Journal of Biopharmaceutical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Utilizing stratified generalized propensity score matching to approximate blocked randomized designs with multiple treatment levels

Nathan Corder & Shu Yang

To cite this article: Nathan Corder & Shu Yang (2022) Utilizing stratified generalized propensity score matching to approximate blocked randomized designs with multiple treatment levels, Journal of Biopharmaceutical Statistics, 32:3, 373-399, DOI: 10.1080/10543406.2022.2065507

To link to this article: https://doi.org/10.1080/10543406.2022.2065507

Published online: 19 Jun 2022.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2022, VOL. 32, NO. 3, 373–399
https://doi.org/10.1080/10543406.2022.2065507

Utilizing stratified generalized propensity score matching to approximate blocked randomized designs with multiple treatment levels

Nathan Corder and Shu Yang

Statistics, NC State University, Raleigh, North Carolina, United States

ABSTRACT

Conducting causal inference in settings with more than one treatment level can be challenging. Classical methods, such as propensity score matching (PSM), are restricted to only a binary treatment. To extend propensity score methods beyond a binary treatment, generalized propensity score methods have been proposed, with generalized propensity score matching (GPSM) standing as the multi-level treatment analog to PSM. One drawback of GPSM is it is only capable of emulating a completely randomized trial (CRT) design and not the more efficient blocked randomized trial design. Motivated by the desire to emulate the more efficient design, we expand on GPSM estimating literature and develop a new estimator incorporating relevant stratifying variables into the GPSM framework. We examine the variance estimation methods available for GPSM and demonstrate how to extend the estimator to one where stratifying variables are included. While it would be straightforward to include relevant stratifying variables as covariates in the propensity score estimation, our method provides for researchers to conduct retrospective analyses more consistently with the prospective experiment they would have designed if permitted. Namely, our method permits researchers to approximate a stratified randomized trial as opposed to the CRT otherwise obtainable by GPSM. We apply our proposed method to an analysis of how the number of children in a household affects systolic blood pressure in adults. We conduct a simulation study assessing how the relationship between response, treatment, and strata affect the performance of our method and compare the results to non-stratified GPSM.

ARTICLE HISTORY
Received 19 June 2021
Accepted 21 March 2022

KEYWORDS
causal inference; propensity score; matching; multiple treatments; stratification

1. Introduction

If it was possible to conduct a randomized clinical trial whenever a new question about a proposed treatment, exposure, or intervention was asked, the job of the researcher would be far easier. Randomized clinical trials, by design, guarantee treatment groups are constructed in such a way as to minimize differences in relevant covariates and thus minimize the potential bias incurred when estimating a causal effect. In a completely randomized trial (CRT) the assignment mechanism is consistent across the whole population. When it is expected or desirable for treatment assignment or outcomes to differ depending on categorical pre-treatment strata, a blocked randomized design is preferred. Unfortunately, in practice, researchers frequently must rely on observational data to answer their questions of interest as often randomized clinical trials are impractical or even impossible to conduct. Observational data brings with it new challenges, as the lack of randomization obfuscates treatment effect estimates behind potential confounding between the response and treatment assignment. Thankfully, this concern is not new to researchers, and a multitude of methods have been

CONTACT Shu Yang syang24@ncsu.edu NC State University, 2311 Stinson Drive, 5109 SAS Hall Campus Box 8203, Raleigh, 27695, North Carolina

proposed to address confounding in observational data. Matching estimators are one class of methods routinely used by researchers to address confounding concerns in observational data; see Stuart (2010) for a comprehensive review of the class, and within the class, the subset of propensity score matching (PSM) methods are arguably the most common.

PSM is based on the seminal paper of Rosenbaum and Rubin (1983), wherein the authors demonstrate that conditioning on the propensity score (the probability of receiving treatment given the set of pretreatment covariates) is sufficient to remove treatment assignment bias when estimating treatment effects from observational data. Despite its popularity, critiques of PSM have focused on its inability to emulate the clinical trial that would have been designed if possible. For instance, Rosenbaum and Rubin (1983) considered only the case of binary treatment, but often researchers are interested in either continuous treatments or the contrast between multiple treatments or treatment levels (see Hirano and Imbens (2004); Imbens (2000); Lu et al. (2001); Wu et al. (2018); Yang et al. (2016) for solutions). Here, we use binary treatment to indicate any treatment schema where exactly two levels of treatment exist (as opposed to a multinomial treatment setting where more than two levels of treatment could exist). The most common binary treatment setting would be the case where the researcher is comparing a set of observational units that did receive some treatment to a set of units that did not, though studies comparing two active treatments would also be considered binary. More recently, King and Nielsen (2019) critiqued PSM as being unable to emulate a trial design other than a CRT and showed how 1:1 PSM could actually increase bias if data was pruned beyond the point where the approximate CRT was obtained. King and Nielsen implore researchers to use other matching methods such as Mahalanobis matching (Rubin 1976) or coarsened exact matching (Iacus et al. 2012) that are better able to emulate the superior trial design of a blocked randomized trial. In this article, we set out to solve both critiques (multi-level treatments and blocked randomized design) simultaneously while still constraining ourselves within the familiar class of PSM methods.

In our review of the literature, we have found no method for PSM that has been both extended to multi-level treatments and approximated the blocked randomized trial design that motivated this article. Methods outside of the PSM class have achieved, to some extent, one or both of the goals set out by King and Nielsen (e.g., Iacus et al. 2011; Imai and Ratkovic 2014; Lopez and Gutman 2017; Lunceford and Davidian 2004), but our motivation here is to develop a method within the PSM class to borrow the common understanding of traditional PSM on binary treatments which may aid adoption by researchers.

Related Works. *In the context of our intent to develop a PSM-based method addressing both of King and Nielsen's critiques of traditional PSM, it is still important to note the important contributions of the example alternatives and demonstrate how close they each come to our goal. Lunceford and Davidian* (2004) *demonstrate doubly-robust methods and stratification methods for estimating causal effects in a binary treatment setting, but the stratification implemented is on the propensity score and not on specified pre-treatment strata. Iacus et al.* (2011) *introduce coarsened exact matching which will ensure (to the researcher's level of tolerance) important strata are matched exactly and mimic a blocked randomized trial, but it has not been extended to multiple treatments. Imai and Ratkovic* (2014) *target the estimation of propensity scores in such a manner to optimize balance in the covariates, and feasibly, the method could be tailored to make strata balancing a required part of the optimization. Furthermore, they propose how their method might be extended to multiple treatments. On the other hand, the extension to multiple treatments goes unexamined in their paper, and they intentionally avoid applications to matching methods. Finally, Lopez and Gutman* (2017) *do focus on matching methods in multi-level treatment settings, but any guarantees that important strata will be balanced are not available.*

We build on the multi-level treatment effect estimation results (Yang et al. 2016) devised via generalized propensity score matching (GPSM) and weak unconfoundedness. We introduce to their method the requirement that the GPSM matched pairs must be constrained within the same level of pretreatment strata. This approach is motivated as an attempt to emulate (in the observational setting) the blocked randomized trial design often seen in Phase I/II dose-finding studies within the pharmaceutical industry. Often these earlier phase trials have multiple doses of the same treatment randomized to subjects within relevant pretreatment strata (such as measures of disease severity, geography, or race). When clinical trials move into later stages, study sponsors may reduce the analysis of treatments to a binary setting of the most successful dose from Phase I/II versus standard of care, though the stratified randomization is likely to remain. Here, stratified GPSM still may provide value, as the binary treatment setting is a special case of the generalized propensity score. Motivated to emulate the stratified blocked randomized trial designs, we propose stratified GPSM as a tool to design observational studies. Like non-stratified GPSM, stratified GPSM still estimates the generalized propensity score at the population level, but in the matching step, we constrain matches only to units within the same strata. We establish theoretical guarantees of stratified GPSM including valid inference on average treatment effects and an asymptotically normal estimator. Our findings also demonstrate the stratified GPSM approach is capable of reducing variance among multiple treatment effect contrasts beyond the results from non-stratified GPSM when the chosen strata are relevant to the treatment assignment mechanism. Our results are further bolstered by the recent findings of Wang et al. (2021) who showed in the experimental/clinical trial setting that when an estimator is consistent and asymptotically normally distributed under simple randomization, then it is consistent and asymptotically normally distributed under stratified randomization with equal or smaller asymptotic variance. As non-stratified GPSM has already been shown to be consistent and asymptotically normally distributed under standard continuity assumptions, the introduction of a stratifying variable to the matching processes to mimic a stratified random sample should also lead to a consistent and asymptotically normally distributed estimator with equal or smaller variances to be expected.

This article is laid out as follows. In Section 2, we begin with a review of potential outcomes framework and its extension into multiple treatments. In Section 3, we detail the usage of GPSM to construct causal effect estimates and characterize their application in the presence of stratifying variables. In Section 4, we utilize simulation results to examine the performance of stratified GPSM over non-stratified GPSM. Also in this section, we analyze the risk associated with stratifying on non-relevant variables. In Section 5, we examine stratified GPSM's performance in a real-world setting utilizing the National Health and Nutrition Examination Survey data. We conclude in Section 6 with a discussion of stratified GPSM's applicability in light of our simulation and real-world data findings and make recommendations for further study.

2. Causal inference for multiple-level treatments

2.1. Potential outcomes

The concept of potential outcomes is ingrained in the causal inference literature (Rubin 1974). We mimic the same framework but with the modifications necessary to extend the potential outcomes to multiple treatments (Imbens 2000; Lechner 2001; McCaffrey et al. 2013). Let W_i represent the treatment for individual i from the set of available treatments $\mathbb{W} = \{1, \ldots, T\}$, where $T \geq 2$. Let $Y_i(w)$ be the potential response values for individual i at each of $w \in \mathbb{W}$ treatment levels. We assume responses within and between units do not interfere, following the stable-unit-treatment-value assumption of Rubin (1980) Commonly, only one treatment level is observable for any individual at any time, so we let $Y_i = Y_i(W_i)$ be the realized and observed response. Allow X_i to be a vector of pretreatment covariates potentially relevant to the assignment of W_i. Finally, let S_i be a vector of stratifying, pre-treatment, categorical variables which may or may not affect treatment assignment but are anticipated to affect the response.

Remark 1. Our notation may imply S_i to be composed of a separate set of categorical variables than X_i, but this is not necessarily the case. We do require S_i to be categorical, but the values of S_i may be derived from one or more covariates in X. For instance, if investigating a medical treatment, S_i might contain a categorical variable for patient sex, but might also contain a variable for disease prognosis (i.e.: $S_{i;severe} \in \{poor, fair, good\}$) derived from pre-treatment lab values contained in X_i.

We next assume the observed set of data $\{(X_i, W_i, S_i, Y_i)\}_{i=1}^N$ are independent draws from the population of interest. Our focus in the paper is on average treatment effects between multiple treatment levels. Because average treatment effect in most causal inference literature refers specifically to the effect of being treated versus not being treated, we will from here-onward instead use the term average treatment contrast to refer to the average treatment effect between any two treatment levels w and w' and define the average treatment contrast

$$\tau(w, w') = E[Y_i(w') - Y_i(w)]. \tag{1}$$

Note this expectation is taken at the population level. Under the potential outcomes framework, this expectation exists even for individuals not treated with either treatment level w or w'. Commonly, if researchers want to estimate a treatment contrast, they might do so by including only patients in either treatment group (Lechner 2001) in a pairwise manner. We will let

$$\tau_{PAIR}(w, w') = E[Y_i(w') - Y_i(w) | W_i \in \{w, w'\}]$$

denote the average treatment contrast estimated via pairwise conditional estimation. While convenient, pairwise conditional estimation of average treatment contrasts poses two problems. First, the estimates are non-comparable in that the populations are not the same for $\tau_{PAIR}(w, w')$ and $\tau_{PAIR}(w', w'')$. Secondly, the estimates are non-transmutable in that $\tau_{PAIR}(w, w') - \tau_{PAIR}(w', w'') \neq \tau_{PAIR}(w, w'')$. Of these two problems, the first is more important as it risks a fundamental change to the estimand that may go unnoticed by the researcher. Recall under the potential outcomes framework, the (population) ATE is defined in (1) as the expected difference between the counterfactuals $Y_i(w)$ and $Y_i(w')$ – which is to say the expected difference at the population level between the responses when *all* subjects are treated with $W = w$ and the responses when *all* subjects are treated with $W = w'$. When estimating the population ATE via pairwise treatment contrasts, each pairwise ATE changes the population of interest only to subjects observed treated with *only* the treatment levels in the current contrast. To put this in practical terms suppose a doctor wanted to know if a particular surgical intervention was superior to some prescription-based (R_x) therapy. The pairwise contrast $\tau_{PAIR}(Surgery, R_x)$ ignores all counterfactuals of patients in the population and sample who had been eligible for therapy but who have not yet undergone either therapy. Combine this problem with the second (that estimates are non-transmutable) it is conceivable that the collection of pairwise contrasts could lead to a paradoxical outcome where every therapy is better than at least one other in some sub-population. Since the population of patients in each pairwise contrasts changes, it would be possible that all three statements below are true:

1. Surgery is better than R_x among patients with either treatment.
2. R_x is better than doing nothing in patients who did not undergo surgery.
3. Doing nothing is better than surgery among patients that did not undergo R_x therapy.

Unless careful attention is paid to the reshuffling of populations under examination in each pairwise contrast, the pairwise results might be used to inform a (combined) population-level decision to change nothing when their may actually be a superior treatment protocol at the

(combined) population level. Therefore, we must find a way to estimate potential outcomes for all treatments for all individuals simultaneously in order to take the estimate at the population level.

Remark 2. We include pairwise PSM as a juxtaposing method only because of its common use in practice. Yang et al. (2016) has already demonstrated non-stratified GPSM's superiority to pairwise PSM.

2.2. Generalized propensity scores

The approach put forth in Rosenbaum and Rubin (1983) is a good starting point for estimating multiple potential outcomes. In their binary treatment setting they defined the propensity score as $e(x) = Pr(W_i = 1|X_i = x)$. The authors show under strong unconfoundedness (i.e., $Y_i(1), Y_i(0) \perp\!\!\!\perp W_i|X_i$ where $\perp\!\!\!\perp$ denotes conditional independence) and sufficient overlap (i.e., $e(x)$ is bounded away from 0 and 1 for all x such that $f(x) > 0$) that

$$E[E\{Y|W = 1, e(X)\} - E\{Y|W = 0, e(X)\}|e(X)] = E\{Y(1) - Y(0)\},$$

thus reducing the need to condition on the full covariate space X to only the scalar balancing score $e(X)$. Imbens (2000) expanded the propensity score definition to allow multiple treatments under the Generalized Propensity Score (GPS), of which the binary propensity score $e(X)$ is a special case:

Definition 1 (Generalized Propensity Score). *The generalized propensity score is the probability of receiving treatment level w conditional on the pretreatment covariates x*:

$$p(w|x) = Pr(W_i = w|X_i = x). \tag{2}$$

Under the GPS, strong unconfoundedness would require $\{Y_i(1), \ldots, Y_i(T)\} \perp\!\!\!\perp W_i|X_i$, which is a complex assumption to make when the number of treatment levels is high. It requires the researcher to assume their available covariates are sufficient to remove all confounding between treatment assignment and all T potential outcomes simultaneously. To gather all relevant covariates informative of *any* treatment assignment may drastically increase the number of measured covariates. Similarly, unlike the binary case where $e(X)$ is a scalar balancing score, the analogous form of conditioning on the GPS in a multiple treatment level setting would result in

$$E(E[Y(1) - Y(0)|\{p(1|X), \ldots, p(T - 1|X)\}]).$$

Ironically, in the setting where there are more treatment levels than covariates, matching on the GPS while requiring strong unconfoundedness would actually increase the dimensionality of the problem instead of decreasing it. Thankfully, Imbens (2000) also introduced the concept of weak unconfoundedness to retain the benefits of conditioning on a scalar balancing score. We adopt the weak unconfoundedness assumption here-onward. To assume weak unconfoundedness, let $D_i(w) \in \{0, 1\}$ be treatment indicators constructed as

$$D_i(w) = \begin{cases} 1 & \text{if } W_i = w, \\ 0 & \text{otherwise.} \end{cases}$$

Assumption 1 (Weak unconfoundedness). *Let treatment level w be weakly unconfounded with response* $Y(w)$ *w conditional on the pretreatment covariates as in* $D_i(w) \perp\!\!\!\perp Y_i(w)|X_i$.

Under weak unconfoundedness, the expected value for any individual for any treatment can be obtained by

$$E\{Y_i(w)\} = E[E\{Y_i|W_i = w, p(w|X_i)\}], \tag{3}$$

and the average treatment contrast becomes

$$E\{Y_i(w') - Y_i(w)\} = E[E\{Y_i|W_i = w', p(w'|X_i)\}] - E[E\{Y_i|W_i = w, p(w|X_i)\}]. \quad (4)$$

Assuming weak unconfoundedness (instead of strong unconfoundedness) importantly limits the focus of estimation to only one conditional on a scalar balancing equation, a problem well suited for matching-based estimators.

As in the binary propensity score case, GPS must also assume sufficient overlap for equation (3) to be estimable. Otherwise, there will be points in the covariate space where $E[Y_i(w)]$ can not exist for one or more treatment levels. To extend the sufficient overlap assumption to the GPS notation, we make the following assumption.

Assumption 2 (Sufficient Overlap). $p(w|x) > 0 \quad w, x.$

3. Propensity score matching for multi-level treatments

3.1. Generalized propensity score matching

GPSM (Yang et al. 2016) extends traditional PSM on binary treatments to multi-level treatments. Like other matching methods, GPSM seeks to impute the non-observed potential outcomes by matching units in one treatment group to others with similar covariates in the other treatment group(s). The matching is done in such a fashion as to minimize covariate imbalance among the matches. Unlike other PSM-based methods for multi-level treatment contrast estimation that either match pairwise between treatments (Lechner 2001) or match using all estimated GPS simultaneously (Rassen et al. 2013, 2011; Tu et al. 2013), GPSM uses the weak unconfoundedness assumption to preserve the dimension reduction features seen in PSM for binary treatments to a scalar balancing score while still allowing for comparable and transmutable treatment contrast estimates. To do so, GPSM begins the same way as other GPS-based methods, by positing a model for GPS in equation (2) and simultaneously estimating all GPS for all treatments and all individuals. This is commonly done via a multinomial regression model, though other models can be incorporated. Matching takes place with replacement along the GPSM matching function

$$m_{\text{GPS}}(w, p) = \underset{j:W_j=w}{\arg\min} ||p(w|X_j) - p|| \quad (5)$$

leading to an estimate for equation (3) as

$$\hat{Y}_i(w) = Y_{m_{\text{GPS}}(w,p(w|X_i))} \overset{\text{def}}{=} Y^{(w)}_{\text{GPS}_i}, \quad (6)$$

where $\overset{\text{def}}{=}$ is used to define an equivalent shorthand representation. In words, GPSM imputes the unobserved potential outcome for unit i and treatment w as the observed outcome for unit j where $W_j = w$ and unit j has the closest estimated GPS for treatment level w to the estimated GPS of unit i for treatment level w.

Given the imputed potential outcome values via GPSM through equation (6), we can define the GPSM treatment effect estimate for a given contrast as

$$\hat{\tau}_{GPSM}(w, w') = N^{-1} \sum_{i=1}^{N} \left\{ \hat{Y}_i(w') - \hat{Y}_i(w) \right\} = N^{-1} \sum_{i=1}^{N} \left\{ Y^{(w')}_{\text{GPS}_i} - Y^{(w)}_{\text{GPS}_i} \right\}.$$

Note that the index i is over the whole population and not just the subset where $W_i \in (w, w')$. This means all treatment contrast estimates share the same support and are thus directly comparable and transmutable. It also means model dependence increases as there is further reliance on a correctly specified GPS model when individuals in neither treatment group being contrasted are contributing to the treatment effect estimate. Yang et al. (2016) found the performance of GPSM deteriorates when the GPS model is incorrectly specified. We continue forward assuming the GPS model is correct.

Assumption 3 (Correctly specified GPS model). Let $\hat{p}(w|X_i;\hat{\theta})$ be the estimated GPS model under a specified parameterization θ. We assume $\hat{p}(w|X_i;\hat{\theta})$ is consistent for $p(w|X_i)$.

3.2. Stratified GPSM

One of the critiques in King and Nielsen (2019) of traditional PSM on binary treatments is that at some point as more and more bad matches are pruned, PSM approximates a CRT, and any further pruning will increase the imbalance. Even when no pruning of bad matches is done, Wang (2020) highlights PSM can still increase imbalance in covariates, noting in practice, when a covariate is originally near balanced, PSM is more likely to worsen its balance rather than to improve it. This feature is of particular concern if we return to our motivating example of a hypothetical Phase I/II dose-finding study where the intended randomized clinical trial would have randomized within relevant pretreatment strata. If pretreatment strata are already balanced at the population level, pairs matched via PSM/GPSM may cross strata and subsequently increase imbalance. To control for this risk, we propose modifying GPSM to force matches to be constrained by desired prespecified strata, in a manner we term stratified GPSM.

Our method is not the first attempt to formalize the inclusion of stratifying variables into a PSM-based method. Included within the approach of Rubin and Thomas (2000) is a method by which researchers can deploy PSM while also matching on a set of special prognostic variables. In their work, PSM's role was limited to reducing the set of candidate matches (via a caliper) prior to a Mahalanobis match on the prognostic variables. Moreover, the combination of the two matching methods (PSM and Mahalanobis matching) was primarily done as a preprocessing step to obtain a matched sample for some subsequent causal analysis to utilize. Nonetheless, if one were to allow those prognostic variables to only be categorical and the matching results were used as the primary causal analysis instead of only a preprocessing step, one would arrive at a method similar in nature to our own about to be discussed. This application of Rubin and Thomas (2000) would better emulate a blocked randomized design for the binary treatment setting than would traditional PSM, though it does not address other critiques in King and Nielsen (2019) such as what would happen when observations are pruned. As pairwise comparisons of treatment contrasts under a multinomial treatment setting would necessitate removing observations not treated with either level of treatment within the contrast. As a result, an estimate of the population-level treatment effect can not be performed under pairwise comparisons of a multinomial treatment. Thus, if a population-level treatment effect estimate is required, further work building on Rubin and Thomas (2000) needs developed.

To implement stratified GPSM, we replace the GPSM matching function in equation (5) with the stratified GPSM matching function

$$m_{\text{STRAT}}(w,p,s) = \underset{j:W_j=w,S_j=s}{\arg\min}\ ||p(w|X_j) - p|| \tag{7}$$

and equation (6) with

$$\hat{Y}_i(w|S_i) = Y_{m_{\text{STRAT}}(w,p(w|X_i),S_i)} \overset{\text{def}}{=} Y^{(w|S_i)}_{\text{STRAT}_i}. \tag{8}$$

such that $\hat{Y}_i(w|S_i)$ is the imputed value of $Y_i(w)$ under its observed strata S_i. The inclusion of the conditioning on S_i is for notational purposes only and is used to signify that the stratified GPSM estimate is constraining the set of available match candidates to those that are observed with the same level of strata as Y_i. The notational addition is only relevant for the estimated value $\hat{Y}(w|S_i)$ where S_i is controlling the available matches. The potential outcomes $\{Y_i(1),\ldots,Y_i(T)\}$ remain unchanged. Note the GPS model $p(w|x)$ is still constructed at the population level and not impacted by strata. This reduces the need to construct multiple propensity models covering different sub-populations. The choice to not recalculate the GPS model within each stratum has other advantages which will become evident in Section 3.2.1.

Once imputed potential outcomes are obtained, we can estimate the average treatment effect for different treatment contrasts via stratified GPSM

$$\hat{\tau}_{STRAT}(w, w') = N^{-1} \sum_{i=1}^{N} \left\{ \hat{Y}_i(w'|S_i) - \hat{Y}_i(w|S_i) \right\}$$

$$= N^{-1} \sum_{i=1}^{N} \left\{ Y_{\text{STRAT}_i}^{(w'|S_i)} - Y_{\text{STRAT}_i}^{(w|S_i)} \right\}.$$

This estimator is easy to implement in practice and functions similarly to placing a caliper on the GPSM matching function where pairs are penalized with an added distance of ∞ if the candidate unit is from a different strata than the unit needing imputed.

3.2.1. Asymptotic distribution and variance estimation

Because of our choice to not recalculate the GPS model within each stratum, variance estimation of $\hat{\tau}_{STRAT}$ is made much simpler. Following Yang et al. (2016), we introduce the following assumptions:

Assumption 4. We have a random sample of size N from a large population

Assumption 5. *Let* $\mu(w, x, s)$ be the conditional response means given the treatment, the covariates, and specified strata. X has a continuous distribution with compact support $\mathbb{X}$ with a continuous density function. $\mu(w, x, s)$ is Lipschitz-continuous in x. $E\left\{ |Y_i|^{2+\delta} | W_i = w, p(w|X_i) = p, S_i = s \right\}$ is uniformly bounded for some $\delta > 0$.

Assumption 6. Let $\bar{\mu}(w, p, s)$ be the conditional response means given the treatment, the GPS, and specified strata. $p(w|X_i)$ has a continuous distribution with compact support $\left[\underline{p}, \overline{p}\right]$ with a continuous density function. $\bar{\mu}(w, p, s)$ is Lipschitz-continuous in p. $E\left\{ |Y_i|^{2+\delta} | W_i = w, p(w|X_i) = p, S_i = s \right\}$ is uniformly bounded for some $\delta > 0$.

Assumption 4 ensures we have a sample of sufficient size to evaluate the asymptotic result and are not constrained by finite population concerns. Because components of the variance are calculated at the strata-treatment level, a larger N will be needed to satisfy Assumption 4 in stratified GPSM than in the non-stratified setting such that there are still no finite populations concerns in the smallest strata-treatment combination. Assumption 5 is used to invoke the central limit theorem when balance on covariates is achieved. Likewise, Assumption 6 is used to establish the limiting distribution of the stratified GPSM via the central limit theorem. Our introduction of a requirement to match within selected pretreatment strata has no effect on Assumption 1 or 2, because we still estimate the GPS at the population level. Assumption 3 does change from Yang et al. (2016), replacing $\mu(w, x)$, the conditional response mean given the covariates, with $\mu(w, x, s)$ Similarly, Assumption 4 must replace $\bar{\mu}(w, p)$ with $\bar{\mu}(w, p, s)$. The key finding here, however, is the continuity requirement in each assumption was within x and p respectively. This means the introduction of strata does not violate any assumptions underlying the limiting distribution of non-stratified GPSM. Thus, it is straight forward to extend the findings of Yang et al. (2016) to stratified GPSM with only need to make the appropriate notational changes to index the finite set of strata.

Theorem 1. Under Assumptions 1–6, the stratified GPSM estimator is root-N consistent and asymptotically normal,

$$N^{-1/2}\{\hat{\tau}_{STRAT}(w,w') - \tau(w,w')\} \to \mathcal{N}(0, \sigma^2_{STRAT}(w,w'))$$

in distribution as $N \to \infty$, where

$$\begin{aligned}\sigma^2_{STRAT}(w,w') = \quad & E_{X,S}\left[\{\bar{\mu}(w', p(w'|X_i), S_i) - \bar{\mu}(w, p(w|X_i), S_i) - \tau(w,w')\}^2\right] \\ & + E_{X,S}[\bar{\sigma}^2(w, p(w|X_i), S_i) \times \{3/(2p(w|X_i)) - p(w|X_i)/2\}] \\ & + E_{X,S}[\bar{\sigma}^2(w', p(w'|X_i), S_i) \times \{3/(2p(w'|X_i)) - p(w'|X_i)/2\}]\end{aligned} \tag{9}$$

with $\bar{\sigma}^2(w,p,s) = \mathbb{V}_{Y|W,p,S}[W_i = w, p(w|X_i) = p, S_i = s]$ under the true GPS model.

By forcing matches within relevant strata, we would expect in many instances for variability in Y to improve, in much the same way. Wang et al. (2021) showed to be true for experimental data when applying a consistent estimator to data derived from a stratified randomization scheme to that of a simple randomization scheme. In fact, in stratified GPSM, if the inequality $\bar{\sigma}^2\{w, p(w|X_i), S_i\} \leq \bar{\sigma}^2\{w, p(w|X_i)\}$ does hold when S_i is derived based on some variables in X_i, then the inequality $\sigma^2_{STRAT}(w) \leq \sigma^2_{NON-STRAT}(w)$ will hold as well. See the Appendix for a proof.

Commonly, the true GPS model is not known and must be estimated. Suppose we let θ parameterized the proposed estimated GPS model as in Assumption 3 be estimated by $\hat{p}(w|X_i;\hat{\theta})$. Let I_θ be the information matrix from estimating θ via the chosen GPS model and

$$\hat{\tau}_{STRAT;\hat{\theta}}(w,w') = N^{-1}\sum_{i=1}^{N}\left\{Y^{(w'|S_i)}_{\text{STRAT}_i;\hat{\theta}} - Y^{(w|S_i)}_{\text{STRAT}_i;\hat{\theta}}\right\}.$$

be the estimated stratified GPSM treatment contrast estimate under $\hat{\theta}$ and

$$Y^{(w|S_i)}_{\text{STRAT}_i;\hat{\theta}} = Y_{m_{\text{STRAT}}(w,p(w|X_i;\hat{\theta}),S_i)}.$$

Also define $c(w,w')$ as

$$\begin{aligned}c(w,w') = E[Cov\{X_i, \mu(w', X_i, S_i)|p(w'|X_i;\theta)\} \times p'(w'|X_i;\theta)/p(w'|X_i;\theta)] - \\ E[Cov\{X_i, \mu(w, X_i, S_i)|p(w|X_i;\theta)\} \times p'(w|X_i;\theta)/p(w|X_i;\theta)]\end{aligned}$$

where $p'(w|X_i;\theta)$ denotes the derivative of $p(w|X_i;\theta)$.

Theorem 2. Under Assumptions 1–6, the stratified GPSM estimator with the estimated GPS is root-N consistent and asymptotically normal,

$$N^{-1/2}\left\{\hat{\tau}_{STRAT;\hat{\theta}}(w,w') - \tau(w,w')\right\} \to \mathcal{N}(0, \sigma^2_{STRAT;\hat{\theta}}(w,w'))$$

in distribution as $N \to \infty$, where $\sigma^2_{STRAT;\hat{\theta}}(w,w') = \sigma^2_{STRAT}(w,w') - c(w,w')^T I_\theta c(w,w')$.

Our estimator of $\sigma^2_{STRAT;\hat{\theta}}(w,w')$ follows Abadie and Imbens (2016) and Yang et al. (2016) but where responses were previously imputed using only the GPS and covariates, the imputed values must be constrained within strata. Most importantly, in estimating the covariance contributing to $c(w,w')$ and the within treatment variance $\bar{\sigma}^2(w, p(w|X_i), S_i)$, the estimation matches within treatment and within strata as in

$$\widehat{Cov}\left\{X_i, \mu(w, X_i, S_i)|p(w|X_i;\hat{\theta})\right\} = \frac{1}{L-1}\sum_{j\in\mathcal{S}_L(i,\hat{\theta})}\left(X_j - \frac{1}{L}\sum_{k\in\mathcal{S}_L(i,\hat{\theta})}X_k\right)\left(Y_j - \frac{1}{L}\sum_{k\in\mathcal{S}_L(i,\hat{\theta})}Y_k\right),$$

and

$$\bar{\tilde{\sigma}}^2(w, p(w|X_i, \hat{\theta}), S_i) = \frac{1}{L-1} \sum_{j \in \mathcal{S}_L(i,\hat{\theta})} \left(Y_j - \frac{1}{L} \sum_{k \in \mathcal{S}_L(i,\hat{\theta})} Y_k \right)^2,$$

where $\widehat{Cov}\left\{X_i, \mu(w, X_i, S_i)|p(w|X_i; \hat{\theta}\right\}$ estimates $Cov(X_i, \mu(w, X_i, S_i)|p(w|X_i; \theta, \bar{\tilde{\sigma}}^2(w, p(w|X_i), S_i)$ estimates $\bar{\sigma}^2(w, p(w|X_i), S_i)$, and $\mathcal{S}_L(i, \theta)$ indexes a set of L nearest neighbor matches for unit i within treatment and strata. $\mathcal{S}_L(i, \theta)$ is defined

$$\mathcal{S}_L(i, \theta) = \Big[j = 1, \ldots, N : W_j = W_i, S_j = S_i,$$

$$\sum_{k: W_k = W_i, S_k = S_i} \mathbb{I}(\cdot)\{|\hat{\mu}(\mathbb{W}_j|\mathbb{X}_k \not\Rightarrow \theta) - \hat{\mu}(\mathbb{W}_i|\mathbb{X}_i \not\Rightarrow \theta)| \leq |\hat{\mu}(\mathbb{W}_j|\mathbb{X}_j \not\Rightarrow \theta) - \hat{\mu}(\mathbb{W}_j|\mathbb{X}_i \not\Rightarrow \theta)|\}],$$

where $\mathbb{I}(\cdot)$ is an indicator function returning 1 if the given condition $(\cdot)$ is true and 0 if false. The resulting estimator for the variance of $\sigma^2_{STRAT;\hat{\theta}}(w, w')$ is thus

$$\begin{aligned} \hat{\sigma}^2_{STRAT;\hat{\theta}}(w, w') = & \sum_{i=1}^{N} \left[\left\{ Y^{(w'|S_i)}_{\text{STRAT}_i;\hat{\theta}} - Y^{(w|S_i)}_{\text{STRAT}_i;\hat{\theta}} - N^{-1} \sum_{i=1}^{N} (Y^{(w'|S_i)}_{\text{STRAT}_i;\hat{\theta}} - Y^{(w|S_i)}_{\text{STRAT}_i;\hat{\theta}}) \right\}^2 \right] \\ & + \sum_{i=1}^{N} \left[\bar{\tilde{\sigma}}^2(w, p(w|X_i, \hat{\theta}), S_i) \times \left\{ 3/(2p(w|X_i; \hat{\theta})) - p(w|X_i; \hat{\theta})/2 \right\} \right] \\ & + \sum_{i=1}^{N} \left[\bar{\tilde{\sigma}}^2(w', p(w'|X_i; \hat{\theta}, \hat{\theta}), S_i) \times \left\{ 3/(2p(w'|X_i; \hat{\theta})) - p(w'|X_i; \hat{\theta})/2 \right\} \right] \\ & - \hat{c}(w, w')^T I_{\hat{\theta}} \hat{c}(w, w'), \end{aligned} \tag{10}$$

where

$$\hat{c}(w, w') = N^{-1} \sum_{i=1}^{N} \sum_{s=1}^{S} \mathbb{I}(\cdot)(S_i = s) \Big[\left\{ \widehat{Cov}(X_i, \mu(w', X_i, s)|p(w'|X_i; \hat{\theta})) \times p'(w'|X_i; \hat{\theta})/p(w'|X_i; \hat{\theta}) \right\} -$$

$$\left\{ \widehat{Cov}(X_i, \mu(w, X_i, s)|p(w|X_i; \hat{\theta})) \times p'(w|X_i; \hat{\theta})/p(w|X_i; \hat{\theta}) \right\} \Big].$$

4. Simulation study

To evaluate the extent to which constraining GPS matches to within pretreatment strata affects the precision and accuracy of treatment contrast estimates, we chose to emulate the simulation set-up of Yang et al. (2016) but with the addition of a stratifying variable introduced with varying levels of dependence on both the response and propensity scores. We examine the differences in stratified GPSM and non-stratified GPSM in the settings where

1. Y and W depend on S
2. $Y \perp\!\!\!\perp S$ but W depends on S
3. $Y \perp\!\!\!\perp S$ and $W \perp\!\!\!\perp S$.

We furthermore seek to confirm the effect of Stratified GPSM on the relevant strata variable beyond what might occur from stratifying on random noise. For the remainder of our simulation results, we will abbreviate Stratified GPSM as (STRAT), non-stratified GPSM as (NON-STRAT), and randomly stratified GPSM as (RAND).

In each of the three design settings we construct a finite population (X, W, S, Y) of size N = 100,000 and take samples of size n = 2,000. Within the population $(X_{1i}, X_{2i}, X_{3i}, X_{Si})$ are multivariate normal with means (1,2,1,1) and covariance matrix

$$\Sigma_X = \begin{pmatrix} 1 & 0.7 & 0.4 & 0.1 \\ 0.7 & 2 & 1 & -1 \\ 0.4 & 1 & 1 & -0.5 \\ 0.1 & -1 & -0.5 & 1 \end{pmatrix}.$$

Stratifying variable S_i is derived from $X_{S,i}$ as

$$S_i = \begin{cases} 1 & \text{if } X_{S,i} < -0.5 \\ 2 & \text{if } -0.5 <= X_{S,i} < 0.5 \\ 3 & \text{if } 0.5 <= X_{S,i}. \end{cases}$$

When evaluating STRAT, $S = (S_i)_{i=1}^N$. When evaluating NON-STRAT, $S = (1)_{i=1}^N$. When evaluating RAND, $S = (S_i^*)_{i=1}^N$, where S_i^* is a discrete uniform variable drawing from $\{1, 2, 3\}$. Additional covariates $X_{4i} \sim U[-3, 3]$, $X_{5i} \sim \chi^2(1)$, and $X_{6i} \sim$ Bernoulli(0.5) are also generated, resulting in $X_i^T = (1, X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}, X_{Si})$. Treatment groups are formed via multinomial regression

$$D_i(1), D_i(2), D_i(3)) \sim \text{Multinom}\,(p(1|X_i), p(2|X_i), p(3|X_i)),$$

where $D_i(w) = 1$ indicates unit i belongs to treatment group $W_i = w$ and

$$p(w|X_i) = \exp(X_i^T \beta_w) / \sum_{w'=1}^{3} \exp(X_i^T \beta_{w'})$$

with parameter vectors $\beta_1^T = 0.5 \times (1, 0, 0, 0, 0, 0, 0, \gamma_1)$, $\beta_2^T = 0.1 \times (0, 1, 1, 1, 1, 1, 1, \gamma_2)$, $\beta_3^T = 0.1 \times (0, 1, 1, 1, -1, -1, -1, \gamma_3)$, and $\gamma^T = (\gamma_1, \gamma_2, \gamma_3)$ is the set of coefficients controlling the impact of the stratifying variable X_S on the propensity scores. The outcome $Y_i(w) = X_i^T \alpha_w + \in_i$ where $\in_i \sim \mathcal{N}(0, 1)$, $\alpha_1^T = (0, 1, 1, 1, -1, 1, 1, \eta_1)$, $\alpha_2^T = (0, 2, 1, 3, 2, 1, 1, \eta_2)$, $\alpha_3^T = (0, 3, 1, 2, -3, -1, -2, \eta_3)$, and $\eta^T = (\eta_1, \eta_2, \eta_3)$ is the set of coefficients controlling the impact of the stratifying variable X_S on Y. Each simulation setting is replicated 2,000 times. Within each simulation we evaluate bias, MSE, and coverage under both the naive and corrected variance estimates. By the naive variance estimate we mean the estimate based on Abadie and Imbens (2006) only taking into account the number of times an observation is used in matching. By the corrected we mean the variance estimate derived in Section 3.2.1.

4.1. *Simulation setting: Y and W* depend on *S*

Let $\gamma^T = (0, -2, 2)$ and $\eta^T = (2, 1, -2)$. Under these settings, the population level treatment, strata, and response levels are displayed in Table 1. Note that while the marginal distributions can be approximately considered uniform among treatments and strata, within strata or treatments, the distributions are not uniform.

For the results depicted in Table 2, the same pattern emerges for all three treatment contrasts. In all cases, STRAT, NON-STRAT, and RAND are all approximately unbiased and well covered under the corrected variances. Under the naive variance estimate, all methods are over-covered with STRAT

Table 1. Population share and average response by treatment and strata for Setting 1.

	% of Population				*Average Response*			
Strata	W = 1	W = 2	W = 3	Total	W = 1	W = 2	W = 3	Total
S = 1	11.8%	11.3%	7.9%	*31.0%*	3.652	7.620	5.625	*5.606*
S = 2	13.2%	11.9%	13.0%	*38.1%*	7.381	11.365	6.155	*8.207*
S = 3	9.6%	7.8%	13.5%	*30.9%*	10.843	14.877	6.916	*10.144*
Total	*34.6%*	*31.0%*	*33.4%*	*100.0%*	*7.080*	*10.882*	*6.332*	*8.001*

Table 2. Treatment contrast results when Y and W depend on S.

Treatment Contrast	Method	Bias	MSE	Coverage (Naive)	Coverage (Corrected)
$\tau(1,2) =$ 2.98	STRAT	0.007	0.038	96.8%	94.3%
	NON-STRAT	0.009	0.046	99.2%	95.5%
	RAND	0.003	0.045	99.4%	95.3%
$\tau(1,3) =$ −2.48	STRAT	0.022	0.042	98.1%	94.7%
	NON-STRAT	0.013	0.047	98.7%	95.0%
	RAND	0.023	0.048	98.5%	94.2%
$\tau(2,3) =$ −5.46	STRAT	0.015	0.079	96.4%	94.8%
	NON-STRAT	0.004	0.092	97.7%	94.9%
	RAND	0.021	0.097	97.7%	94.0%

Table 3. Population share and average response by treatment and strata for Setting 2.

	% of Population				*Average Response*			
Strata	W = 1	W = 2	W = 3	Total	W = 1	W = 2	W = 3	Total
S = 1	11.8%	11.3%	7.9%	*31.0%*	3.954	7.795	5.482	*5.750*
S = 2	13.2%	11.9%	13.0%	*38.1%*	5.400	10.376	8.190	*7.908*
S = 3	9.6%	7.8%	13.5%	*30.9%*	6.642	12.790	11.316	*10.231*
Total	*34.6%*	*31.0%*	*33.4%*	100.0%	*5.260*	*10.041*	*8.795*	7.958

being the closest to nominal coverage under the naive estimate each time. Likewise, Stratified GPSM has the smallest MSE in all contrasts by 12–17%. Interestingly, conducting GPSM under random strata performed reasonably well. This observation will be addressed in more detail in Section 4.4.

4.2. *Simulation setting*: $Y \perp\!\!\!\perp S$ *but* W depends on S

Let $\gamma^T = (0, -2, 2)$ and $\eta^T = (0, 0, 0)$. As the choice of γ is the same between Simulation Setting 1 and Simulation Setting 2, the percentage of the population in each treatment/strata is the same as Table 1; however, the changes in η cause an update in the response distribution as displayed in Table 3. Despite no remaining explicit relationship between Y and S, there is still a clear pattern observed where Y tends to be greater for higher strata. This would suggest that even without an explicit link between the response and the strata, STRAT still may provide value over NON-STRAT. From the results in Table 4 that does prove true.

Table 4. Treatment contrast results when $Y \perp\!\!\!\perp S$ but W depends on S.

Treatment Contrast	Method	Bias	MSE	Coverage (Naive)	Coverage (Corrected)
$\tau(1,2) =$ 3.98	STRAT	0.006	0.032	96.8%	94.3%
	NON-STRAT	0.006	0.038	98.9%	95.1%
	RAND	0.003	0.037	98.9%	94.7%
$\tau(1,3) =$ 1.51	STRAT	0.015	0.034	98.5%	95.4%
	NON-STRAT	0.008	0.036	99.0%	94.9%
	RAND	0.021	0.036	99.2%	94.9%
$\tau(2,3) =$ −2.46	STRAT	0.009	0.070	96.7%	95.1%
	NON-STRAT	0.002	0.076	98.0%	95.5%
	RAND	0.018	0.079	98.1%	95.1%

Table 5. Population share and average response by treatment and strata for Setting 3.

	% of Population				*Average Response*			
Strata	W = 1	W = 2	W = 3	Total	W = 1	W = 2	W = 3	Total
S = 1	11.8%	11.0%	8.1%	*30.9%*	3.935	7.902	5.341	*5.716*
S = 2	13.1%	14.4%	10.6%	*38.1%*	5.415	10.286	8.263	*8.044*
S = 3	9.8%	12.0%	9.1%	*30.9%*	6.695	12.660	11.501	*10.427*
Total	*34.7%*	*37.4%*	*27.8%*	*100.0%*	*5.272*	*10.346*	*8.473*	*8.061*

Again, STRAT, NON-STRAT, and RAND are all approximately unbiased and well covered under the corrected variances. Under the naive variance estimate, all methods are over-covered with STRAT being the closest to nominal coverage under the naive estimate each time. Stratified GPSM again has the smallest MSE in all contrasts, but this time the gain over Non-Stratified GPSM has shrunk to 6–15%. RAND continues, confoundingly, to perform reasonably well, even slightly outperforming Non-Stratified GPSM.

4.3. *Simulation setting:* $Y \perp\!\!\!\perp S$ and $W \perp\!\!\!\perp S$

In our final simulation setting, we let $\gamma^T = (0, 0, 0)$ and $\eta^T = (0, 0, 0)$. The change in γ results in new treatment, strata, and response distributions as shown in Table 5.

Compared to the marginal distributions in Simulation Setting 1 (Table 1), the marginal strata distribution is mostly unchanged; however, the treatment distribution now heavily favors treatment 2 at the expense of treatment 3. The variability in average response across strata has also increased considerably, especially within treatment 3. Despite this pronounced increase in range, the choice between STRAT a NON-STRAT turns out to be irrelevant when both $Y \perp\!\!\!\perp S$ and $P(W = w) \perp\!\!\!\perp S$.

Using Stratified GPSM when $Y \perp\!\!\!\perp S$ and $P(W = w) \perp\!\!\!\perp S$ did tend to have the best coverage of any of the three methods; however, the gain in efficiency is now almost inconsequential. The relative improvement in MSE between STRAT and NON-STRAT has fallen to below 5%. All methods are still unbiased, though, indicating you are no worse off if you did use Stratified GPSM when the underlying relationship between the strata and the response or propensities did not support doing so. Unsurprisingly, Simulation Setting 3 is the closest STRAT and RAND have come to mimicking the results of each other as demonstrated in the results in Table 6.

4.4. *Investigating the impact of the number of strata*

In the above simulation settings, performing GPSM on randomized strata did at least as good a job in estimating treatment effect contrasts as did performing GPSM with no strata. In some instances, coverage and variance properties improved when implementing RAND. At face value, this would seem counter-intuitive, as how could adding random noise improve variability and coverage? One hypothesis is the random stratification is functioning as a make-shift caliper function, creating *de facto*

Table 6. Treatment contrast results when $Y \perp\!\!\!\perp S$ and $W \perp\!\!\!\perp S$.

Treatment Contrast	Method	Bias	MSE	Coverage (Naive)	Coverage (Corrected)
$\tau(1,2) =$ 3.98	STRAT	0.016	0.025	96.5%	94.2%
	NON-STRAT	0.005	0.026	97.3%	93.4%
	RAND	0.012	0.026	97.3%	93.1%
$\tau(1,3) =$ 1.51	STRAT	0.018	0.037	98.8%	96.5%
	NON-STRAT	(0.000)	0.039	98.7%	96.6%
	RAND	0.018	0.038	99.1%	97.0%
$\tau(2,3) =$ −2.46	STRAT	0.002	0.067	97.3%	95.6%
	NON-STRAT	(0.005)	0.069	98.0%	96.1%
	RAND	0.006	0.069	97.7%	95.7%

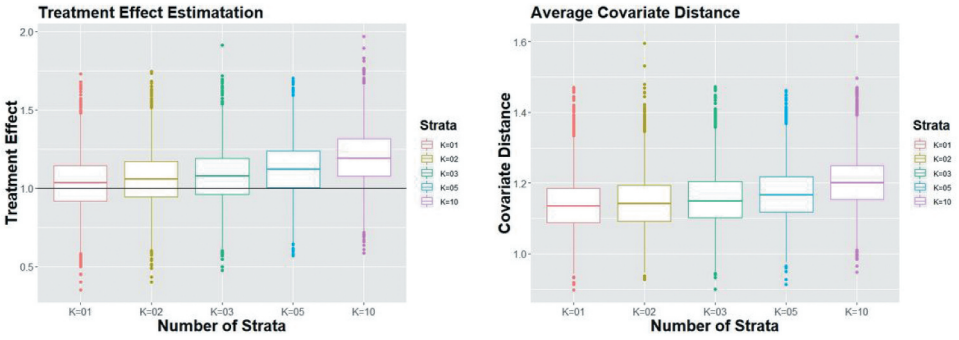

Figure 1. Treatment contrast estimate and covariate distance distributions when conducting GPSM with random strata. Left: Boxplot of treatment contrast estimates by strata. Black line represents true value of τ. Right: Boxplot of average covariate distance of matched pairs by strata

constraints on the max distance between two matched points. As the number of available matches decreases, the limits of the observed covariate space decreases as well (or more accurately is split over multiple strata). By randomly assigning units into K strata, the researcher is effectively pruning $(1 - 1/K) \times 100\%$ randomly and taking the average result of doing so over K non-overlapping sets.

Unlike the random pruning effect on PSM discussed by King and Nielsen (2019), randomly stratified GPSM still retains all treated and control units. However, it could be argued that within each random strata, the results in Simulation Settings 1–3 could simply be experiencing the initial decrease in bias and variance King and Nielson observed prior to the point where a CRT was approximated. To examine the effect of randomized strata on GPSM, we conduct a secondary simulation analysis varying the number of randomized strata and examining the distribution of distances between matched pairs.

To construct our data for the simulation we let X_1 and X_2 be independent random variables distributed $\mathcal{N}(0, 1)$. As our interest in this secondary simulation is in the variation of $S = \{1, \ldots, K\}$ and not $W = \{1, \ldots, T\}$, we will let $T = 2$ and draw $W \sim Binomial(\phi)$ and $\phi = \text{logit}^{-1}(X_1 + X_2)$. Response $Y = X_1 - X_2 + W$, and strata are randomly assigned with equal probability $Pr(S = s) = 1/K$. We vary $K \in \{1, 2, 3, 5, 10\}$. Note that $K = 1$ is the same as non-stratified GPSM. We construct data sets of $n = 500$ repeatedly for $5,000$ simulations. Results for bias and average covariate distance (the absolute Euclidean distance in the covariate space between matched units) are in Figure 1.

What becomes quickly apparent is that as the number of random strata increases so too does the bias. Likewise, as the number strata increases, so too does the average distance between observations in a matched pair but at a slower rate than the bias. At K = 3 (the number of strata used in our prior simulation results), the true value of $\tau(1, 2)$ is still contained within the inter-quartile range of the distribution; however, the covariate distance distributions is mostly unchanged from K = 1 to K = 3. Compare this to K = 10 where the IQR no longer contains the true value of *tau* and the distribution of distances has shifted noticeably upwards. Figure 1

Table 7. Frequency chart by strata and percentile for the occurrence that the covariate distance for matches under non-stratified GPSM was greater than for randomly stratified GPSM.

Percentile	K = 2	K = 3	K = 5	K = 10
90th	48.9%	46.9%	45.2%	39.1%
75th	45.6%	44.1%	40.2%	30.6%
50th	42.8%	39.2%	31.9%	20.2%
25th	38.7%	34.3%	26.4%	13.9%
10th	36.7%	33.3%	23.3%	10.5%
Average	*43.0%*	*39.0%*	*32.7%*	*19.1%*

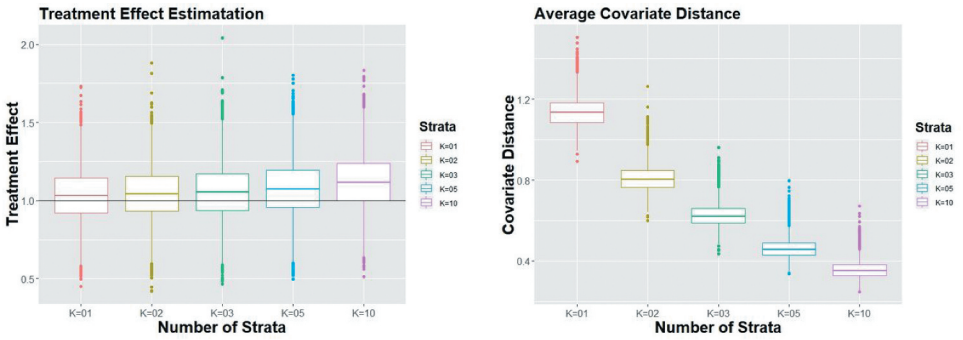

Figure 2. Treatment contrast estimate and covariate distance distributions when conducting GPSM stratified by k quantiles of X_1. Left: Boxplot of treatment contrast estimates by strata. Black line represents true value of τ. Right: Boxplot of average covariate distance of matched pairs by strata

suggests that we should expect bias to increase as additional random strata are introduced but the difference in estimates might not be evident in lower strata counts. Conversely, it does not immediately suggest a rationale for the decreases in variance observed in the simulations comparing STRAT, NON-STRAT, and RAND.

To identify an underlying cause for the decrease in variance for randomly stratified GPSM over non-stratified GPSM, we instead have to examine the within-sample differences of the covariate distance distributions. Table 7 displays the percent of simulations in which the covariate distance between matched units was greater for NON-STRAT than RAND.

As is suggested by Figure 1, on average, the covariate distance for non-stratified GPSM is shorter than for randomly stratified GPSM. What is interesting is that the superiority of NON-STRAT over RAND is not distributed evenly. At higher percentiles (as measured by covariate distance) even K = 10 strata has shorter covariate distances nearly 40% of the time. The fact that the upper tail of the distance distribution is preserved longer than the lower tail gives credence to the intuition that random strata may be mimicking the random pruning phenomena described in King and Nielsen (2019). Random strata are throwing out some of the worst matches non-stratified GPSM would select. Eventually, though, the effect on bias of accepting lower quality matches overcomes the gain from generating shorter-distance matches on the strata-pruned observations.

Comparatively, if conducting the same analysis with a relevant stratifying variable, the increase in bias is not observed, and the average covariate distance between matched units decreases with increasing strata. Figure 2 shows the simulation results of same conditions, except here, S is derived by the k-quantile function over X_1. $K = 1$ represents no stratification. $K = 2$ represents constraining matches above or below the median of X_1. Similarly, for $K \in \{3, 5, 10\}$ representing constraining matches within thirds, quintiles, and deciles respectively along X_1.

Even at $K = 10$ strata, the true value of τ is within the center-mass of the distribution; however, bias is starting to increase slightly. Recall though $n = 500$, which may not allow a sufficient number of treated or control units within each strata if over-coarsened. At such a point, there may be better (nearer) matches just on the other side of the strata boundary than within a sparsely populated strata. Repeating the simulation with larger sample size caused the apparent increase in bias to disappear. The steady decrease in covariate distance also indicates better balance among covariates with higher strata values, though the decrease in covariate distance is diminishing.

5. Application of stratified GPSM on real-world data

In this section, we illustrate the application of stratified GPSM in a real-world data application. We elect to utilize four years of the U.S. Center for Disease Control's National Health and Nutrition Examination Survey (NHANES) to examine the effect of the number of children in a household has on an individual's

systolic blood pressure (SBP). Chronic stress has long been associated with increased blood pressure (Kulkarni et al. 1998; Spruill et al. 2019), and so too has the link between parenting and increased stress (Berry and Jones 1995; Holly et al. 2019). In all of this literature, the causal relationships are admittedly nuanced. The uncertainty is compounded in the parenting stress literature with the need to introduce measures of stress that themselves required validation. Suppose a researcher was interested in using SBP as a biomarker for stress (instead of using a questionnaire) to circumvent the measurement/validation concern and wanted to use the demographic and laboratory in the NHANES data to control for other factors that may influence stress. The NHANES data does not have a continuous variable for the number of children in a household and instead truncates the upper-tail, leading to a categorical view of the treatment variable. We demonstrate how we can still investigate the question of interest via stratified GPSM, as the method can address the multiple treatment levels derived from the NHANES data while also incorporating the available covariates and stratifying variables.

To begin, we combine the two most recent, complete NHANES data sets, which collectively span the survey years 2015–2018. We construct the treatment variable by recombining the truncated variables for household young child count and household youth count variables into a single variable $W \in \{None, One, Multiple\}$ corresponding to how many children ages 0–17 are in the household. For convenience, we will refer to our treatment as child count. The response Y is directly obtained as the SBP lab value. Survey and lab values for general health satisfaction, waist circumference, cholesterol, liver function, BMI, and age are obtained to construct the covariate space X. Respondent's data for marital status ($S_{marital}$), insurance type (S_{payer}), race (S_{race}), and weight group (S_{weight}) are extracted to be used as potential stratifying variables. A complete listing of variables and derivations (where applicable) is provided in the appendix.

For each potential strata variable $S \in \{S_{race}, S_{marital}, S_{weight}, S_{payer}\}$ we conduct Stratified GPSM and compare its treatment contrast estimates and variances to what would be obtained via non-stratified GPSM and via pairwise PSM. Recall, pairwise estimation restricts analysis only to the sub-population actually treated with either treatment in the underlying contrast and not the total population as is the case with STRAT and NON-STRAT. Direct comparisons should not be made between PAIRWISE and the other two methods; however, due to the ubiquity of PAIRWISE estimation and because the population versus sub-population nuance may sometimes be missed in applied situations, we still include PAIRWISE as a juxtaposing method.

We restrict our analysis each time to respondents where Y, W, and S are all observed. To ensure the STRAT results are fairly compared to NON-STRAT and PAIRWISE results using the same data, the deletion of records missing Y, W, or the selected S at the start of each comparing analysis. This does cause the treatment contrast and variance estimates for NON-STRAT and PAIRWISE to differ slightly depending on the stratifying variable utilized by STRAT; however, the effect is minimal. Missing values for covariates are obtained via mean imputation. Generalized propensity scores were estimated via a multinomial model using the selected demographic and lab value covariates as predictors.

In Table 8, we can see the treatment contrasts vary depending on the choice of strata. Graphical depictions of the table's data are available in the appendix. Against expectations, the effect of child count on SBP consistently reported higher SBP for individuals with no children than with one or multiple children. Admittedly the treatment contrast sizes are small. For scale, prior analysis of NHANES data has shown a 10 mm Hg increase in SPB increased the risk of cardiovascular mortality among prehypertensive adults by an odds ratio of 2.11 (Greenberg 2006), so a treatment contrast of +1 mm Hg is unlikely to be medically relevant. Nonetheless, stratified GPSM was able to identify one significant contrasts that went otherwise undetected by NON-STRAT and PAIRWISE. When stratifying on insurance type, moving from no children to one child in the household significantly decreased respondents' average SBP (the same effect would be significant at the $\alpha = 0.1$ level for marital status).

Taking aside medical relevance, we also see a pattern emerge similar to the results in Section 4.4. In most instances stratifying decreases the variance of treatment contrasts. This is true for all contrasts for insurance type. Conversely, stratifying on race increased variance more times than it decreases. Recall from Section 4.4 that stratifying on random noise is expected to increase bias and variance as the amount of noise increases.

Table 8. Treatment effect, 95% CI, and variance estimates for the effect of child count on respondent's SPB broken out by stratifying variable, estimation method, and treatment contrast.

Strata = Marital Status				Strata = Insurance Type			
Contrast	Estimation Method	Treatment Effect	Variance	Contrast	Estimation Method	Treatment Effect	Variance
Multiple vs One	NO_STRAT	0.234 (−1.327, 1.795)	0.634	Multiple vs One	NO_STRAT	−0.561 (−2.178, 1.056)	0.681
Multiple vs One	PAIRWISE	−0.173 (−1.041, 0.695)	0.196	Multiple vs One	PAIRWISE	−0.037 (−0.912, 0.838)	0.199
Multiple vs One	STRAT	1.236 (−0.346, 2.818)	0.651	Multiple vs One	STRAT	0.687 (−0.872, 2.246)	0.633
None vs Multiple	NO_STRAT	0.788 (−0.954, 2.529)	0.789	None vs Multiple	NO_STRAT	1.342 (−0.209, 2.893)	0.626
None vs Multiple	PAIRWISE	0.240 (−1.448, 1.928)	0.742	None vs Multiple	PAIRWISE	0.671 (−1.149, 2.490)	0.862
None vs Multiple	STRAT	0.293 (−1.276, 1.862)	0.641	None vs Multiple	STRAT	1.064 (−0.360, 2.488)	0.528
None vs One	NO_STRAT	1.022 (−0.638, 2.683)	0.718	None vs One	NO_STRAT	0.781 (−0.730, 2.292)	0.594
None vs One	PAIRWISE	0.623 (−0.886, 2.132)	0.593	None vs One	PAIRWISE	0.298 (−1.358, 1.953)	0.713
None vs One	STRAT	1.529 (−0.015, 3.073)	0.621	None vs One	STRAT	1.751 (0.488, 3.014)	0.415
Strata = Race				**Strata = Weight Group**			
Contrast	**Estimation Method**	**Treatment Effect**	**Variance**	**Contrast**	**Estimation Method**	**Treatment Effect**	**Variance**
Multiple vs One	NO_STRAT	−0.429 (−2.057, 1.199)	0.690	Multiple vs One	NO_STRAT	0.06 (−1.575, 1.695)	0.696
Multiple vs One	PAIRWISE	−0.043 (−0.928, 0.842)	0.204	Multiple vs One	PAIRWISE	−0.147 (−1.008, 0.715)	0.193
Multiple vs One	STRAT	0.243 (−1.641, 2.128)	0.925	Multiple vs One	STRAT	0.287 (−1.429, 2.003)	0.766
None vs Multiple	NO_STRAT	1.333 (−0.271, 2.936)	0.669	None vs Multiple	NO_STRAT	0.994 (−0.564, 2.551)	0.632
None vs Multiple	PAIRWISE	0.437 (−1.173, 2.046)	0.674	None vs Multiple	PAIRWISE	0.316 (−1.327, 1.959)	0.703
None vs Multiple	STRAT	0.736 (−1.074, 2.546)	0.853	None vs Multiple	STRAT	−1.271 (−2.809, 0.267)	0.616
None vs One	NO_STRAT	0.903 (−0.571, 2.378)	0.566	None vs One	NO_STRAT	1.053 (−0.463, 2.569)	0.598
None vs One	PAIRWISE	−0.222 (−1.78, 1.336)	0.632	None vs One	PAIRWISE	0.144 (−1.477, 1.765)	0.684
None vs One	STRAT	0.979 (−0.406, 2.364)	0.499	None vs One	STRAT	−0.984 (−2.155, 0.188)	0.357

Because stratifying on race increases variance over non-stratified GPSM, it should be assumed that it is contributing little to no information for estimating the propensity score or response and can be discarded in favor on the available non-stratified GPSM estimate.

Finally, of narrative interest are the results when stratifying by insurance type. The variable S_{payer} categorizes respondents into having commercial, government, or no insurance. Because in the United States, commercial insurance is typically obtained through an employer, whereas government insurance is typically extended in response to financial need (Medicaid) or age (Medicare), the single stratifying variable S_{payer} contains information relating to a wider range of latent covariates (income, age, access to healthcare, etc.). It is no surprise then that insurance has the largest changes between NON-STRAT and STRAT for both the treatment contrasts and variance estimates. Full comparisons of covariate balance pre/post-matching are available in the Appendix.

6. Conclusions

In this article we show how GPSM can be easily extended to formally incorporate constraining pretreatment strata, allowing GPSM to better approximate a blocked randomized trial. We show how the estimation of the GPSM variance is minimally affected when switching to stratified GPSM but through simulation study show how if selected strata are informative of treatment assignment that stratified GPSM is more efficient than non-stratified GPSM. From the results where $Y \perp\!\!\!\perp S$ and $W \perp\!\!\!\perp S$ and the investigation of random stratification we illustrate how constraining matches to non-relevant strata may not pose a risk to bias or variance when the number of strata levels are small, but also how the risk increases as the number of strata levels increases. Because of how easily stratified GPSM can be rerun to generate non-stratified GPSM results (by providing a strata variable with a single level), if there is any prior uncertainty about the strength of association between a chosen strata and treatment assignment we recommend researchers obtain results for both stratified and non-stratified GPSM to ensure efficiency is being improved by the strata's inclusion. Finally, we demonstrate how stratified GPSM can be deployed in a real-world data setting to identify treatment contrasts non-stratified GPSM may overlook.

Due to stratified GPSM's ability to handle multiple treatment levels and emulate a blocked randomized design used in many clinical trials, it would be a natural extension to consider GPSM in the context of ANCOVA testing. Because stratified GPSM has forced matches within strata, any categorical pre-treatment variables would be balanced by design, leaving only continuous variables left to balance. Recent draft guidance from the United States Food and Drug Administration recommends expanding ANCOVA's usage within randomized trials (Center for Drug Evaluation and Research 2019), and Wang et al. (2019) have found ANCOVA to be consistent for estimating treatment effect estimates in randomized clinical trials (even when the ANCOVA model is misspecified). It may be possible for the benefits of ANCOVA testing to be applicable within the emulated randomized trials generated by stratified-GPSM. Yang and Kim (2019) and Yang and Kim (2020) considered the prognostic score instead of the propensity score in matching. Additionally, Yang and Zhang (2022); Zhang et al. (2021) have proposed a double score matching (DSM) algorithm that combines propensity score and prognostic score estimates to improve the efficiency of the propensity score. If DSM can be extended to work with GPS, stratified GPSM may prove to be a special case of DSM. Either way the integration of the prognostic scores into the double score may function similarly to stratified GPSM and allow for ANCOVA testing via DSM. Further work to confirm the possibility of applying ANCOVA in either stratified or prognostic score-supported GPSM is interesting and will be investigated in the future.

As a closing comment and recommendation, we advise researchers to be aware stratified GPSM still relies on a correctly specified propensity score model, so some of the critiques of traditional PSM on binary treatments from King and Nielsen (2019) are still not addressed. We would direct researchers to Wang (2020) where a broader set of recommendations is provided for using common propensity score-based matching methods in clinical practice. In their listing, stratified GPSM provides a means for incorporating their final two suggestions and does so while also extending to multiple treatments.

We hope our findings are useful for researchers and get them closer to being able to emulate the clinical trials they would have constructed if only using observational data had been avoidable. Moreover, like most matching approaches, the proposed matching estimator requires all confounders to be measured, which however is not verifiable empirically. In the future, we will develop sensitivity analysis (Yang and Lok 2018) for the stratified GPSM framework to assess the robustness of the study conclusion to the key unverifiable assumptions.

Notes

1. Data available at: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx

Acknowledgments

Yang is partly supported by NIH 1R01AG066883, 1R01ES031651, and NSF DMS 1811245.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation [DMS 1811245]; National Institutes of Health [1R01AG066883, 1R01ES031651].

ORCID

Shu Yang http://orcid.org/0000-0001-7703-707X

References

Abadie, A., and G. W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74 (1):235–267. doi:10.1111/j.1468-0262.2006.00655.x.

Abadie, A., and G. W. Imbens. 2016. Matching on the estimated propensity score. *Econometrica* 84 (2):781–807. doi:10.3982/ECTA11293.

Berry, J. O., and W. H. Jones. 1995. The parental stress scale: Initial psychometric evidence. *Journal of Social and Personal Relationships* 12 (3):463–472. doi:10.1177/0265407595123009.

Center for Drug Evaluation and Research (2019). *Adjusting for covariates in randomized clinical trials for drugs and biologics with continuous outcomes guidance for industry*.

Greenberg, J. 2006. Are blood pressure predictors of cardiovascular disease mortality different for prehypertensives than for hypertensives? *American Journal of Hypertension* 19 (5):454–461. doi:10.1016/j.amjhyper.2005.10.023.

Hirano, K., and G. W. Imbens. 2004. The propensity score with continuous . In *Treatments, Chapter 7*, ed. A. Gelman, and Meng. X, 73–84. John Wiley & Sons, Ltd. doi:10.1002/0470090456.

Holly, L. E., A. R. Fenley, T. K. Kritikos, R. A. Merson, R. R. Abidin, and D. A. Langer. 2019. Evidence- base update for parenting stress measures in clinical samples. *Journal of Clinical Child & Adolescent Psychology* PMID: 31393178. 48 (5):685–705. doi:10.1080/15374416.2019.1639515.

Iacus, S. M., G. King, and G. Porro. 2011. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* 106 (493):345–361. doi:10.1198/jasa.2011.tm09599.

Iacus, S. M., G. King, and G. Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 20 (1):1–24. doi:10.1093/pan/mpr013.

Imai, K., and M. Ratkovic. 2014. Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 76 (1):243–263. doi:10.1111/rssb.12027.

Imbens, G. W. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87 (3):706–710. doi:10.1093/biomet/87.3.706.

King, G., and R. Nielsen. 2019. Why propensity scores should not be used for matching. *Political Analysis* 27 (4):435–454. doi:10.1017/pan.2019.11.

Kulkarni, S., I. O'Farrell, M. Erasi, and M. Kochar. December 1998. Stress and hypertension. *WMJ: Official Publication of the State Medical Society of Wisconsin* 97(11):34–38.

Lechner, M. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market*, ed. M. Lechner and F. Pfeiffer, 43–58. Policies, Heidelberg: Physica-Verlag HD.

Lopez, M. J., and R. Gutman. 2017. Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science* 32 (3):432–454. doi:10.1214/17-STS612.

Lu, B., E. Zanutto, R. Hornik, and P. R. Rosenbaum. December 2001. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 96(456):1245–1253. doi:10.1198/016214501753381896.

Lunceford, J. K., and M. Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23 (19):2937–2960. doi:10.1002/sim.1903.

McCaffrey, D. F., B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette. 2013. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* 32 (19):3388–3414. doi:10.1002/sim.5753.

Rassen, J. A., D. H. Solomon, R. J. Glynn, and S. Schneeweiss. July 2011. Simultaneously assessing intended and unintended treatment effects of multiple treatment options: A pragmatic "matrix design". *Pharmacoepidemiology and Drug Safety* 20(7):675–683. doi: 10.1002/pds.2121.

Rassen, J. A., A. A. Shelat, J. M. Franklin, R. J. Glynn, D. H. Solomon, and S. Schneeweiss. 2013. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* 24 (3):401–409. doi:10.1097/EDE.0b013e318289dedf.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1):41–55. doi:10.1093/biomet/70.1.41.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5):688–701. doi:10.1037/h0037350.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63 (3):581–592. doi:10.1093/biomet/63.3.581.

Rubin, D. B. 1980. Randomization analysis of experimental data: the fisher randomization test comment. *Journal of the American Statistical Association* 75 (371):591–593.

Rubin, D. B., and N. Thomas. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95 (450):573–585. doi:10.1080/01621459.2000.10474233.

Spruill, T. M., M. J. Butler, S. J. Thomas, G. S. Tajeu, J. Kalinowski, S. F. Castañeda, A. T. Langford, M. Abdalla, C. Blackshear, M. Allison, et al. 2019. Association between high perceived stress over time and incident hypertension in black adults: findings from the jackson heart study. *Journal of the American Heart Association* 8 (21):e012139. doi:10.1161/JAHA.119.012139.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25 (1):1. doi:10.1214/09-STS313.

Tu, C., S. Jiao, and W. Y. Koh. 2013. Comparison of clustering algorithms on generalized propensity score in observational studies: A simulation study. *Journal of Statistical Computation and Simulation* 83 (12):2206–2218. doi:10.1080/00949655.2012.685169.

Wang, B., E. L. Ogburn, and M. Rosenblum. 2019. Analysis of covariance in randomized trials: more precision and valid confidence intervals, without model assumptions. *Biometrics* 75 (4):1391–1400. doi:10.1111/biom.13062.

Wang, J. 2020. To use or not to use propensity score matching? *Pharmaceutical* 20 (August):Statistics 20. doi:10.1002/pst.2051.

Wang, B., R. Susukida, R. Mojtabai, M. Amin-Esmaeili, and M. Rosenblum. 2021. Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association* 1–12.

Wu, X., F. Mealli, M.-A. Kioumourtzoglou, F. Dominici, and D. Braun. 2018. Matching on generalized propensity scores with continuous exposures. *arXiv preprint arXiv:1812.06575*.

Yang, S., G. W. Imbens, Z. Cui, D. E. Faries, and Z. Kadziola. 2016. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* 72 (4):1055–1065. doi:10.1111/biom.12505.

Yang, S., and J. J. Lok. 2018. Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica* 28 (4):1703–1723. doi:10.5705/ss.202016.0133.

Yang, S., and J. K. Kim. 2019. *Nearest neighbor imputation for general parameter estimation in survey sampling. In the econometrics of complex survey data*. Bingley, England: Emerald Publishing Limited. doi:10.1108/S0731-905320190000039012.

Yang, S., and J. K. Kim. 2020. Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scandinavian Journal of Statistics* 47 (3):839–861. doi:10.1111/sjos.12429.

Yang, S., and Y. Zhang. 2022. Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics*. doi:10.1111/sjos.12585.

Zhang, Y., S. Yang, W. Ye, D. E. Faries, I. Lipkovich, and Z. Kadziola. 2021. Practical recommendations on double score matching for estimating causal effects. *Statistics in Medicine*. doi:10.1002/sim.9289. (Accessed March 20, 2022).

A1 Appendix

A1.1 Improvement in variance of stratified GPSM over non-stratified GPSM

To show that the stratified GPSM estimator is more efficient than the non-stratified GPSM, we will compare their asymptotic variances. First, we write

$$\sigma^2_{STRAT}(w) = E_{X,S}\left[\{\bar{\mu}(w, p(w|X_i), S_i) - \tau(w)\}^2\right]$$

$$+ E_{X,S}\left[\bar{\sigma}^2(w, p(w|X_i), S_i) \times \{3/(2p(w|X_i)) - p(w|X_i)/2\}\right]$$

$$+ \bar{\sigma}^2(w, p(w|X_i), S_i) - \bar{\sigma}^2(w, p(w|X_i), S_i)$$

$$= E_{X,S}\left[\{\bar{\mu}(w, p(w|X_i), S_i) - \tau(w)\}^2 + \bar{\sigma}^2(w, p(w|X_i), S_i)\right]$$

$$+ E_{X,S}\left[\bar{\sigma}^2(w, p(w|X_i), S_i) \times \{3/(2p(w|X_i)) - p(w|X_i)/2 - 1\}\right]$$

$$= \mathbb{V}\{Y(w)\} + E_{X,S}\left[\bar{\sigma}^2(w, p(w|X_i), S_i) \times \{3/(2p(w|X_i)) - p(w|X_i)/2 - 1\}\right]. \quad \text{(A1)}$$

Similarly it can be shown

$$\sigma^2_{NON-STRAT}(w) = \mathbb{V}\{Y(w)\} + {}_{X,S}\left[\bar{\sigma}^2(w, p(w|X_i)) \times \{3/(2p(w|X_i)) - p(w|X_i)/2 - 1\}\right]. \quad \text{(A2)}$$

The first terms are the same between equations (A1) and (A2), and due to the choice to model the GPS at the population level instead of within each strata, so too are the last terms. By assumption $\bar{\sigma}^2\{w, p(w|X_i), S_i\} \leq \bar{\sigma}^2\{w, p(w|X_i)\}$ implies $\sigma^2_{STRAT}(w) \leq \sigma^2_{NON-STRAT}(w)$.

A1.2 NHANES Data Descriptions and Derivations

Table A1 catalogs the variable, usage, description, and derivation of each variable used in the analysis of child count on systolic blood pressure. Variables in all capital letters in the Source/Derivation column indicate variable names as they appear in the publicly available NHANES data.[1]

Table A1: Description and Derivations of NHANES variables used in the analysis of the effect of child count on systolic blood pressure.

Variable	Use	Description	Source/Derivation
bp_sys	Response	Systolic blood pressure in mm Hg	BPXSY1
Child_Group	Treatment	Treatment indicator for if a household has	child_count=DMDHHSZA+DMDHHSZB;
		0, 1, or 2+ children living in the household who are ages 0-17	if child_count=0 then Child_group='No Child';
			else if Child_count=1 then Child_group='1 Child';
			else if Child_count≥2 then Child_group='2+ Child';
health_status	Covariate	Self-assessed satisfaction of current health (5 point scale)	if 0<HSD010≤5 then health_status=HSD010;
waist	Covariate	Waist circumference in cm	BMXWAIST
liv_alt	Covariate	Alanine Aminotransferase (ALT) in U/L. ALT is a common	LBXSATSI
		biomarkder of liver function with higher levels indicative of	
		worse liver function.	
chol_tot	Covariate	Total Cholesterol in mg/dL. Cholesterol is a common	LBXSCH
		biomarker of cardiac health with higher leverls	
		indicative of worse cardiovascular disease prognosis.	
BMI	Covariate	Body mass index	BMXBMI
age	Covariate	Chronological age in years	RIDAGEYR
marital_status	Strata	Categorization of marital status into Married/ Cohabitating,	if year=2018 then do;
		Widowed/Divorced/Separated, Never Married, or Other	if DMDHRMAZ = 1 then marital_status='Married/ Cohab';
			else if DMDHRMAZ = 2 then marital_status='Widow/ Div/Sep';
			else if DMDHRMAZ = 3 then marital_status='Never Married';
			else if DMDHRMAZ > 3 then marital_status='Other';
			end;
			if year=2016 then do;
			if DMDHRMAR = 1 or DMDHRMAR=6 then marital_status='Married/Cohab';
			else if DMDHRMAR = 2 or DMDHRMAR=3
			or DMDHRMAR=4 then marital_status='Widow/Div/ Sep';
			else if DMDHRMAR = 5 then marital_status='Never Married';
			else if DMDHRMAR > 6 then marital_status='Other';
			end;
payer_type	Strata	Categorization of source of health insurance into	if hiq011=2 then payer_type='No Ins';
		commercial, government, and no insurance	else if HIQ031A=14 THEN payer_type='Comm';

(*Continued*)

(Continued).

Variable	Use	Description	Source/Derivation
			else if HIQ031B=15 THEN payer_type='Govt';
			else if HIQ031C=16 THEN payer_type='Govt';
			else if HIQ031D=17 THEN payer_type='Govt';
			else if HIQ031E=18 THEN payer_type='Govt';
			else if HIQ031F=19 THEN payer_type='Govt';
			else if HIQ031H=21 THEN payer_type='Govt';
			else if HIQ031I=22 THEN payer_type='Govt';
race	Strata	Reported race of survey respondent	if RIDRETH1 = 1 then race='Mex.Amer';
			else if RIDRETH1 = 2 then race='Oth.Hisp';
			else if RIDRETH1 = 3 then race='White';
			else if RIDRETH1 = 4 then race='Black';
			else race='Other';
weight_group	Strata	Categorization of BMI into Under/Normal weight,	if MISSING(bmi) then weight_group="";
		Overweight, and Obese	else if 0≤bmi<25 then weight_group='1.UNDER/NORMAL';
			else if bmi<30 then weight_group='2.OVER';
			else if bmi≥30 then weight_group='3.OBESE';

A1.3 Pre/Post-matching Balance of NHANES Covariates

In table A2 we see initial and post-matching covariate balance for each variable in each contrast in terms of standard deviations for all four proposed stratifying variables. Both stratified and non-stratifed GPSM improve overall bias over the initial sample means. In all four instances, stratified-GPSM results in a more-balanced overall than for non-stratified GPSM. Indicator arrows are provided to show which column has better balance within each contrast-variable pair.

Table A2: Initial and post-matching covariate balance of NHANES variables used in the analysis of the effect of child count on systolic blood pressure.

Strata	Variable	Contrast	Initial Balance	Non-Strat Balance	More Balanced	Strat Balance	Strata	Variable	Contrast	Initial Balance	Non-Strat Balance	More Balanced	Strat Balance
marital	age	Multiple vs One	0.36	0.01	&	0.02	payer_type	age	Multiple vs One	0.36	0.01	%	0.00
		None vs One	1.00	0.05	=	0.04			None vs One	0.99	0.04	%	0.03
		None vs Multiple	1.40	0.04	%	0.02			None vs Multiple	1.39	0.03	=	0.03
marital	health_status	Multiple vs One	0.10	0.06	&	0.13	payer_type	health_status	Multiple vs One	0.10	0.07	%	0.03
		None vs One	0.06	0.16	&	0.25			None vs One	0.06	0.18	&	0.23
		None vs Multiple	0.16	0.10	&	0.12			None vs Multiple	0.16	0.12	&	0.21
marital	waist	Multiple vs One	0.17	0.01	%	0.01	3*payer_type	waist	Multiple vs One	0.18	0.01	&	0.02
		None vs One	0.24	0.13	%	0.12			None vs One	0.23	0.16	%	0.14
		None vs Multiple	0.42	0.15	%	0.12			None vs Multiple	0.42	0.16	&	0.16
marital	liv_alt	Multiple vs One	0.04	0.03	%	0.00	payer_type	liv_alt	Multiple vs One	0.04	0.04	%	0.00
		None vs One	0.01	0.00	=	0.00			None vs One	0.00	0.01	&	0.02
		None vs Multiple	0.04	0.03	%	0.00			None vs Multiple	0.04	0.03	%	0.02
marital	chol_tot	Multiple vs One	0.15	0.06	&	0.06	payer_type	chol_tot	Multiple vs One	0.15	0.06	%	0.04
		None vs One	0.16	0.13	%	0.12			None vs One	0.16	0.15	%	0.12
		None vs Multiple	0.31	0.07	%	0.05			None vs Multiple	0.31	0.08	=	0.08
marital	BMI	Multiple vs One	0.13	0.02	%	0.00	payer_type	BMI	Multiple vs One	0.13	0.02	&	0.03
		None vs One	0.06	0.29	%	0.17			None vs One	0.06	0.27	%	0.19
		None vs Multiple	0.20	0.31	%	0.16			None vs Multiple	0.19	0.26	%	0.23
Sum of Imbalance			*5.00*	*1.66*	%	*1.40*	***Sum of Imbalance***			*4.96*	*1.70*	%	*1.58*
race	age	Multiple vs One	0.35	0.01	=	0.01	weight_group	age	Multiple vs One	0.35	0.01	=	0.01
		None vs One	0.98	0.04	=	0.04			None vs One	0.98	0.04	%	0.03
		None vs Multiple	1.37	0.03	=	0.03			None vs Multiple	1.37	0.03	%	0.02
race	health_status	Multiple vs One	0.09	0.11	%	0.06	weight_group	3*health_status	Multiple vs One	0.10	0.09	%	0.07
		None vs One	0.06	0.23	&	0.24			None vs One	0.05	0.20	%	0.19
		None vs Multiple	0.15	0.13	&	0.19			None vs Multiple	0.15	0.12	&	0.13
race	waist	Multiple vs One	0.17	0.02	%	0.00	weight_group	waist	Multiple vs One	0.18	0.02	%	0.01
		None vs One	0.23	0.17	=	0.16			None vs One	0.23	0.18	%	0.01
		None vs Multiple	0.41	0.15	&	0.17			None vs Multiple	0.41	0.16	%	0.00
race	liv_alt	Multiple vs One	0.05	0.01	=	0.01	weight_group	liv_alt	Multiple vs One	0.05	0.02	=	0.03
		None vs One	0.00	0.02	&	0.04			None vs One	0.00	0.03	&	0.02
		None vs Multiple	0.04	0.01	&	0.05			None vs Multiple	0.04	0.00	&	0.05
race	chol_tot	Multiple vs One	0.15	0.02	&	0.04	weight_group	chol_tot	Multiple vs One	0.15	0.02	=	0.02
		None vs One	0.16	0.13	%	0.07			None vs One	0.17	0.10	&	0.14

(*Continued*)

(Continued).

Strata	Variable	Contrast	Initial Balance	Non-Strat Balance	More Balanced	Strat Balance	Strata	Variable	Contrast	Initial Balance	Non-Strat Balance	More Balanced	Strat Balance
		None vs Multiple	0.31	0.10	=	0.10			None vs Multiple	0.31	0.08	&	0.11
race	BMI	Multiple vs One	0.13	0.03	%	0.01	weight_group	BMI	Multiple vs One	0.13	0.01	=	0.01
		None vs One	0.06	0.27	%	0.20			None vs One	0.06	0.25	%	0.05
		None vs Multiple	0.19	0.25	%	0.20			None vs Multiple	0.19	0.25	%	0.05
Sum of Imbalance			*4.92*	*1.73*	*%*	*1.62*	***Sum of Imbalance***			*4.93*	*1.62*	*%*	*0.96*

A1.4 NHANES Treatment and Variance Results

Depictions of the treatment contrast estimates, variances, and 95% confidence intervals for the effect of child count on respondent's SPB broken out by stratifying variable, estimation method, and treatment contrast are depicted in Figure A1 – A3.

Figure A1. Treatment contrast estimates for the effect of child count on respondent's SPB broken out by stratifying variable, estimation method, and treatment contrast. Rows denote treatment contrast, and columns denote stratifying variable.

Figure A2. Variance estimates for the effect of child count on respondent's SPB broken out by stratifying variable, estimation method, and treatment contrast. Rows denote treatment contrast, and columns denote stratifying variable.

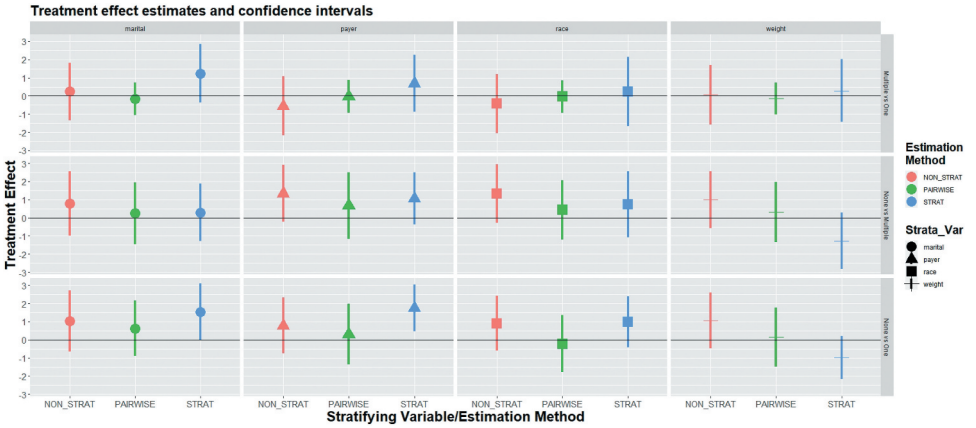

Figure A3. Treatment contrast and 95% confidence interval (CI) for the effect of child count on respondent's SPB broken out by stratifying variable, estimation method, and treatment contrast. Rows denote treatment contrast, and columns denote stratifying variable.