effects: a latent class modeling strategy. In: van der Ark LA, Bolt DM, Wang W-C, Douglas JA, Chow S-M, eds. *Quantitative Psychology Research*. New York, NY: Springer International Publishing; 2015:293-306. doi:10.1007/978-3-319-19977-1_21
31. Kim J-S, Lim W-C, Steiner PM. Causal inference with observational multilevel data: investigating selection and outcome heterogeneity. In: van der Ark LA, Wiberg M, Culpepper SA, Douglas JA, Wang W-C, eds. *Quantitative Psychology*. New York, NY: Springer International Publishing; 2017:287-308.
32. Suk Y, Kim J-S. Measuring the heterogeneity of treatment effects with multilevel observational data. In: Wiberg M, Culpepper S, Janssen R, González J, Molenaar D, eds. *Quantitative Psychology*. New York, NY: Springer International Publishing; 2019:265-277.
33. Lee Y, Nguyen TQ, Stuart EA. Partially pooled propensity score models for average treatment effect estimation with multilevel data. *J R Stat Soc Ser A*. 2021;184:1578-1598. doi:10.1111/rssa.12741
34. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal*. 2007;15(3):199-236. doi:10.1093/pan/mpl013
35. Austin PC, Stuart EA. Optimal full matching for survival outcomes: a method that merits more widespread use. *Stat Med*. 2015;34(30):3949-3967. doi:10.1002/sim.6602
36. Rosenbaum PR. Overt bias in observational studies. In: Rosenbaum PR, ed. *Observational Studies*. New York, NY: Springer; 2002:71-104. doi:10.1007/978-1-4757-3692-2_3
37. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Stat*. 1993;2(4):405-420. doi:10.1080/10618600.1993.10474623
38. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10(2):150-161. doi:10.1002/pst.433
39. Kim J-S, Steiner PM, Hall C, Thoemmes FJ. *Within-Cluster and across-Cluster Matching with Observational Multilevel Data*. Washington, DC: Society for Research on Educational Effectiveness; 2013.
40. Rickles JH, Seltzer M. A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *J Educ Behav Stat*. 2014;39(6):612-636. doi:10.3102/1076998614559748
41. Stuart EA, Rubin DB. Matching with multiple control groups with adjustment for group differences. *J Educ Behav Stat*. 2008;33(3):279-306. doi:10.3102/1076998607306078
42. Arpino B, Cannas M. Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the apgar score. *Stat Med*. 2016;35(12):2074-2091. doi:10.1002/sim.6880
43. Cannas MC. Matching: matching algorithms for causal inference with clustered data. R package version 2.3.0; 2019.
44. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962-973. doi:10.1111/j.1541-0420.2005.00377.x
45. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol*. 2011;173(7):761-767. doi:10.1093/aje/kwq439
46. Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. *Econometrica*. 2008;76(6):1537-1557.
47. Neuhäuser M, Thielmann M, Ruxton GD. The number of strata in propensity score stratification for a binary outcome. *Arch Med Sci*. 2018;14(3):695-700. doi:10.5114/aoms.2016.61813
48. Orihara S, Hamada E. Determination of the optimal number of strata for propensity score subclassification. *Stat Probab Lett*. 2021;168:108951. doi:10.1016/j.spl.2020.108951
49. Rosenbaum PR. A characterization of optimal designs for observational studies. *J R Stat Soc Ser B*. 1991;53(3):597-610. doi:10.1111/j.2517-6161.1991.tb01848.x
50. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99(467):609-618. doi:10.1198/016214504000000647
51. Xiang Y, Tarasawa B. Propensity score stratification using multilevel models to examine charter school achievement effects. *J Sch Choice*. 2015;9(2):179-196. doi:10.1080/15582159.2015.1028862
52. Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *J Educ Behav Stat*. 2010;35(5):499-531. doi:10.3102/1076998609359785
53. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.
54. Schuler MS, Chu W, Coffman D. Propensity score weighting for a continuous exposure with multilevel data. *Health Serv Outcomes Res Methodol*. 2016;16(4):271-292. doi:10.1007/s10742-016-0157-5
55. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937-2960. doi:10.1002/sim.1903
56. Mercatanti A, Li F. Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Ann Appl Stat*. 2014;8(4):2485-2508.
57. Su YS, Cortina J. What do we gain? Combining propensity score methods and multilevel modeling. Proceedings of the Annual Meeting of the American Political Science Association; 2009; Toronto, Canada.
58. Suk Y, Kang H, Kim J-S. Random forests approach for causal inference with clustered observational data. *Multivar Behav Res*. 2021;56(6):829-852. doi:10.1080/00273171.2020.1808437
59. Athey S, Wager S. Estimating treatment effects with causal forests: an application [preprint]. arXiv. 2019. 10.48550/arXiv.1902.07409. Accessed April 20, 2022.

60. Keele L, Lenard M, Page L. Matching methods for clustered observational studies in education. *J Res Educ Eff*. 2021;14(3):696-725. doi:10.1080/19345747.2021.1875527
61. Stuart EA. Estimating causal effects using school-level data sets. *Educ Res*. 2007;36(4):187-198. doi:10.3102/0013189X07303396
62. Steiner PM. *Matching Strategies for Observational Data with Multilevel Structure*. Washington, DC: Society for Research on Educational Effectiveness; 2011.
63. Zubizarreta JR, Keele L. Optimal multilevel matching in clustered observational studies: a case study of the effectiveness of private schools under a large-scale voucher system. *J Am Stat Assoc*. 2017;112(518):547-560. doi:10.1080/01621459.2016.1240683
64. Leyrat C, Caille A, Donner A, Giraudeau B. Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Stat Med*. 2013;32(19):3357-3372. doi:10.1002/sim.5795
65. Balzer LB, Zheng W, van der Laan MJ, Petersen ML. A new approach to hierarchical data analysis: targeted maximum likelihood estimation for the causal effect of a cluster-level exposure. *Stat Methods Med Res*. 2019;28(6):1761-1780. doi:10.1177/0962280218774936
66. Sobel ME. What do randomized studies of housing mobility demonstrate? *J Am Stat Assoc*. 2006;101(476):1398-1407. doi:10.1198/016214506000000636
67. Eckardt P. Propensity score estimates in multilevel models for causal inference. *Nurs Res*. 2012;61(3):213-223. doi:10.1097/NNR.0b013e318253a1c4
68. Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. *Stat Methods Med Res*. 2012;21(1):55-75. doi:10.1177/0962280210386779
69. Forastiere L, Airoldi EM, Mealli F. Identification and estimation of treatment and interference effects in observational studies on networks. *J Am Stat Assoc*. 2021;116(534):901-918. doi:10.1080/01621459.2020.1768100
70. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546-555. doi:10.1002/pds.1555
71. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346. doi:10.1002/sim.3782
72. Han H, Kwak M. An alternative method in estimating propensity scores with conditional inference tree in multilevel data: a case study. *J Korean Data Inf Sci Soc*. 2019;30:951-966. doi:10.7465/jkdi.2019.30.4.951
73. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol*. 2019;188(1):250-257. doi:10.1093/aje/kwy201
74. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199. doi:10.1093/biomet/asn055
75. Elliott MR. Model averaging methods for weight trimming in generalized linear regression models. *J Off Stat*. 2009;25(1):1-20.
76. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655. doi:10.1002/sim.7084
77. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Stat Med*. 2014;33(24):4306-4319. doi:10.1002/sim.6276
78. Scott M, Diakow R, Hill J, Middleton J. Potential for bias inflation with grouped data: a comparison of estimators and a sensitivity analysis strategy. *Obs Stud*. 2018;4(1):111-149. doi:10.1353/obs.2018.0016
79. Carnegie NB, Harada M, Hill JL. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *J Res Educ Eff*. 2016;9(3):395-420. doi:10.1080/19345747.2015.1078862
80. Middleton JA, Scott MA, Diakow R, Hill JL. Bias amplification and bias unmasking. *Polit Anal*. 2016;24(3):307-323. doi:10.1093/pan/mpw015

How to cite this article: Chang T-H, Stuart EA. Propensity score methods for observational studies with clustered data: A review. *Statistics in Medicine*. 2022;41(18):3612-3626. doi: 10.1002/sim.9437